The Black Box Problem: AI Decision-Making in Critical Infrastructure and Its Implications

Dmytro Batishchev and Motaz Saad *

Posted Date: 28 November 2025

doi: 10.20944/preprints202511.2276.v1

Keywords: Explainable AI; XAI; AI decision-Making; critical infrastructure

 doi:10.20944/preprints202511.2276.v1

Article

The Black Box Problem: AI Decision-Making in Critical Infrastructure and Its Implications

Dmytro Batishchev and Motaz Saad *

Dept of Engineering and Innovations, University of Salento, 73100 Lecce, Italy
* Correspondence: motaz.saad@gmail.com

Abstract

This report analyzes the critical risks associated with deploying opaque "black box" AI systems, particularly deep neural networks, within Critical Infrastructure sectors such as energy, transportation, and healthcare. It highlights how the inherent lack of transparency in these models creates significant challenges regarding safety, security, and accountability, particularly in time-critical scenarios where human intervention is impossible. The study evaluates current Explainable AI (XAI) techniques, concluding that they currently lack the robustness and stability necessary to guarantee safety in high-stakes environments. Consequently, the report recommends a multi-layered mitigation strategy that includes rigorous Verification and Validation (V&V) protocols, sector-specific safety standards, and strict limitations on the use of autonomous, opaque AI in high-consequence roles.

Keywords: Explainable AI; XAI; AI decision-Making; critical infrastructure

---

1. Introduction

1.1. The Indispensable Role of Critical Infrastructure

Modern society is fundamentally reliant on a complex web of interconnected systems known as critical infrastructure (CI). These are the assets, systems, and networks, both physical and virtual, deemed so vital that their incapacitation or destruction would inflict debilitating consequences on national security, economic stability, public health, or safety.[1] The U.S. Cybersecurity and Infrastructure Security Agency (CISA) identifies 16 such sectors, encompassing domains like Energy, Transportation Systems, Water and Wastewater Systems, Healthcare and Public Health, Financial Services, and Nuclear Reactors, Materials, and Waste.[1] These sectors do not operate in isolation; they exhibit profound interdependencies.[2] The Energy Sector, for instance, provides an essential "enabling function" across all other critical infrastructure sectors, highlighting how disruptions in one area can rapidly cascade, potentially leading to widespread systemic failure.[2] The smooth, reliable, and secure operation of these infrastructures is not merely a convenience but a prerequisite for societal functioning and national well-being.[3] Any significant disruption, whether accidental or malicious, carries the potential for severe and far-reaching negative impacts.[1]

1.2. The Rise of AI in Critical Systems Operation and Management

In recent decades, Artificial Intelligence (AI) and its subfield, Machine Learning (ML), have transitioned from theoretical concepts to powerful tools deployed across numerous industries. Critical infrastructure sectors are increasingly integrating AI/ML technologies to enhance operational efficiency, optimize resource allocation, improve predictive maintenance, enable greater automation, and bolster security and resilience.[3] The allure of AI lies in its capacity to analyze vast datasets, identify complex patterns invisible to human analysts, and make predictions or decisions at speeds exceeding human capabilities.[4]

Across CI domains, the applications are diverse and growing. Energy grids leverage AI for predictive maintenance to anticipate equipment failures, optimize power flow, forecast demand, and

 doi:10.20944/preprints202511.2276.v1

detect anomalies that might indicate cyber or physical threats.[5] Transportation systems employ AI for autonomous vehicle navigation, intelligent traffic management, logistics optimization, and enhancing safety features.[6] Healthcare networks utilize AI for medical diagnosis support, personalized treatment planning, accelerating drug discovery, and managing patient data.[6] Water management systems benefit from AI in real-time quality monitoring, leak detection, demand forecasting, and optimizing treatment processes.[7] Financial systems use AI for fraud detection, credit scoring, and investment management.[6] Even the highly regulated nuclear power sector is exploring AI for potential benefits in operational efficiency and cost reduction, although currently focused on non-safety-critical applications.[8] The U.S. Department of Energy's initial assessment highlights AI's potential to dramatically improve nearly all aspects of the energy sector, including security, reliability, and resilience.[5] This drive towards AI adoption is fueled by the compelling promise of significant performance improvements in systems where efficiency and reliability are paramount.

1.3. Defining the Core Challenge: The AI "Black Box Problem"

Despite the compelling benefits, the increasing reliance on sophisticated AI models, particularly those based on deep learning and neural networks, introduces a fundamental challenge: the "black box problem".[6] This term refers to the inherent lack of transparency and interpretability in the decision-making processes of these complex algorithms.[4] Users, developers, and regulators often cannot fully understand *how* a model arrives at a specific conclusion or prediction, or *why* it made a particular decision.[9] The internal workings—the intricate interplay of millions or billions of parameters within deep neural networks—are opaque and resist straightforward human comprehension.[10]

This opacity is not merely an academic curiosity; it poses significant practical barriers, especially in high-stakes domains like critical infrastructure.[6] The inability to understand AI reasoning hinders trust, making stakeholders reluctant to cede control to systems whose workings are inscrutable.[9] It complicates debugging and validation, as identifying the root cause of errors becomes difficult.[10] It raises serious questions about accountability when failures occur.[6] Furthermore, it impedes efforts to detect and mitigate biases embedded within the models or the data they were trained on.[9] This contrasts sharply with "white-box" systems, such as simpler rule-based systems or decision trees, where the decision logic is inherently transparent and traceable.[11] The challenge is further compounded by "update opacity," where understanding *why* a model's behavior changes after retraining or updates remains difficult, undermining user confidence and mental models developed through familiarity with previous versions.[12] The fundamental tension arises because the very AI models offering the greatest performance gains in complex tasks are often the least transparent, creating a direct conflict between the drive for efficiency and the non-negotiable requirements for safety, reliability, and accountability in critical systems[13]

The opacity of AI directly contributes to and amplifies a range of risks, which are especially critical in high-stakes CI environments. These risks include:

- Safety Risks.[3]
- Security Risks.[9]
- Operational Risks.[14]
- Accountability Risks.
- Time-Criticality Risks.

The complexity and opacity of AI decision-making present significant challenges for governance and regulation (Governance and Regulatory Gaps). Traditional methods of validation and oversight are often insufficient for AI. The speed of AI development and deployment may outpace the ability of regulatory frameworks to adapt. This mismatch can create gaps in ensuring the safe, secure, and responsible use of AI in CI. Specific concerns include:

- **Lack of Standards:** Absence of clear, widely accepted standards for AI assurance, safety, and security.

 doi:10.20944/preprints202511.2276.v1

- **Difficulty in Audit and Compliance:** Challenges in auditing and enforcing compliance with AI-related regulations or guidelines.
- **International Harmonization:** The need for consistent international approaches to AI governance, especially for interconnected CI systems.

AI models are highly dependent on the quality and representativeness of the data they are trained on (Data Dependence and Quality Issues). Issues like biased data, data gaps, or data corruption can lead to flawed AI performance and unforeseen consequences in CI operations. Successfully integrating AI requires careful consideration of how humans will interact with these systems (Human-Machine Interaction Challenges). Automation bias, over-reliance on AI, or inadequate training for human operators can lead to critical failures. The complexity of AI interactions within CI can lead to unexpected and unintended consequences, particularly in systems with multiple interconnected components or under novel conditions (Potential for Unintended Consequences). These challenges are interconnected and must be addressed holistically. The next section of this report will delve into the specific dimensions of AI opacity, the risks it generates, and the essential role of governance in managing these complexities within critical infrastructure.

1.4. Report Focus: Navigating Opacity, Risk, and Governance in High-Stakes AI Applications

This report undertakes a comprehensive analysis of the AI black box problem specifically within the context of critical infrastructure. The central objective is to dissect the multifaceted implications of deploying opaque AI decision-making systems in sectors vital to national and public well-being. The analysis will focus critically on the spectrum of risks introduced or exacerbated by this lack of transparency, with particular emphasis on:

- **Safety:** Failures leading to physical harm, environmental damage, digital assets corruption, or loss of human life.
- **Security:** Vulnerabilities to adversarial manipulation, data breaches, and system integrity compromises.
- **Ethical Concerns:** Issues of bias, fairness, privacy, and the erosion of trust.
- **Accountability:** Difficulties in assigning responsibility for AI-driven failures.
- **Time-Criticality:** The unique dangers posed by opaque AI operating in high-speed control loops with limited or no possibility for human oversight or intervention. (For example, when operating a nuclear reactor (even a research one), the "price" of an error is very high. If, for example, we introduce AI into the system for outputting information about the state of the reactor, for example, to analyze sensor readings and provide recommendations, the operator may simply not have time to analyze the correctness of the given recommendation. How to guarantee that the recommendation is really correct, to double-check the sensor readings this is additional time.)

The report will examine the state of the art in AI deployment across key CI sectors, delve into the methods proposed by Explainable AI (XAI) to mitigate opacity, analyze documented failures and plausible risk scenarios, and explore technical and governance-based mitigation strategies. Special attention will be paid to the regulatory landscape, particularly concerning AI use in nuclear energy and the emerging need for limitations on autonomous, opaque systems in high-risk, time-sensitive applications.

The report is structured as follows: Section 2 reviews the State of the Art, detailing the technical basis of AI opacity, mapping AI applications in CI, introducing XAI techniques, and analyzing key scientific and official reports. Section 3 describes the Research Methodology employed. Section 4 presents the Results, focusing on the identified implications and risks, especially concerning safety, security, and time-criticality. Finally, Section 5 offers Conclusions, synthesizing the findings, discussing potential solutions and regulatory approaches, and outlining future research directions. The analysis aims to provide an authoritative resource for policymakers, regulators, CI operators, and researchers grappling with the profound challenges and opportunities presented by AI in critical systems.

 doi:10.20944/preprints202511.2276.v1

2. State of the Art

2.1. Unpacking the Black Box: Technical Reasons for AI Opacity

The "black box" nature of many contemporary AI systems, particularly those achieving state-of-the-art performance, stems primarily from their inherent complexity.[9] Deep Neural Networks (DNNs), including architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), form the backbone of many advanced AI applications.[9] These networks consist of numerous layers of interconnected nodes ("neurons"), processing information through complex, non-linear transformations.[10] The sheer number of parameters (weights and biases) in these models can range from millions to billions, making it computationally infeasible and conceptually overwhelming to trace the exact path of influence from input features to final output for any given decision.[10]

Furthermore, DNNs learn hierarchical representations of data, meaning features are combined and transformed in intricate ways across multiple layers.[11] Understanding precisely which combinations of low-level features contribute to a high-level decision becomes exceptionally difficult. This contrasts sharply with simpler, intrinsically interpretable models like linear regression, decision trees, or rule-based systems, where the decision logic is explicit and readily understandable.[11] However, these simpler models often lack the capacity to capture the complex patterns present in high-dimensional data (like images, natural language, or complex sensor readings from CI systems), leading to lower predictive accuracy on challenging tasks.[13]

