Cyber Security and Resilience of Artificial Intelligence Systems in the Nuclear Sector: A Literature Review

Tadej Šeruga, Igor Bernik
University of Maribor
Faculty of Criminal Justice and Security
Kotnikova 8, SI-1000 Ljubljana, Slovenia
tadej.seruga@student.um.si,igor.bernik@um.si

Abstract*. This literature review examines the cyber security and resilience of artificial intelligence (AI) systems in the nuclear sector. Based on a structured review of publications from Scopus and Web of Science, the paper analyses how recent literature addresses AI systems as cyber-vulnerable and resilience-critical assets. The findings show that AI is often discussed as a tool for cyber security, while fewer studies examine the security of AI systems themselves. Key concerns include adversarial attacks, data manipulation, model robustness, verification and validation, explainability, secure lifecycle management, and nuclear-specific deployment constraints.*

Keywords. cyber security, resilience, artificial intelligence, machine learning, operational technology, nuclear sector

1 Introduction

The nuclear sector is one of the most safety- and security-sensitive areas of critical infrastructure. It includes nuclear power plants (NPPs), research reactors, fuel cycle facilities, radioactive waste management facilities, and other installations where nuclear or radioactive material is used, processed, stored, or controlled. Nuclear energy is increasingly discussed as a strategic low-carbon energy source that can contribute to energy security and decarbonisation objectives (Huang et al., 2023; Khadka et al., 2024).
At the same time, the safety-sensitive nature of nuclear activities means that failures, accidents, malicious acts, or cyber incidents may have consequences that extend beyond ordinary business disruption and include equipment damage, loss of safety functions, radiological concerns, environmental impact, injury or death, loss of sensitive information, and loss of public confidence (Khadka et al., 2024; Lezzi et al., 2025). Nuclear facilities are therefore operated under strict safety, security, and regulatory requirements.

Although cyber security in nuclear facilities includes the protection of digital assets, control networks, communication pathways, safety-related systems, and operational technology (OT) environments, it cannot be reduced to conventional information technology security.

Digital systems may support nuclear safety, nuclear security, physical protection, emergency preparedness, operational monitoring, and control functions. As a result, a cyber incident may affect not only the confidentiality, integrity, or availability of information, but also the behaviour of physical processes and the ability of the facility to maintain safe and secure operation (Ayodeji et al., 2023; Lezzi et al., 2025).

As the sector undergoes a gradual but significant digital transformation, it is also adopting more advanced digital and data-driven technologies such as artificial intelligence. In this paper, AI is used as a broad term covering data-driven and learning-based systems, including machine learning (ML) and deep learning. In the nuclear sector, AI is increasingly discussed in relation to monitoring, diagnostics, anomaly detection, predictive maintenance, digital twins, remote inspection, decision support, autonomous control, and cyber security (Arhouni et al., 2025; Cancila et al., 2024; Huang et al., 2023; Khadka et al., 2024; Mendoza et al., 2026; Park & Koo, 2026). Within cyber security, AI-based and machine-learning-based methods are being used for anomaly detection, intrusion detection, attack classification, threat hunting, vulnerability assessment, attack path analysis, risk assessment, and predictive intelligence (Blakely et al., 2025; Lezzi et al., 2025). AI for cyber security is therefore a growing research area in industrial control system (ICS) security and nuclear cyber security.

The nuclear sector adds several constraints that make this topic distinct from general AI cyber security or general ICS cyber security. Nuclear systems are characterised by strict safety requirements, regulatory oversight, long system lifecycles, limited tolerance for unverified changes, restricted access to operational data, and the need to distinguish between equipment faults, normal transients, and malicious cyber activity. These characteristics affect how AI systems can be

designed, tested, deployed, and trusted (Yockey et al., 2023).

Existing literature has already examined the use of AI-based methods for cyber security in nuclear power plants, particularly for threat detection, cyber risk assessment, and AI-based countermeasures (Lezzi et al., 2025). However, less attention has been given to the opposite perspective: the cyber security and resilience of AI systems themselves. This review therefore focuses on AI systems as cyber-vulnerable and resilience-critical assets in the nuclear sector.

2 Background

For the purpose of this review, the term AI system is used broadly and refers not only to a trained model, but also to the wider technical and organisational environment required for its development, operation and maintenance. This includes data, model architecture and parameters, software libraries, deployment platforms, interfaces with instrumentation and control (I&C) systems, monitoring mechanisms, update processes, and human interpretation of outputs. This broader view is important because cyber security risks may arise across the AI lifecycle, from data collection and model development to deployment, operation, maintenance, and retirement.

AI systems can support nuclear applications in different ways. Relevant applications include monitoring, diagnostics, predictive maintenance, digital twins, remote inspection, reactor design support, anomaly detection, cyber response, and operational decision support (Huang et al., 2023; Khadka et al., 2024; Mendoza et al., 2026). In some cases, AI may become part of operational or safety-relevant decision-making chains.

AI assurance in the nuclear sector is further complicated by data and explainability challenges. Many AI applications in nuclear reactor research rely on simulated data because real plant data are difficult to obtain. This may lead to problems such as data imbalance, limited representativeness, and uncertainty regarding generalisability to operational environments (Huang et al., 2023). At the same time, high-performing AI models may behave as black boxes, making them difficult to explain, validate, or justify in safety-sensitive contexts.

In the nuclear sector, cyber security is commonly addressed through structured computer security programmes subject to regulatory oversight. International guidance and national regulatory frameworks emphasise the need to identify, protect, and maintain computer-based systems and digital assets that support safety, security, and emergency preparedness functions. The International Atomic Energy Agency addresses computer security as part of nuclear security in its implementing guidance and provides more detailed technical guidance for nuclear facilities (IAEA, 2021a, 2021b).

A central concept in this context is the identification of cyber-relevant and critical (sensitive) digital assets. Critical digital assets may include digital components and systems used in instrumentation and control, reactor protection, engineered safety features, distributed control, feedwater control, turbine control, emergency diesel generator systems, physical protection, emergency preparedness, communication, monitoring, and network security (Eggers, 2021; IAEA, 2021b; Lezzi et al., 2025).

AI systems that support safety-relevant, security-relevant, emergency preparedness, operational monitoring, autonomous control, or cyber defence functions may therefore need to be assessed as cyber-relevant digital assets. Depending on their function and on the potential consequences of compromise, some AI systems may also need to be considered as critical digital assets. As AI adoption increases, governance, transparency, robustness, and cyber security of AI systems are also likely to become increasingly relevant to