This creates a well-documented trade-off: increasing model complexity often yields higher performance but simultaneously reduces interpretability.[13] Models that are easy to understand (white-box) may not be powerful enough for the demands of complex CI applications, while the high-performing models (black-box) resist explanation. This fundamental tension is a core driver of the challenges addressed in this report. The difficulty is not just in understanding a static model, but also in comprehending how its internal logic shifts during training or after updates, a phenomenon termed "update opacity," which further complicates trust and verification.[12]

2.2. AI Across Critical Infrastructure Sectors: Applications and Use Cases

The integration of AI is rapidly expanding across the 16 critical infrastructure sectors identified by CISA [1], driven by the pursuit of enhanced efficiency, reliability, safety, and security. The following table summarizes key applications in the sectors most relevant to this report's focus:

Table 1. Critical Infrastructure Sectors and Representative AI Applications

Critical Sector	Description	Example AI Use Cases	Key Benefits Sought	Primary Challenges / Risks
Energy [2]	Production, transmission, distribution of electricity, oil, natural gas. Enables all other sectors.	Predictive maintenance [15], Grid optimization & control [16], Load forecasting [15], Anomaly/Threat detection [17], Operational awareness [15], High-complexity modeling. [15]	Reliability, Resilience, Security, Efficiency, Cost reduction.[5]	Cybersecurity threats [8], Safety risks from control errors [16], Data availability/quality, Complexity of interconnected systems.
Nuclear Reactors, Materials, and Waste [1]	Nuclear power plants, handling of nuclear materials/waste.	*Potential/Emerging:* Predictive maintenance, Operational optimization (cost reduction) [8], Anomaly detection [18], Support for safety/security analysis & regulation. [19] *Current use mainly non-safety.* [20]	Efficiency, Cost reduction, Enhanced safety analysis (potential) .[8]	Extreme safety criticality, Regulatory hurdles [19], Need for high assurance/verification [21], Public trust, Security concerns (physical & cyber).

 doi:10.20944/preprints202511.2276.v1

Critical Sector	Description	Example AI Use Cases	Key Benefits Sought	Primary Challenges/Risks
Transportation Systems [1]	Movement of people and goods via aviation, highways, rail, maritime, mass transit.	Autonomous Vehicles (AVs) [6], Traffic Management [22], Logistics & Route Optimization[23], Predictive Maintenance [23], Advanced Driver-Assistance Systems (ADAS)[24], Infrastructure Monitoring. [24]	Safety (reduce human error) [25], Efficiency, Reduced congestion [22], Mobility access[25], Cost reduction.	Safety risks (AV accidents) [26], Cybersecurity vulnerabilities [27], Regulatory uncertainty [28], Ethical dilemmas (AV decision-making), Public acceptance. [29]
Water and Wastewater Systems [1]	Provision of safe drinking water and wastewater treatment.	Water Quality Monitoring [7], Leak Detection [30], Demand Forecasting [30], Resource Optimization (distribution, energy) [30], Predictive Maintenance [31], Flood Prediction. [32]	Efficiency, Water conservation [31], Cost reduction, Resilience, Safety (water quality).[7]	Data quality/availability [33], System integration [7], Ensuring reliability of predictions, Cybersecurity of control systems.[34]
Healthcare and Public Health [1]	Hospitals, clinics, public health agencies, medical device manufacturers.	Medical Diagnosis (imaging, pathology) [35], Treatment Planning [6], Drug Discovery [36], Patient Data Analysis/Risk Prediction [36], Resource Management [37], Disease Surveillance. [38]	Diagnostic accuracy [35], Personalized care [36], Efficiency [35], Faster drug development [36], Improved public health response. [38]	Patient safety risks (misdiagnosis) [39], Data privacy (HIPAA) [40], Algorithmic bias [38], Regulatory approval (FDA/EMA) [41], Clinician trust/adoption. [42]
Financial Services [1]	Banks, insurance, investment firms, markets.	Fraud Detection [6], Credit Scoring/Lending [6], Algorithmic Trading [43], Investment Management [6], Risk Assessment. [9]	Efficiency, Fraud prevention, Risk management, Profitability, Regulatory compliance. [6]	Fairness/Bias in lending [6], Market stability risks (flash crashes) [43], Security vulnerabilities [44], Lack of transparency for consumers/regulators. [6]
Information Technology [1]	IT hardware, software, services underpinning other sectors.	Cybersecurity Threat Detection/Response [45], Network Optimization, Cloud resource management.	Enhanced security posture, Efficiency, Automation.	AI model security (poisoning, evasion) [46], Supply chain security [5], Explainability for security analysts.[47]
Communications [1]	Wired, wireless, satellite, broadcast networks.	Network traffic management, Predictive maintenance for infrastructure, Anomaly detection for service quality, Cybersecurity.	Network reliability, Service quality, Efficiency, Security.	Ensuring network resilience, Security of AI-managed network functions, Data privacy.

This table illustrates the pervasive nature of AI adoption across vital sectors. While the potential benefits are substantial, the inherent risks, particularly those stemming from the black box problem, necessitate careful consideration and robust mitigation strategies, as explored later in this report. The variation in risk profiles and regulatory maturity across sectors, evident even in this brief overview, suggests that sector-specific considerations are crucial for effective AI governance. For example, the extreme caution observed in the nuclear sector contrasts with the more rapid, though increasingly regulated, adoption in transportation and healthcare.

 doi:10.20944/preprints202511.2276.v1

2.3. The Pursuit of Transparency: Explainable AI (XAI)

In response to the challenges posed by opaque AI models, the field of Explainable AI (XAI) has emerged with the primary goal of rendering the decisions and predictions of AI systems understandable to humans.[9] XAI seeks to "open the black box" [11] by providing insights into the internal workings or justifying the outputs of complex models.

The motivations for pursuing XAI are multifaceted and critical, especially in the context of CI. Key goals include:

- **Building Trust:** Transparency is fundamental for users and stakeholders to trust AI systems, particularly when they automate critical decisions.[6] Understanding the 'why' behind a decision fosters confidence in its reliability.
- **Enabling Debugging and Validation:** Explanations help developers identify flaws, errors, or unintended behavior in models, facilitating debugging and improvement.[10] It supports validation that the model operates as intended.[10]
- **Ensuring Fairness and Detecting Bias:** XAI techniques can help uncover whether a model relies on sensitive attributes or reflects biases present in the training data, enabling mitigation efforts.[9]
- **Facilitating Accountability:** When failures occur, explanations can aid in determining the cause and assigning responsibility, addressing the accountability gap created by opacity. [6]
- **Meeting Regulatory Requirements:** Increasingly, regulations like the EU's GDPR emphasize transparency and may imply a "right to explanation" for automated decisions, making XAI necessary for compliance. [6]

Significant research initiatives, such as the Defense Advanced Research Projects Agency's (DARPA) XAI program, aim to create machine learning techniques that produce more explainable models while maintaining high performance, enabling humans to understand, trust, and manage AI partners effectively.[47] The intended output includes a toolkit of ML and human-computer interface modules.[48] Similarly, the National Institute of Standards and Technology (NIST) has proposed four fundamental principles for explainable AI systems: **Explanation** (system must provide reasons), **Meaningful** (explanations understandable to the user), **Explanation Accuracy** (explanations correctly reflect the system's process), and **Knowledge Limits** (system operates only within its designed conditions and confidence levels).[49]

A variety of XAI techniques have been developed, broadly falling into several categories:

While XAI offers valuable tools, a critical perspective is necessary. The proliferation of methods lacks uniformity, making it hard for practitioners to choose appropriately.[50] More importantly, as detailed in Table 2 and supported by numerous studies [51], current popular methods like LIME and SHAP have significant limitations regarding stability, fidelity, computational cost, and robustness. This gap between the promise of XAI and the reliability of current techniques is a central theme, particularly when considering their application in safety-critical CI domains where erroneous explanations could be misleading or even dangerous.[51] The focus on post-hoc explanations, while convenient, explains decisions *after* they are made, rather than guaranteeing safe behavior by design, which may be insufficient for the highest-risk applications.[47]

 doi:10.20944/preprints202511.2276.v1

Table 2. Comparative Overview of Key XAI Techniques

Technique	Type	Mechanism Description	Key Strengths	Critical Limitations / Weaknesses	Applications
LIME (Local Interpretable Model-agnostic Explanations) [52]	Model-Agnostic, Post-Hoc	Approximates black-box model locally around a specific prediction using a simpler, interpretable model (e.g., linear regression) based on perturbed input samples.	Applicable to any model type; Provides instance-specific explanations.	Instability (explanations vary with perturbations) [53]; Sensitive to sampling/perturbation strategy [54]; Local focus may miss global context [55]; Computationally intensive for many perturbations [56]; Assumes local linearity, which may fail for complex boundaries.[57]	Tabular data, Text, Images
SHAP (SHapley Additive exPlanations) [52]	Model-Agnostic, Post-Hoc	Attributes prediction contribution to each feature based on Shapley values from cooperative game theory, averaging marginal contributions across feature coalitions.	Strong theoretical foundation (fairness, consistency) [58]; Provides both local and global explanations; Model-agnostic versions exist.	Computationally expensive, especially for non-tree models or many features [54]; Can be sensitive to feature collinearity (may assign low importance to correlated features) [57]; Interpretation of values can still be complex; Approximation methods needed in practice. [54]	Tabular data, Tree models, Images, NLP
Counterfactual Explanations [4]	Model-Agnostic, Post-Hoc	Identifies the minimal changes to an input instance that would alter the model's prediction to a different outcome. Answers "what-if" questions.	Intuitive for users ("What needs to change?"); Useful for actionable recourse; Highlights decision boundaries.	Finding the *minimal* or *most plausible* counterfactual can be computationally hard; May generate unrealistic or infeasible input changes; Multiple counterfactuals might exist.	Tabular data, Images, Text
Attention Mechanisms [59]	Model-Specific (esp. Transformers)	Internal model components that assign weights to different parts of the input sequence (e.g., words in a sentence) based on their relevance for predicting the output.	Provides direct insight into model's focus during processing; Can be visualized easily; Integral part of many state-of-the-art NLP/Vision models.	Attention weights may not always equate to true feature importance/explanation [60]; Primarily applicable to specific architectures (Transformers).	NLP, Computer Vision

 doi:10.20944/preprints202511.2276.v1

Saliency Maps / Gradient-based / CAM [61]	Model-Specific (esp. CNNs)	Visualizes the importance of input features (e.g., pixels) by computing gradients of the output with respect to the input, or using activation map weights (CAM, Grad-CAM).	Visually intuitive for image data; Computationally relatively efficient; Highlights influential input regions.	Can be noisy or unstable [51]; Susceptible to gradient saturation issues [62]; May not reflect true model reasoning (can be fooled); Primarily for differentiable models (CNNs).	Computer Vision (Images)
LRP / DTD [62]	Model-Specific (esp. DNNs)	Propagates prediction relevance backwards through the network layers to attribute relevance scores to input features.	Provides detailed layer-wise insights; Can handle non-linearities better than simple gradients.	Can be complex to implement; Theoretical justification varies between propagation rules; Computationally more intensive than simple gradients.	Computer Vision, DNNs
Inherently Interpretable Models [11]	Intrinsic	Models whose structure is inherently understandable (e.g., linear regression coefficients, decision tree paths, rule lists).	High transparency by design; Easy to understand decision logic; Verification might be simpler.	Often lower predictive performance on complex tasks compared to black-box models [13]; May oversimplify complex relationships.	Simpler classification/regression tasks, Rule discovery

2.4. Analysis of Seminal Works: Insights from Key Scientific and Official Reports

Several key reports and initiatives provide authoritative context on the state of AI, its risks, and the pursuit of explainability, particularly concerning critical infrastructure:

- **U.S. Department of Energy (DOE) AI Risk Assessment (April 2024) [5]:** This assessment, focused on critical energy infrastructure, acknowledges AI's tremendous potential benefits for security, reliability, and resilience. However, it crucially identifies four major categories of risk:
  1. *Unintentional failure modes:* Including issues like bias, extrapolation errors beyond training data, and misalignment with intended goals.
  2. *Adversarial attacks against AI:* Highlighting vulnerabilities like data poisoning and evasion attacks targeting AI models used in energy systems.
  3. *Hostile applications of AI:* Recognizing the potential for adversaries to use AI tools to facilitate attacks on energy infrastructure.
  4. *AI software supply chain compromise:* Addressing risks embedded within the development and deployment pipeline. The report strongly emphasizes the need for regularly updated, risk-aware best practices to guide the safe and secure deployment of AI in this critical sector, signaling a move towards structured risk management.
- **NIST AI Risk Management Framework (AI RMF 1.0) (Jan 2023) [63]:** This voluntary framework provides a cross-sectoral approach to managing AI risks. It is structured around four functions: **Govern** (establishing risk management culture and processes), **Map** (contextualizing risks and benefits), **Measure** (analyzing and tracking risks), and **Manage** (prioritizing and responding to risks). Central to the AI RMF is the concept of **trustworthy AI**, defined by characteristics including validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy-enhancement, and fairness with bias management.[64] While voluntary, it provides a comprehensive structure for organizations, including CI operators, to think systematically about AI risks throughout the system lifecycle. Its existence underscores the recognized need for structured approaches to AI assurance.

 doi:10.20944/preprints202511.2276.v1

- **IAEA / Nuclear Regulators (NRC, ONR, CNSC) on AI (Ongoing) [20] :** The nuclear sector exhibits a highly cautious and collaborative approach. Key insights include:
  1. *Current Use Limitation:* AI is presently focused on non-safety-related applications within nuclear power plants. [20]
  2. *Regulatory Readiness:* Agencies like the NRC have strategic plans (FY2023-2027) to ensure readiness for reviewing AI applications. [19] Initiatives like the NEA's RegLab aim to facilitate safe AI application for Small Modular Reactors (SMRs).[65]
  3. *International Collaboration:* Strong emphasis on international cooperation (IAEA, bilateral agreements) to share best practices, conduct research, and influence standards. [19]
  4. *Adapting Existing Frameworks:* Recognition that new AI-specific standards may take time, thus requiring adaptation of existing nuclear standards coupled with consideration of unique AI attributes.[66]
  5. *Guiding Principles:* Trilateral papers (UK, US, Canada) outline principles focusing on managing AI based on failure consequences, autonomy levels, human factors (including trust and oversight), AI lifecycle management, and the need for robust safety cases.[67] The opacity of generative AI currently limits its use in operations.[66] This measured, principle-based approach reflects the sector's extremely low tolerance for risk.
- **DARPA Explainable AI (XAI) Program [47]:** This foundational program aimed to create ML techniques yielding explainable models without sacrificing performance, enabling human users to understand, trust, and manage AI partners.[48] Its strategy involved developing a portfolio of XAI methods and integrating them with human-computer interfaces for effective explanation delivery.[48] The program's goal was to produce a toolkit library to facilitate future explainable AI system development.[48] Its existence and goals highlight the long-standing recognition within defense and research communities of the critical need for AI transparency.

These seminal works collectively illustrate a growing awareness of AI risks across different domains, the development of frameworks for managing these risks (NIST AI RMF), the cautious, principle-driven approach in highly regulated sectors like nuclear energy, and the ongoing research efforts (DARPA XAI) to develop technical solutions for the black box problem. However, they also implicitly reveal the gap between the ambition for trustworthy AI and the current capabilities and limitations of AI and XAI technologies, particularly for the most demanding critical infrastructure applications.

3. Description of Work

3.1. Research Methodology

The research presented in this report was conducted through a comprehensive literature review and synthesis methodology. Given the interdisciplinary nature of the subject matter—spanning artificial intelligence, machine learning, cybersecurity, safety engineering, critical infrastructure operations, ethics, and regulatory policy—a broad review of diverse, high-quality sources was deemed essential.

The primary corpus for this analysis comprises the research material provided.[9] This material includes a wide range of source types, which were prioritized based on their relevance and authority:

- **Peer-Reviewed Scientific Papers:** Publications sourced from academic databases and preprint servers (e.g., ArXiv, IEEE Xplore, ACM Digital Library, PubMed Central, ResearchGate) detailing technical aspects of AI/ML, XAI methods, cybersecurity threats, formal verification, robust ML, safe RL, and specific applications in CI sectors. [17]
- **Official Government and Agency Reports:** Documents published by national and international bodies responsible for CI oversight, regulation, and research (e.g., U.S. Department of Energy (DOE), Cybersecurity and Infrastructure Security Agency (CISA), National Institute of Standards and Technology (NIST), National Highway Traffic Safety Administration (NHTSA), Food and Drug Administration (FDA), International Atomic Energy Agency (IAEA), Nuclear Regulatory

 doi:10.20944/preprints202511.2276.v1

Commission (NRC), Office for Nuclear Regulation (ONR), Canadian Nuclear Safety Commission (CNSC), European AI Office (part of European Commission), European Medicines Agency (EMA)). [1]

- **Technical Standards Documents:** References to standards developed by organizations like the International Organization for Standardization (ISO) and SAE International relevant to AI safety and specific applications like autonomous vehicles.[68]
- **Reputable Expert Analyses:** Reports and publications from recognized research institutions and think tanks specializing in AI safety, security, and policy (e.g., RAND Corporation, Center for Security and Emerging Technology (CSET), Belfer Center, European AI Office).[69]
- **Documented Incident Reports and Databases:** Information from sources tracking real-world AI failures, near-misses, or security incidents, including the AI Incident Database and reports analyzing specific events.[70]

The synthesis involved systematically extracting relevant information pertaining to the core themes of the user query, cross-referencing findings across different source types, identifying points of consensus and divergence, and analyzing the implications of the collected evidence.

3.2. Analytical Focus

The analysis derived from the literature review concentrated specifically on addressing the key elements outlined in the user query. The core areas of investigation were:

- **The Black Box Problem:** Defining the phenomenon, exploring its technical origins in complex AI models (especially DNNs), and understanding why it hinders trust, accountability, and validation.[9]
- **AI in Critical Infrastructure:** Identifying and describing the deployment of AI across various CI sectors, with specific attention to energy (including nuclear power plants), transportation systems, water/wastewater systems, healthcare networks, and financial systems.[2]
- **Explainable AI (XAI):** Researching the field of XAI, detailing its goals, common techniques (e.g., LIME, SHAP, attention, counterfactuals, intrinsic methods), and critically evaluating their capabilities and documented limitations.[53]
- **Risks and Consequences:** Investigating the specific risks, vulnerabilities, and potential negative consequences associated with using non-transparent AI in CI. This included a primary focus on safety failures (especially those posing a risk to human life), security breaches (including adversarial attacks), biased outcomes, lack of accountability, and operational disruptions.[70] *Particular emphasis was placed on analyzing risks in time-critical scenarios where human oversight is inherently limited or impossible.*
- **Illustrative Examples:** Finding and analyzing documented case studies, significant incidents (real-world failures or cyberattacks), or well-analyzed hypothetical scenarios that demonstrate the potential impact of the black box problem in CI contexts.[70]
- **Mitigation Strategies:** Researching proposed technical solutions (including advanced XAI, robust ML design, formal verification, safe RL, AI assurance frameworks), best practices, ethical guidelines, and regulatory frameworks aimed at managing these risks.[69] *Specific attention was given to documents related to the regulation of AI in nuclear energy and the need for limitations in high-risk, time-sensitive applications.*
- **Synthesis:** Consolidating the findings to construct the core arguments and evidence base for each section of the report structure (Introduction, State of the Art, Description of Work, Results, Conclusions), ensuring logical flow and comprehensive coverage of the research theme.

The adoption of a broad literature review methodology was necessitated by the topic's inherent complexity, touching upon deep technical specifics of AI, operational realities of diverse CI sectors, principles of safety and security engineering, ethical considerations, and the dynamic landscape of global regulation. No single empirical study could encompass this breadth. Furthermore, triangulating information from scientific literature (providing technical depth and methodological rigor),

 doi:10.20944/preprints202511.2276.v1

official reports (offering policy context, authoritative risk assessments, and regulatory stances), and incident/case study analyses (grounding the discussion in real-world events or plausible scenarios) allows for a more robust and credible analysis than relying on any single type of source material alone. This approach enables the synthesis of diverse perspectives into a coherent understanding of the challenges and potential pathways forward.

4. Results: Implications and Risks of Black Box AI in Critical Infrastructure

The deployment of AI systems, particularly opaque "black box" models, within critical infrastructure introduces a complex landscape of risks that extend beyond traditional software vulnerabilities. The lack of transparency into AI decision-making processes has profound implications for safety, security, ethics, and accountability, especially in systems where failures can have catastrophic consequences and human oversight is limited.

4.1. Safety Implications: Failures, Human Harm, and Environmental Risks

The foremost concern regarding opaque AI in CI is the potential for safety failures leading to direct harm.[3] When the reasoning behind an AI's decision is not understandable, it becomes exceedingly difficult to anticipate, detect, or diagnose potential failure modes before they manifest as adverse events.[26] AI systems, especially those based on ML, learn patterns from data, but they may fail unpredictably when encountering situations outside their training distribution or when subtle changes in input lead to large errors – a property often described as "brittleness".[71]

In safety-critical CI applications, such failures can have dire consequences:

- **Healthcare:** An AI diagnostic tool misinterpreting a medical image (e.g., radiology, pathology) could lead to a missed diagnosis or incorrect treatment, directly impacting patient health and potentially causing serious injury or death.[39] Biased algorithms might allocate resources unfairly, exacerbating health disparities.[38]
- **Transportation:** Errors in perception, prediction, or planning by an autonomous vehicle's AI could result in collisions, endangering passengers, pedestrians, and other road users.[25] Failures in AI-driven traffic management could lead to increased congestion or accidents.[72]
- **Energy Systems:** Malfunctions in AI controlling power grid operations could lead to blackouts, equipment damage, or instability. [34] While currently limited in nuclear safety systems, a future failure in AI-based monitoring or control could have catastrophic consequences.[67]
- **Water Systems:** Errors in AI managing water treatment processes could compromise water quality, posing public health risks.[34] Failure to detect leaks or manage distribution effectively could lead to resource shortages or infrastructure damage.[31]

These "AI accidents" represent unintended and harmful behavior stemming from factors like incomplete specifications, unforeseen interactions with the environment, or limitations in the AI model itself. [26] The opacity of the AI makes root cause analysis difficult, hindering the ability to learn from failures and prevent recurrence.[73] Furthermore, verifying the safety of these complex, often adaptive systems using traditional methods is challenging; testing alone cannot guarantee the absence of critical flaws, especially for rare but high-consequence events.[16] Formal verification methods, while promising for providing stronger guarantees, face significant scalability challenges when applied to complex AI models.[74]

4.2. Security Vulnerabilities: Adversarial Threats and System Integrity

The black box nature of AI systems creates novel attack surfaces and vulnerabilities that differ fundamentally from traditional cybersecurity threats.[75] Adversaries can exploit the opacity and data dependency of AI models to manipulate their behavior in ways that are difficult to detect.[44] Key threats include:

 doi:10.20944/preprints202511.2276.v1

- **Adversarial Attacks:** These involve crafting malicious inputs, often subtly modified from legitimate ones, designed specifically to deceive an AI model into making incorrect predictions or classifications.[5]
- **Evasion Attacks:** Inputs are altered at inference time to cause misclassification. Examples include adding small patches to stop signs to make an AV misread them [71], or subtly modifying network traffic data to bypass an AI-based intrusion detection system.[76]
- **Poisoning Attacks:** Malicious data is injected into the training set to corrupt the learned model, potentially creating backdoors or degrading performance.[5] An attacker could poison data for grid control AI to make it misinterpret normal operations or ignore signs of equipment failure [77], or poison malware datasets to cause classifiers to mislabel malicious software as benign.[76]
- **Model Inversion/Extraction:** Attackers query a model to infer sensitive information about the training data (e.g., patient records in a healthcare AI) or to steal the model itself.[44]
- **AI Software Supply Chain Compromises:** Vulnerabilities or malicious code can be introduced through third-party libraries, pre-trained models, or data sources used in AI development.[5]

These attacks can directly compromise the integrity, availability, and confidentiality of CI systems.[44] A successful adversarial attack could cause an AI system responsible for safety to fail (blurring the line between security and safety incidents), disrupt critical operations, or lead to data breaches.[44] The opacity of the models makes detecting these subtle manipulations extremely challenging.[78]

Furthermore, AI itself is becoming a tool for attackers, lowering the barrier for sophisticated attacks against CI by enabling automated vulnerability discovery, crafting highly convincing phishing emails, or coordinating physical attacks using AI-enhanced systems like drones.[8] Ironically, the very techniques developed for XAI, while intended to increase transparency, could potentially be exploited by attackers to better understand system vulnerabilities and design more effective attacks – the "double-edged sword" of explainability.[47] Addressing these security risks requires not only traditional cybersecurity measures but also novel defenses specifically tailored to the unique vulnerabilities of AI models and the potential for AI-driven attacks.

4.3. Ethical and Accountability Challenges: Bias, Fairness, Trust, and Privacy

Beyond direct safety and security failures, the opacity of AI in critical infrastructure raises significant ethical and accountability concerns:

- **Bias and Fairness:** AI models learn from data, and if that data reflects historical biases or is unrepresentative, the AI system can perpetuate or even amplify unfairness and discrimination.[9] In CI, this could manifest as biased allocation of healthcare resources disadvantaging certain demographic groups [38], discriminatory credit scoring in financial services [6], or potentially inequitable distribution of energy or water resources. The black box nature makes it difficult to audit models for such biases or understand how they influence outcomes.[1]
- **Erosion of Trust:** As repeatedly noted, the inability to understand how an AI system arrives at a decision fundamentally undermines trust among users, operators, regulators, and the public.[4] This is particularly problematic in CI where reliability and public confidence are essential. Lack of trust can lead to underutilization of beneficial AI tools or, conversely, dangerous over-reliance if explanations create a false sense of security.[11]
- **Accountability Gap:** When an opaque AI system fails, determining causality and assigning responsibility becomes extremely difficult.[6] Was the failure due to flawed design, biased data, an unforeseen environmental factor, operator error, or a malicious attack? Without transparency, investigations are hampered, potentially leaving victims without recourse and preventing effective learning from failures. This ambiguity challenges existing legal and ethical frameworks built on clear lines of responsibility.[79] Human operators may find themselves in a "moral crumple zone," unfairly blamed for failures of autonomous systems they could neither fully understand nor control.[80]

 doi:10.20944/preprints202511.2276.v1

- **Privacy Concerns:** Training effective AI models often requires vast amounts of data, which in CI sectors like healthcare, finance, and potentially energy or water usage, can include sensitive personal information.[11] The collection, storage, and processing of this data raise significant privacy risks, including potential breaches, misuse, or re-identification, even from model outputs via inference attacks.[44] Ensuring compliance with privacy regulations like HIPAA or GDPR while leveraging data for AI development is a major challenge. [6]

Addressing these ethical challenges requires not only technical solutions like bias detection and mitigation techniques or privacy-preserving ML, but also robust governance frameworks, clear ethical guidelines, and ongoing stakeholder dialogue.

4.4. The Time-Criticality Dilemma: High-Risk Decisions with Limited Human Intervention

The risks associated with black box AI are significantly amplified in applications involving time-critical decision-making where human oversight or intervention is impractical or impossible due to speed constraints.[12] Many emerging AI applications in CI fall into this category:

- **Autonomous Systems:** Collision avoidance maneuvers in autonomous vehicles must occur in fractions of a second.[25]
- **Power Grid Control:** Maintaining grid stability (e.g., frequency and voltage control) requires responses much faster than human operators can typically provide, especially during cascading failures. [17]
- **Cyber Defense:** Automated systems may need to respond instantly to detected cyber threats to prevent intrusion or damage.[81]
- **Financial Trading:** High-frequency trading algorithms operate at microsecond timescales. [6]
- **Potential Future Nuclear Safety:** While currently avoided [20], future AI-driven safety systems might need to react instantly to anomalies.[67]

In these scenarios, the opacity of AI models presents a profound dilemma. If the AI makes an incorrect decision, there is no time for a human to review the reasoning, question the output, or take corrective action.[21] Operators may be forced to trust the AI's judgment implicitly, even if it feels counterintuitive. This can lead to "automation bias" or complacency, where humans overly rely on the automated system and fail to monitor its performance adequately or intervene when necessary, potentially accepting incorrect AI suggestions.[11]

Furthermore, the inability to understand *why* an AI might fail in a specific split-second scenario makes it incredibly difficult to design robust safety mechanisms or predict behavior under novel, high-stress conditions.[16] Traditional testing and validation methods struggle to cover the vast state space of possible real-time interactions, meaning performance guarantees are often lacking for unforeseen edge cases.[16] The combination of high stakes (potential for immediate catastrophic failure), decision speed precluding human review, and the inherent unpredictability and opacity of complex AI models makes time-critical control one of the most dangerous applications of black box AI in critical infrastructure. This necessitates extremely high levels of assurance or potentially imposes fundamental limits on the safe deployment of such systems.

4.5. Illustrative Cases: Documented Incidents and Plausible Scenarios

Abstract discussions of risk become more concrete when examined through the lens of real-world incidents and plausible scenarios. While attributing failures solely to AI opacity can be complex, several cases highlight the types of problems that can arise from automation complexity, data issues, or unexpected AI behavior in critical systems:

- **Documented Incidents & Failures:**
    1. *Autonomous Vehicle Accidents:* Several crashes involving vehicles with advanced driver-assistance systems (often marketed with terms implying autonomy, like Tesla's Autopilot) have raised questions about system limitations, sensor failures, unexpected behavior, and

 doi:10.20944/preprints202511.2276.v1

the interaction between the AI and the human driver (automation bias).[26] The fatal Uber self-driving test vehicle crash in Tempe, Arizona (2018) highlighted failures in perception and decision-making software under specific conditions.[80]

2. *Cyberattacks on Critical Infrastructure:* The 2015 cyberattack on the Ukrainian power grid, using malware like BlackEnergy, demonstrated the potential for remote disruption of critical control systems, leaving hundreds of thousands without power.[70] While not directly an AI failure, it shows the vulnerability of automated CI control. The 2017 TRITON/TRISIS malware specifically targeted industrial safety systems, aiming to disable safety functions while causing physical disruption.[82] Tampering with a Kansas water treatment system in 2019 by a former employee highlights the risks of unauthorized remote access to control systems.[34]
3. *AI Bias and Harm:* Facial recognition systems have led to wrongful arrests due to higher error rates for certain demographics. [26] Chatbots deployed for sensitive tasks, like the National Eating Disorders Association's Tessa chatbot, have given harmful advice.[83] AI translation errors have led to misunderstandings with serious consequences, such as an arrest based on a mistranslated social media post.[43]
4. *Financial System Disruptions:* Algorithmic trading has been implicated in market "flash crashes," where automated systems reacted rapidly and unexpectedly to market conditions, causing severe volatility.[43]

- **Analyzed Hypothetical Scenarios:** Risk assessments often utilize plausible scenarios to explore potential failure modes:
  1. *Adversarial Attacks:* Scenarios involve attackers using manipulated inputs (e.g., stickers on signs, altered sensor data) to cause widespread AV malfunction[27], or poisoning data to compromise AI-based grid control or cybersecurity defenses. [15] DHS and CISA conduct exercises exploring AI-powered cyberattacks on CI.[84]
  2. *Cascading Failures:* Scenarios model how an initial AI-related failure in one sector (e.g., a power grid control error due to flawed AI prediction or undetected anomaly) could propagate through interconnected systems (transportation, communications, healthcare, water) leading to widespread disruption.[14]
  3. *Nuclear Safety Scenarios:* While AI is not currently used for direct safety control, hypothetical scenarios explore risks if opaque AI were integrated into monitoring or response systems, potentially leading to delayed or incorrect actions due to misinterpretation of complex sensor data or unforeseen failure modes under stress.[67]
  4. *Healthcare Resource Allocation:* Scenarios examine how biased AI algorithms used for allocating scarce resources (like ventilators or vaccines during a pandemic) could lead to systematically inequitable outcomes if not carefully designed and audited.[37]

These examples, both real and hypothetical, underscore the tangible nature of the risks associated with deploying complex, often opaque, AI systems in critical infrastructure. They highlight the convergence of safety and security concerns, the potential for cascading impacts due to interconnectivity, and the unique challenges posed by time-critical autonomous decision-making. The difficulty in fully understanding or predicting the behavior of these systems, especially under novel or adversarial conditions, remains a central vulnerability. A detailed examination of the risks associated with the AI 'black box' problem in critical infrastructure is presented in Table 3. This table outlines the various risk categories, their potential impacts, and the elements that can intensify these risks within critical infrastructure contexts.

 doi:10.20944/preprints202511.2276.v1

Table 3. Typology of Risks Associated with Black Box AI in Critical Infrastructure

Risk Category	Description	Specific Manifestations in CI	Potential Consequences	Amplifying Factors
Safety Failures	Unintended system behavior leading to harm or damage.	Misdiagnosis / treatment errors (Healthcare) [39]; AV collisions [26] ; Power grid instability / outages [17] ; Water contamination/system failure [34]; Incorrect nuclear plant monitoring/response (hypothetical). [21]	Loss of life, Injury, Environmental damage, Property damage, Service disruption.[3]	Time-criticality, Complexity of interaction with environment, Lack of robustness/brittleness [71], Inadequate V&V.[16]
Security Vulnerabilities	Susceptibility to malicious manipulation or compromise.	Adversarial attacks (evasion, poisoning) on sensors/control systems [77]; Data breaches via model inversion [44]; AI supply chain attacks [5]; Use of AI by attackers .[8]	Sabotage, Espionage, Service denial, Data theft, Physical damage triggered by cyber means.[34]	Opacity hindering attack detection, Data dependency (poisoning target), Interconnectivity (attack propagation), Use of open-source models/data. [85]
Ethical Concerns	Violations of fairness, privacy, or societal values.	Biased resource allocation (Healthcare, Finance, potentially Energy/Water) [38]; Discriminatory outcomes [79]; Privacy violations from data collection/use [40]; Erosion of public trust.[9]	Discrimination, Inequality, Loss of autonomy, Public backlash, Reduced adoption of beneficial tech.[79]	Opacity hiding biases, Dependence on large, potentially sensitive datasets, Lack of clear ethical guidelines/audits, Algorithmic complexity.
Operational Disruptions	Failures impacting the reliable functioning of CI services.	Unexpected system downtime [86] ; Cascading failures across sectors.[2] ; Reduced efficiency due to model errors or instability [12]; Difficulty in diagnosing/repairing AI-related faults.[10]	Economic losses, Service unavailability (power, water, transport, healthcare), Public inconvenience, Loss of productivity. [14]	Interconnectivity, Time-criticality, Opacity hindering diagnostics, Update opacity [12], Lack of skilled personnel.
Accountability Gaps	Inability to determine cause or assign responsibility for failures.	Difficulty in post-incident analysis due to opacity [9]; Ambiguity in liability (developer vs. operator vs. user) [80]; "Moral crumple zone" effect on human operators.[73]	Lack of legal recourse for victims, Hindered learning from failures, Erosion of trust in governance/oversight.[10]	Opacity, Complexity of AI systems and CI interactions, Lack of standardized logging/auditing for AI decisions, Distributed responsibility across lifecycle actors.[80]

This typology underscores the multifaceted nature of risks emanating from opaque AI in critical systems. Addressing these requires a holistic approach encompassing technical robustness, security hardening, ethical design principles, rigorous operational procedures, and clear governance structures, as discussed in the following section. The amplification of these risks by factors like time-criticality and interconnectivity highlights the unique challenges posed by AI in the CI context.

5. Conclusions and Recommendations

The integration of Artificial Intelligence into critical infrastructure presents a paradigm shift, offering transformative potential for efficiency, reliability, and resilience. However, this potential is shadowed by the significant challenge posed by the "black box" nature of many advanced AI systems. This report has synthesized evidence from scientific literature, official reports, and incident analyses

 doi:10.20944/preprints202511.2276.v1

to illuminate the pervasive risks associated with deploying opaque AI decision-making systems in sectors vital to national security, economic stability, and public well-being.

5.1. Synthesis of Core Findings: The Pervasive Challenge of Opaque AI in Critical Systems

The core findings of this analysis converge on several key points. Firstly, the adoption of AI in CI is accelerating across sectors like energy, transportation, healthcare, water management, and finance, driven by demonstrable benefits in optimization, prediction, and automation.[5] Secondly, the most powerful AI models driving these advancements, particularly deep neural networks, often suffer from a lack of transparency – the black box problem – making their internal decision-making processes difficult or impossible for humans to fully comprehend.[9]

Thirdly, this opacity introduces substantial risks that are particularly acute in the context of CI due to the potential for catastrophic consequences. These risks span safety (unpredictable failures leading to harm or loss of life), security (novel vulnerabilities to adversarial manipulation and AI-driven attacks), ethics (bias, fairness, privacy violations), and accountability (difficulty in assigning responsibility for failures).[3] Fourthly, these risks are significantly amplified in interconnected CI systems where failures can cascade rapidly, and especially in time-critical applications where the speed of AI decision-making precludes meaningful human oversight or intervention.[2] The convergence of safety and security threats, where attacks can cause safety failures and system fragility can enable attacks, further complicates the risk landscape.

Finally, while Explainable AI (XAI) offers valuable tools for providing partial insights into AI behavior, current methods (like LIME and SHAP) have documented limitations regarding stability, fidelity, computational cost, and robustness.[51] They are not, in their current state, a panacea capable of guaranteeing the safety and trustworthiness required for the most demanding, high-consequence CI applications, particularly those operating under severe time constraints. This necessitates a broader perspective on mitigation that encompasses technology, process, and governance.

5.2. Pathways to Mitigation: Technical Solutions and Best Practices

Addressing the complex risks of opaque AI in CI requires a multi-layered strategy combining advancements in AI technology with rigorous engineering practices. No single solution is sufficient; rather, a portfolio of approaches is needed:

- **Technical Solutions:**
    1. *Advanced and Robust XAI:* Research should concentrate on creating XAI techniques which provide both stability through consistent explanations for similar inputs and faithfulness by accurately showing AI model reasoning. The system needs to operate at high computational efficiency to deliver real-time or near-real-time explanations because this capability is essential for time-sensitive critical infrastructure scenarios. These techniques need to withstand manipulation and small input changes because adversarial attacks should not be able to exploit vulnerabilities in the explanation process.[51] The evaluation of explanation quality and reliability requires systematic methods that go beyond explanation generation to ensure explanatory validity. The development of standardized evaluation metrics and benchmarks for XAI assessment needs to happen to measure both explanation accuracy and understandability and utility for various stakeholder groups.[87] The field needs to understand and address Explainability Pitfalls (EPs) because these pitfalls create misleading explanations that give users false security. Research into human operator explanation perception and validation needs active investigation because it affects explanation validation. The use of rule-based explanations combined with visualization tools for AI decision-making and feature attribution techniques presents potential strategies. The dynamic nature of critical infrastructure requires essential online or adaptive XAI methods which can adapt to AI model updates.

 doi:10.20944/preprints202511.2276.v1

2. *Robust Machine Learning:* Emphasis should shift towards designing ML models that are inherently more robust from the outset.[16] This includes techniques like adversarial training (exposing models to adversarial examples during training), robust optimization, certified defenses, data augmentation to cover edge cases, and rigorous uncertainty quantification to understand when a model's prediction is unreliable.[88] Building resilience against data distribution shifts and noisy inputs is critical for real-world CI environments.[89]
3. *Formal Verification and Synthesis:* For components where safety is paramount and behavior can be mathematically specified, formal methods offer the potential for provable guarantees of correctness.[74] Techniques like model checking, theorem proving, and reachability analysis can verify properties like robustness or adherence to safety rules.[90] While scalability remains a major challenge for complex DNNs [90], research into abstraction techniques and verification of specific properties (rather than full functional correctness) shows promise.[74] Formal inductive synthesis aims to generate components that are correct by construction.[90]
4. *Safe Reinforcement Learning (Safe RL):* For AI systems learning control policies through interaction (common in robotics, grid control, autonomous systems), Safe RL techniques aim to ensure that safety constraints are satisfied throughout the learning process and during deployment.[91] Methods include incorporating safety constraints into the optimization objective (e.g., via constrained MDPs), using safety layers or shields to filter unsafe actions, and employing Lyapunov functions or control barrier functions to guarantee stability and constraint satisfaction.[92]
5. *AI Assurance and V&V:* Comprehensive Verification and Validation (V&V) frameworks specifically tailored for AI systems are essential.[93] This involves integrating various techniques—rigorous testing (including adversarial testing and stress testing [94]), simulation, performance monitoring, formal methods where applicable, and potentially the use of assurance cases to structure safety arguments.[93] Frameworks like the NIST AI RMF provide a structure for managing these activities throughout the AI lifecycle.[63] Continuous monitoring and validation post-deployment are crucial for adaptive AI systems.[95]

- **Best Practices:**
  1. *Human-Centric Design:* Maintain human oversight and control wherever feasible, especially for critical decisions.[96] Design interfaces and processes that support effective human-AI teaming and mitigate automation bias.[67]
  2. *Rigorous Testing and Evaluation (T&E):* Implement comprehensive T&E protocols that go beyond standard accuracy metrics to assess robustness, safety, fairness, and security under realistic and stressful conditions.[16]
  3. *Data Governance:* Ensure high-quality, representative, and secure training data. Implement processes for detecting and mitigating bias in datasets.[94]
  4. *Secure Development Lifecycle:* Integrate security considerations throughout the AI development process, from design to deployment and maintenance.[85]
  5. *Incident Reporting and Learning:* Establish mechanisms for reporting AI incidents, failures, and near-misses, and foster a culture of learning from these events to improve future systems.[43]
  6. *Transparency and Documentation:* Maintain clear documentation regarding AI system design, training data, limitations, intended use, and performance evaluations.[97]

Implementing these technical solutions and best practices requires significant investment, expertise, and a cultural shift towards prioritizing safety and trustworthiness alongside performance.

5.3. Addressing Accountability and Liability in AI-Driven Critical Infrastructure

The increasing deployment of AI systems within critical infrastructure necessitates a thorough examination of accountability and liability frameworks. The "black box" nature of complex AI models poses unique challenges to traditional notions of responsibility, as it can be difficult to discern how specific decisions are made and who should be held accountable when failures occur. This subsection

 doi:10.20944/preprints202511.2276.v1

addresses the crucial need for clear lines of responsibility, mechanisms for redress, and proactive strategies for managing liability in the context of AI-driven critical systems.

It's important to note that efforts to regulate these issues have already begun. For instance, the European Union is developing the **AI Liability Directive**, which aims to establish clear rules regarding liability for damage caused by AI systems. This directive seeks to ensure fair compensation for victims and incentivize the development of safe and reliable AI technologies.

Key Considerations:

- **Defining Roles and Responsibilities:** Establish clear roles and responsibilities for all stakeholders involved in the lifecycle of AI systems in CI, including developers, operators, and regulators. Determine who is accountable for the design, implementation, maintenance, and deployment of these systems.
- **Legal and Regulatory Frameworks:** Adapt existing legal and regulatory frameworks to address the specific challenges of AI accountability, considering international initiatives and directives such as the AI Liability Directive. This may include developing new regulations or clarifying existing laws regarding liability for AI-related failures.
- **Auditing and Transparency:** Implement robust auditing mechanisms to track AI decision-making processes where feasible. Promote transparency in AI systems to facilitate investigation and accountability in the event of incidents.
- **Redress and Compensation:** Establish mechanisms for redress and compensation for individuals or entities harmed by AI-driven failures in CI. This may involve creating dedicated funds or insurance schemes.
- **Risk Management and Mitigation:** Encourage proactive risk management strategies to identify and mitigate potential liability issues associated with AI deployment. This includes conducting thorough testing, implementing safety measures, and developing contingency plans.
- **Human Oversight and Intervention:** Emphasize the importance of human oversight and intervention in critical AI decision-making processes. Clearly define the roles and responsibilities of human operators and ensure they have the authority to override AI decisions when necessary.

Recommendations:

1. Develop clear legal and ethical guidelines for AI accountability in critical infrastructure sectors, clarifying liability for AI-related incidents, and taking into account international directives such as the AI Liability Directive.
2. Establish mandatory reporting mechanisms for AI incidents and near-misses to facilitate investigations and identify patterns of failure.
3. Promote research and development of explainable AI (XAI) techniques to enhance transparency and accountability.
4. Require comprehensive documentation and auditing of AI systems, including training data, algorithms, and decision-making processes.
5. Explore insurance mechanisms or liability funds to compensate victims of AI-related failures in CI.
6. Invest in training and education programs for CI operators and regulators to ensure they understand the complexities of AI systems and their implications for accountability.

By proactively addressing accountability and liability, stakeholders can build trust in AI-driven critical infrastructure and ensure that these systems are deployed safely and responsibly.

5.4. Governing AI in Critical Infrastructure: Regulatory Landscape and Policy Recommendations

Effective governance is crucial for managing the risks of AI in CI. The regulatory landscape is evolving rapidly, with a mix of general frameworks and emerging sector-specific approaches.

Regulatory Landscape Overview:

 doi:10.20944/preprints202511.2276.v1

- **General Frameworks:** The EU AI Act adopts a risk-based approach, classifying AI systems based on potential harm and imposing stricter requirements (including transparency and human oversight) for high-risk applications, many of which fall within CI sectors.[6] The US has pursued a combination of executive orders (e.g., EO 14110 on Safe, Secure, and Trustworthy AI) directing agencies to develop guidelines and standards, alongside voluntary frameworks like the NIST AI RMF.[98]
- **Sector-Specific Approaches:**
  1. *Nuclear Energy:* Characterized by intense regulatory scrutiny and international collaboration (IAEA, NRC, ONR, CNSC). [19] Focus is on adapting existing rigorous safety standards, developing specific guiding principles for AI (considering failure consequence, autonomy, human factors, safety cases), and currently limiting AI use primarily to non-safety applications.[66]
  2. *Autonomous Vehicles:* NHTSA provides voluntary guidance (e.g., ADS: A Vision for Safety) and oversees testing, while states enact varying legislation covering testing, deployment, liability, and operational rules.[99] Standards bodies like SAE and ISO define automation levels and develop technical standards.[100]
  3. *Medical Devices:* The FDA regulates AI-enabled medical devices (SaMD) using existing pathways (510(k), De Novo, PMA) but is developing AI-specific guidance on aspects like predetermined change control plans (PCCPs) for adaptive algorithms and lifecycle management.[101] Emphasis is on safety, effectiveness, and managing modifications.
  4. *Cybersecurity:* General CI cybersecurity regulations (like CIRCIA for incident reporting [102]) apply, but specific regulations targeting AI security vulnerabilities are less mature, often relying on guidance from agencies like CISA and NIST.[69]

Table 4. Overview of Relevant Regulatory Frameworks and Guidelines for AI in Critical Infrastructure

Framework / Body	Scope / Applicability	Key Tenets related to AI Risk / Safety / Transparency	Approach to Black Box Problem	Status
EU AI Act [103]	General (Cross-Sector)	Risk-based tiers (unacceptable, high, limited, minimal); Strict requirements for high-risk systems (data quality, documentation, transparency, human oversight, robustness, accuracy, cybersecurity).	Mandates transparency & explainability for high-risk systems; Requires human oversight capabilities.	Enacted (phased implementation).
US AI Executive Order (14110) & Agency Actions [98]	General (US Federal Gov & influenced sectors)	Directs agencies (NIST, HHS, DOT, DOE, DHS etc.) to develop standards, guidelines, and tools for safe, secure, trustworthy AI; Emphasizes safety, security, privacy, equity, civil rights.	Promotes transparency; Directs NIST to develop explainability guidelines; Agencies developing sector-specific approaches.	Executive Order issued; Agency actions ongoing.
NIST AI RMF 1.0 [63]	General (Cross-Sector)	Voluntary framework; Govern, Map, Measure, Manage functions; Defines Trustworthy AI characteristics (validity, reliability, safety, security, fairness, transparency, explainability, privacy).	Explicitly includes Explainability & Interpretability as trustworthiness characteristics; Provides structure for assessing and managing risks related to opacity.	Published (Jan 2023); Voluntary.

 doi:10.20944/preprints202511.2276.v1

Nuclear Regulators (IAEA, NRC, ONR, CNSC) [20]	Sector-Specific (Nuclear Energy)	Extreme focus on safety & security; Adaptation of existing nuclear standards; Guiding principles (failure consequence, autonomy, human factors, safety cases); International collaboration.	Cautious approach; Transparency and V&V paramount for safety-critical use (currently limited); Requires robust safety cases addressing AI uncertainties.	Ongoing development of guidelines, strategic plans, collaborative initiatives (RegLab).
Automotive (NHTSA, SAE, ISO) [99]	Sector-Specific (Transportation/AVs)	NHTSA: Voluntary guidance (ADS 2.0/AV Policy), safety assessment elements, state guidance, incident reporting (SGO). SAE/ISO: Automation levels (J3016), technical standards for ADAS/ADS functions, safety (ISO 26262, SOTIF/ISO 21448).	Focus on functional safety, operational design domains (ODD), human-machine interface (HMI), cybersecurity; Less explicit focus on algorithmic transparency itself, more on system-level safety assurance.	NHTSA guidance issued; State laws vary; SAE/ISO standards evolving.
Medical Devices (FDA, EMA) [101]	Sector-Specific (Healthcare/Medical Devices)	Regulation as medical devices (SaMD); Risk-based classification; Requirements for safety & effectiveness; Guidance on PCCPs for adaptive AI/ML; Lifecycle management considerations.	Requires sufficient transparency for validation & clinical use; Focus on performance validation, bias assessment, and managing changes in adaptive models.	Existing device regulations apply; AI-specific guidance developing (drafts & final versions issued).
Cybersecurity (CISA, CIRCIA) [69]	Cross-Cutting (CI Cybersecurity)	CIRCIA mandates reporting of significant cyber incidents for CI entities; CISA provides guidance, threat intelligence, promotes secure-by-design (including for AI).	Focus on securing AI systems from attack & preventing malicious use of AI; Transparency needed for incident analysis & defense, but less regulated from algorithmic perspective.	CIRCIA reporting rules proposed/finalizing; CISA guidance ongoing.

Policy Recommendations: Based on the identified risks and the limitations of current technical solutions, particularly the inadequacy of XAI for guaranteeing safety in high-risk, time-critical scenarios, the following policy recommendations are proposed:

- **Mandate Rigorous V&V for High-Risk CI AI:** Require stringent V&V processes, potentially incorporating formal methods where feasible, for AI systems deployed in safety-critical CI functions. This goes beyond standard testing to provide higher assurance levels.
- **Develop Enforceable Sector-Specific AI Safety Standards:** Build upon the cautious, principle-based approach of the nuclear sector to develop mandatory, sector-specific standards for AI safety, tailored to the unique risks and operational contexts of energy, transportation, healthcare, etc.
- **Establish Clear Limitations for Opaque, Autonomous AI in Time-Critical, High-Consequence Roles:** Explicitly define boundaries for the deployment of fully autonomous AI systems whose decisions cannot be adequately verified or explained *before* potentially catastrophic failure occurs, and where human intervention is impossible within the necessary timeframe. This may necessitate prohibiting such systems in specific applications (e.g., primary nuclear safety control, certain autonomous weapon functions) until assurance methods mature significantly. This acknowledges the current limitations of XAI and verification techniques.[16]
- **Strengthen and Standardize AI Incident Reporting:** Enhance mandatory incident reporting requirements (building on CIRCIA and FDA/NHTSA models) to specifically include failures and near-misses related to AI components, facilitating cross-sector learning and proactive risk identification.[102]
- **Increase Investment in Trustworthy AI R&D:** Fund research focused on improving the robustness, explainability, verifiability, and safety of AI systems specifically for CI applications.[5]

 doi:10.20944/preprints202511.2276.v1

- **Promote Adoption of AI Assurance Frameworks:** Encourage or mandate the use of comprehensive risk management frameworks like the NIST AI RMF by CI operators developing or deploying AI systems.[63]

These recommendations aim to foster responsible innovation by ensuring that safety, security, and trustworthiness keep pace with AI capabilities, particularly where critical infrastructure and human lives are at stake. The current reactive approach to regulation needs to shift towards proactivity, informed by the inherent risks of opacity and the lessons learned from high-assurance sectors.

5.5. Future Outlook: Advancing Research in Trustworthy and Safe AI for Critical Systems

Significant research challenges remain in ensuring the safe and trustworthy deployment of AI in critical infrastructure. Future efforts should prioritize several key areas:

- **Improving XAI Methods:** Developing XAI techniques that are not only more accurate and stable but also scalable to large, complex models and computationally feasible for real-time applications is crucial.[104] Research is needed on methods to robustly evaluate the faithfulness and utility of explanations.[87]
- **Practical Formal Verification:** Bridging the gap between theoretical formal verification techniques and practical application to real-world, large-scale AI systems, particularly DNNs, is a major hurdle.[74] Research into scalable abstraction, compositional verification, and verifying specific critical properties (rather than full functional correctness) is needed.[105]
- **Guaranteed Safe RL:** Enhancing the theoretical guarantees and practical robustness of Safe RL algorithms is essential for their deployment in physical control systems within CI.[91] Methods need to handle complex state/action spaces and provide strong assurances against constraint violations during both training and execution.
- **Standardized Metrics and Benchmarks:** Developing widely accepted metrics and benchmarks for evaluating AI safety, robustness, fairness, and explainability is critical for comparing different approaches and setting clear performance targets.[106]
- **Understanding Long-Term Behavior:** Research is needed to understand how AI systems adapt and evolve over time in dynamic CI environments, including the implications of continuous learning, model drift, and update opacity.[12]
- **Human Factors and Interaction:** Deeper investigation into human interaction with complex and potentially opaque AI systems is required, focusing on trust calibration, mitigating automation bias, designing effective interfaces for shared control, and training operators for AI-augmented roles.[67]

Addressing these research challenges is fundamental to unlocking the full potential of AI in critical infrastructure while managing its inherent risks. A concerted, multidisciplinary effort involving AI researchers, domain experts, safety engineers, ethicists, and policymakers is required to build a future where AI enhances, rather than compromises, the resilience and safety of our most vital systems. The path forward demands not only technical innovation but also a commitment to rigorous validation, transparent governance, and a cautious approach proportionate to the potential consequences of failure.

References

1. Deck, L.; Schoeffer, J.; De-Arteaga, M.; Kühll, N. A Critical Survey on Fairness Benefits of Explainable AI. Available online: https://facctconference.org/static/papers24/facct24-105.pdf.
2. Hunnewell, B. National Security Concerns for Artificial Intelligence and Civilian Critical Infrastructure. Available online: https://muse.jhu.edu/article/950955.
3. Laplante, P.; Milojicic, D.; Serebryakov, S.; Bennet, D. Artificial Intelligence and Critical Systems: From Hype to Reality. Available online: https://www.osti.gov/servlets/purl/1713282.
4. Morandini, S.; Fraboni, F.; Balatti, E.; Hackmann, A.; Brendel, H.; Puzzo, G.; Volpi, L.; et al. Assessing the Transparency and Explainability of AI Algorithms in Planning and Scheduling Tools: A Review of the

 doi:10.20944/preprints202511.2276.v1

Literature. Available online: https://openaccess-api.cms-conferences.org/articles/download/978-1-958651-87-2_65.
5. of Energy (DOE), U.D. DOE Delivers Initial Risk Assessment on Artificial Intelligence for Critical Energy Infrastructure. Available online: https://www.energy.gov/ceser/articles/doe-delivers-initial-risk-assessment-artificial-intelligence-critical-energy.
6. Akther, A.; Arobee, A.; Adnan, A.A.; Auyon, O.; Islam, A.J.; Akter, F. Blockchain as a Platform for Artificial Intelligence (AI) Transparency. Available online: https://www.arxiv.org/pdf/2503.08699.
7. Wastewater, W.I. The Potential of Artificial Intelligence in Water Quality Monitoring: Revolutionizing Environmental Protection. Available online: https://www.waterandwastewater.com/the-potential-of-artificial-intelligence-in-water-quality-monitoring/.
8. Resecurity. Cyber Threats Against Energy Sector Surge as Global Tensions Mount. Available online: https://www.resecurity.com/blog/article/cyber-threats-against-energy-sector-surge-global-tensions-mount.
9. Lin, B.; Bilal, A.; Ebert, D. LLMs for Explainable AI: A Comprehensive Survey. Available online: https://arxiv.org/html/2504.00125v1.
10. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Available online: https://www.researchgate.net/publication/373382966_Interpreting_Black-Box_Models_A_Review_on_Explainable_Artificial_Intelligence.
11. Rosenberger, J.; Kuhlemann, S.; Tiefenbeck, V.; Kraus, M.; Zschech, P. The Impact of Transparency in AI Systems on Users' Data-Sharing Intentions: A Scenario-Based Experiment. Available online: https://arxiv.org/html/2502.20243v1.
12. Hatherley, J. A Moving Target in AI-assisted Decision-making: Dataset Shift, Model Updating, and the Problem of Update Opacity. Available online: https://arxiv.org/html/2504.05210v1.
13. Yang, G.; Ye, Q.; Xia, J. Unbox the Black-box for the Medical Explainable AI via Multi-modal and Multi-centre Data Fusion: A Mini-review, Two Showcases and Beyond. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC8459787/.
14. Baker, G.H.; Volandt, S. Cascading Consequences: Electrical Grid Critical Infrastructure Vulnerability. Available online: https://www.domesticpreparedness.com/articles/cascading-consequences-electrical-grid-critical-infrastructure-vulnerability.
15. Ribeiro, A. US DOE Rolls Out Initial Assessment Report on AI Benefits and Risks for Critical Energy Infrastructure. Available online: https://industrialcyber.co/ai/us-doe-rolls-out-initial-assessment-report-on-ai-benefits-and-risks-for-critical-energy-infrastructure/.
16. Porawagamage, G.; Dharmapala, K.; Chaves, J.S.; Villegas, D.; Rajapakse, A. A Review of Machine Learning Applications in Power System Protection and Emergency Control: Opportunities, Challenges, and Future Directions. Available online: https://www.frontiersin.org/journals/smart-grids/articles/10.3389/frsgr.2024.1371153/full.
17. Zhai, Z.M.; Moradi, M.; Lai, Y.C. Detecting Attacks and Estimating States of Power Grids from Partial Observations with Machine Learning. Available online: https://link.aps.org/doi/10.1103/PRXEnergy.4.013003.
18. Yurman, D. Is the Nuclear Energy Industry Ready for Artificial Intelligence? Available online: https://energycentral.com/c/ec/nuclear-energy-industry-ready-artificial-intelligence.
19. (NRC), U.N.R.C. How The NRC Is Preparing To Review AI Technologies. Available online: https://www.nrc.gov/ai/externally-focused.html.
20. Muhlheim, M. D. Available online: https://info.ornl.gov/sites/publications/Files/Pub201873.pdf.
21. U.S. Nuclear Regulatory Commission (NRC). Available online: https://www.nrc.gov/docs/ML2424/ML24241A252.pdf.
22. Luzniak, K. AI in Transportation Industry: Key Use Cases & the Future of Mobility. Available online: https://neoteric.eu/blog/use-of-ai-in-transportation-industry/.
23. Sameer, S. AI in Transportation: Use Cases, Advantages, and Potential Challenges. Available online: https://www.apptunix.com/blog/ai-in-transportation/.
24. Singh, A. Top 10 Applications of AI in Transportation and Logistics. Available online: https://www.appventurez.com/blog/applications-of-ai-in-transportation-and-logistics.
25. NVIDIA. NVIDIA Autonomous Vehicles Safety Report. Available online: https://images.nvidia.com/aem-dam/en-zz/Solutions/auto-self-driving-safety-report.pdf.
26. Arnold, Z.; Toner, H. AI Accidents: An Emerging Threat. Available online: https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Accidents-An-Emerging-Threat.pdf.

 doi:10.20944/preprints202511.2276.v1

27. Rawat, D.B. Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open Topics, Blockchain, and Future Directions. Available online: https://www.mdpi.com/2624-800X/3/3/25.
28. Vincent, O. Kevin. Available online: https://secureenergy.org/wp-content/uploads/2021/05/Kevin-Vincent-Regulatory-Framework.pdf.
29. Center for Sustainable Systems, U.o.M. Autonomous Vehicles Factsheet. Available online: https://css.umich.edu/publications/factsheets/mobility/autonomous-vehicles-factsheet.
30. Idrica. Five Key Areas in Which Artificial Intelligence Is Set to Transform Water Management in 2025. Available online: https://www.idrica.com/blog/artificial-intelligence-is-set-to-transform-water-management/.
31. DigitalDefynd. 10 Ways AI Is Being Used in Water Resource Management [2025]. Available online: https://digitaldefynd.com/IQ/ai-use-in-water-resource-management/.
32. Numalis. AI Innovations in Water, Sewerage, and Waste Management. Available online: https://numalis.com/ai-in-water-sewerage-and-waste-management/.
33. Hyer, C. Intelligent Water, Improved Systems: The AI Blueprint. Available online: https://www.arcadis.com/en/insights/blog/global/celine-hyer/2024/intelligent-water-improved-systems-the-ai-blueprint.
34. Yigit, Y.; Ferrag, M.A.; Ghanem, M.C.; Sarker, I.H.; Maglaras, L.A.; Chrysoulas, C.; Moradpoor, N.; Tihanyi, N.; Janicke, H. Generative AI and LLMs for Critical Infrastructure Protection: Evaluation Benchmarks, Agentic AI, Challenges, and Opportunities. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC11944634/.
35. OpenMedScience. Artificial Intelligence in Healthcare: Revolutionising Diagnosis and Treatment. Available online: https://openmedscience.com/artificial-intelligence-in-healthcare-revolutionising-diagnosis-and-treatment/.
36. Shuliak, M. AI in Healthcare: Examples, Use Cases & Benefits [2025 Guide]. Available online: https://acropolium.com/blog/ai-in-healthcare-examples-use-cases-and-benefits/.
37. Wu, H.; Lu, X.; Wang, H. The Application of Artificial Intelligence in Health Care Resource Allocation Before and During the COVID-19 Pandemic: Scoping Review. Available online: https://ai.jmir.org/2023/1/e38397.
38. Olawade.; B., D.; Wada, O.J.; David-Olawade, A.C.; Kunonga, E.; Abaire, O.; Ling, J. Using Artificial Intelligence to Improve Public Health: A Narrative Review. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC10637620/.
39. Daley, S. AI in Healthcare: Uses, Examples & Benefits. Available online: https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare.
40. Dunn, P. Leveraging AI To Improve Public Health. Available online: https://statetechmagazine.com/article/2025/02/leveraging-ai-improve-public-health.
41. Meade.; E., H.; Dillon, L.; Palmer, S. AI in Health Care: A Regulatory and Legislative Outlook. Available online: https://www.ey.com/content/dam/ey-unified-site/ey-com/en-us/campaigns/health/documents/ey-ai-in-healthcare.pdf.
42. Rubegni, E.; Ayoub, O.; Rizzo, S.M.R.; Barbero, M.; Bernegger, G.; Faraci, F.; Mangili, F.; et al. Designing for Complementarity: A Conceptual Framework to Go Beyond the Current Paradigm of Using XAI in Healthcare. Available online: https://arxiv.org/html/2404.04638v1.
43. Mcgregor, S. When AI Systems Fail: Introducing the AI Incident Database. Available online: https://partnershiponai.org/aiincidentdatabase/.
44. Othman, A. Ensuring the Safety and Security of AI Systems in Critical Infrastructure and Decision-Making. Available online: https://www.researchgate.net/publication/389988661_Ensuring_the_Safety_and_Security_of_AI_Systems_in_Critical_Infrastructure_and_Decision-Making.
45. Rjoub, G.; Bentahar, J.; Wahab, O.A.; Mizouni, R.; Song, A.; Cohen, R.; Otrok, H.; Mourad, A. A Survey on Explainable Artificial Intelligence for Cybersecurity. Available online: https://arxiv.org/pdf/2303.12942.
46. Li, Y.; Zhang, S.; Li, Y. Adversarial Deep Learning for Robust Short-Term Voltage Stability Assessment under Cyber-Attacks. Available online: https://arxiv.org/pdf/2504.02859.
47. Capuano, N.; Fenza, G.; Loia, V.; Stanzione, C. Explainable Artificial Intelligence in CyberSecurity: A Survey. Available online: https://www.researchgate.net/publication/363314499_Explainable_Artificial_Intelligence_in_Cybersecurity_A_Survey.
48. DARPA. Explainable Artificial Intelligence. Available online: https://www.darpa.mil/research/programs/explainable-artificial-intelligence.
49. Capuano, N. Explainable Artificial Intelligence in CyberSecurity: A Survey. Available online: https://www.capuano.cloud/papers/IEEE_Access_2022.pdf.

 doi:10.20944/preprints202511.2276.v1

50. Clement, T.; Kemmerzell, N.; Abdelaal, M.; Amberg, M. XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process. Available online: https://www.mdpi.com/2504-4990/5/1/6.
51. Chung.; Christopher, N.; Chung, H.; Lee, H.; Brocki, L.; Chung, H.; Dyer, G. False Sense of Security in Explainable Artificial Intelligence (XAI). Available online: https://arxiv.org/html/2405.03820v2.
52. Bhagavatula, A.; Ghela, S.; Tripathy, B. Demystifying the Black Box-Unveiling the Decision-Making Process of AI Systems. Available online: https://www.researchgate.net/publication/382317071_Demystifying_the_Black_Box-Unveiling_the_Decision-Making_Process_of_AI_Systems.
53. Knab, P.; Marton, S.; Schlegel, U.; Bartelt, C. Which LIME Should I Trust? Concepts, Challenges, and Solutions. Available online: https://arxiv.org/html/2503.24365v1.
54. Pradhan, R.; Lahiri, A.; Galhotra, S.; Salimi, B. Explainable AI: Foundations, Applications, Opportunities for Data Management Research. Available online: https://romilapradhan.github.io/assets/pdf/xai-sigmod.pdf.
55. Marshall, K. Are There Any Limitations to Using LIME? Available online: https://www.deepchecks.com/question/are-there-any-limitations-to-using-lime/.
56. Hsieh, W.; Bi, Z.; Jiang, C.; Liu, J.; Peng, B.; Zhang, S.; Pan, X.; et al. A Comprehensive Guide to Explainable AI: From Classical Models to LLMs. Available online: https://arxiv.org/pdf/2412.00800.
57. Salih, A.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Menegaz, G.; Lekadir, K. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. Available online: https://arxiv.org/html/2305.02012v3.
58. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. Available online: https://proceedings.neurips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
59. Wikipedia. Explainable artificial intelligence. Available online: https://en.wikipedia.org/wiki/Explainable_artificial_intelligence.
60. Zheng, H.; Pamuksuz, U. SCENE: Evaluating Explainable AI Techniques Using Soft Counterfactuals. Available online: https://arxiv.org/pdf/2408.04575.
61. Arrighi, L.; de Moraes, I.A.; Zullich, M.; Simonato, M.; Barbin, D.F.; Junior, S.B. Explainable Artificial Intelligence Techniques for Interpretation of Food Datasets: A Review. Available online: https://arxiv.org/pdf/2504.10527.
62. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of Explainable AI Techniques in Healthcare. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC9862413/.
63. Villanueva, E. Safeguard the Future of AI: The Core Functions of the NIST AI RMF. Available online: https://www.auditboard.com/blog/nist-ai-rmf/.
64. Baig, A.; Malik, O.I. NIST AI Risk Management Framework Explained. Available online: https://securiti.ai/nist-ai-risk-management-framework/.
65. (NEA), O.N.E.A. Regulating AI use during SMR deployment. Available online: https://www.oecd-nea.org/jcms/pl_101324/regulating-ai-use-during-smr-deployment.
66. Picot, W. Enhancing Nuclear Power Production with Artificial Intelligence. Available online: https://www.iaea.org/bulletin/enhancing-nuclear-power-production-with-artificial-intelligence.
67. of the National Cyber Director (ONCD), O. New Paper Shares International Principles for Regulating AI in the Nuclear Sector. Available online: https://www.onr.org.uk/news/all-news/2024/09/new-paper-shares-international-principles-for-regulating-ai-in-the-nuclear-sector/.
68. Becker, J. An Overview of Taxonomy, Legislation, Regulations, and Standards for Automated Mobility. Available online: https://www.apex.ai/post/legislation-standards-taxonomy-overview.
69. Sledjeski, C. Principles For Reducing AI Cyber Risk In Critical Infrastructure: A Prioritization Approach. Available online: https://www.mitre.org/sites/default/files/2023-10/PR-23-3086%20Principles-for%20Reducing-AI-Cyber-Risk-in-Critical-Infrastructure.pdf.
70. Collins, S.; Eckert, G.; Hauer, V.; Hubmann, C.; Pilon, J.; Poulton, G.; Polke-Markmann, H. Cyber Attacks on Critical Infrastructure. Available online: https://commercial.allianz.com/news-and-insights/expert-risk-articles/cyber-attacks-on-critical-infrastructure.html.
71. of Standards, N.I.; (NIST), T. Combinatorial Methods for Trust and Assurance. Available online: https://csrc.nist.gov/projects/automated-combinatorial-testing-for-software/autonomous-systems-assurance/explainable-ai.
72. Bharadwaj, C. AI in Transportation: Benefits, Use Cases, and Examples. Available online: https://appinventiv.com/blog/ai-in-transportation/.

 doi:10.20944/preprints202511.2276.v1

73. Macrae, C. Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety and Sociotechnical Sources of Risk. Available online: https://www.researchgate.net/publication/352307955_Learning_from_the_Failure_of_Autonomous_and_Intelligent_Systems_Accidents_Safety_and_Sociotechnical_Sources_of_Risk.
74. Neider, D.; Roy, R. What is Formal Verification without Specifications? A Survey on Mining LTL Specifications. Available online: https://arxiv.org/pdf/2501.16274.
75. Comiter, M. Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It. Available online: https://www.belfercenter.org/publication/AttackingAI.
76. Rosenberg, I.; Shabtai, A.; Elovici, Y.; Rokach, L. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. Available online: https://arxiv.org/pdf/2007.02407.
77. Walton, R. Artificial Intelligence Can Help Manage the Grid but Creates Risks if Deployed 'Naïvely,' DOE Warns. Available online: https://www.utilitydive.com/news/artificial-intelligence-AI-manage-electric-grid-risks-doe/714663/.
78. Ivezic, M.; Ivezi, L. Adversarial Attacks: The Hidden Risk in AI Security. Available online: https://securing.ai/ai-security/adversarial-attacks-ai/.
79. Networks, P.A. AI Risk Management Framework. Available online: https://www.paloaltonetworks.com/cyberpedia/ai-risk-management-framework.
80. Ryan, P.; Porter, Z.; Al-Qaddoumi, J.; McDermid, J.; Habli, I. What's My Role? Modelling Responsibility for AI-based Safety-critical Systems. Available online: https://arxiv.org/html/2401.09459v1.
81. Govea, J.; Gaibor-Naranjo, W.; Villegas-Ch, W. Transforming Cybersecurity into Critical Energy Infrastructure: A Study on the Effectiveness of Artificial Intelligence. Available online: https://www.mdpi.com/2079-8954/12/5/165.
82. Weinberg, A. Analysis of Top 11 Cyber Attacks on Critical Infrastructure. Available online: https://www.firstpoint-mg.com/blog/analysis-of-top-11-cyber-attackson-critical-infrastructure/.
83. Marco, D.P. AI Incidents: A Rising Tide of Trouble. Available online: https://www.ewsolutions.com/ai-incidents-a-rising-tide-of-trouble/.
84. Cybersecurity.; (CISA), I.S.A. 2024 Year in Review. Available online: https://www.cisa.gov/about/2024YIR.
85. Gilkarov, D.; Dubin, R. Zero-Trust Artificial Intelligence Model Security Based on Moving Target Defense and Content Disarm and Reconstruction. Available online: https://arxiv.org/pdf/2503.01758.
86. Almasoudi, F.M. Enhancing Power Grid Resilience through Real-Time Fault Detection and Remediation Using Advanced Hybrid Machine Learning Models. Available online: https://www.mdpi.com/2071-1050/15/10/8348.
87. Chen, J.; Storchan, V. Seven Challenges for Harmonizing Explainability Requirements. Available online: https://smallake.kr/wp-content/uploads/2022/10/2108.05390.pdf.
88. Tehranipoor, M.; Farahmandi, F. Guide #1 AI for Microelectronics Security: National Security Imperatives for the Digital Age. Available online: https://ai.ufl.edu/media/aiufledu/resources/AI-Policy-Guide_One.pdf.
89. Mohseni, S.; Wang, H.; Yu, Z.; Xiao, C.; Wang, Z.; Yadawa, J. Taxonomy of Machine Learning Safety: A Survey and Primer. Available online: https://arxiv.org/pdf/2106.04823.
90. Seshia.; A., S.; Sadigh, D.; Sastry, S.S. Towards Verified Artificial Intelligence. Available online: https://arxiv.org/pdf/1606.08514.
91. Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter, F.; Wang, J.; Knoll, A. A Review of Safe Reinforcement Learning: Methods, Theories and Applications. Available online: https://kclpure.kcl.ac.uk/portal/files/300373453/A_Review_of_Safe_Reinforcement_Learning_Methods_Theories_and_Applications_2_.pdf.
92. Eckel, D.; Zhang, B.; Bödecker, J. Revisiting Safe Exploration in Safe Reinforcement Learning. Available online: https://arxiv.org/html/2409.01245v1.
93. Correa-Jullian, C.; Grimstad, J.N.; Dugan, S. The Safety Case for Autonomous Systems: An Overview. Available online: https://www.researchgate.net/publication/388779912_The_Safety_Case_for_Autonomous_Systems_An_Overview.
94. McGrath, A.; Jonker, A. What Is AI Safety? Available online: https://www.ibm.com/think/topics/ai-safety.
95. Laplante, P.; Kuhn, R. AI Assurance for the Public — Trust but Verify, Continuously. Available online: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=935075.
96. Kuzhanthaivel, A. AI Needs Human Oversight to Safeguard Critical Infrastructure Against New Cyber Threats. Available online: https://www.itnews.asia/news/ai-needs-human-oversight-to-safeguard-critical-infrastructure-against-new-cyber-threats-615635.

 doi:10.20944/preprints202511.2276.v1

97. Food, U.; (FDA), D.A. Artificial Intelligence and Machine Learning in Software as a Medical Device. Available online: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.
98. Anderson., J., A.; C, P. Morgan, Amanda K. Available online: https://crsreports.congress.gov/product/pdf/R/R48319.
99. (NHTSA), N.H.T.S.A. Automated Vehicle Safety. Available online: https://www.nhtsa.gov/vehicle-safety/automated-vehicles-safety.
100. (NHTSA), N.H.T.S.A. Automated Driving Systems. Available online: https://www.nhtsa.gov/vehicle-manufacturers/automated-driving-systems.
101. U.S. Food and Drug Administration (FDA). Available online: https://www.fda.gov/media/184856/download.
102. Register, F. Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) Reporting Requirements. Available online: https://www.federalregister.gov/documents/2024/04/04/2024-06526/cyber-incident-reporting-for-critical-infrastructure-act-circia-reporting-requirements.
103. Parliament, E.; of the Council. "AI Act". Available online: http://data.europa.eu/eli/reg/2024/1689/oj.
104. Yang, W.; Wei, Y.; (repeated), H.W.; Chen, Y. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. Available online: https://www.researchgate.net/publication/373066914_Survey_on_Explainable_AI_From_Approaches_Limitations_and_Applications_Aspects.
105. Pullum, L. Verification and Validation of Systems in Which AI is a Key Element. Available online: https://sebokwiki.org/wiki/Verification_and_Validation_of_Systems_in_Which_AI_is_a_Key_Element.
106. Mohale, V.Z.; Obagbuwa, I.C. A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems to Enhancing Transparency and Interpretability in Cybersecurity. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC11877648/.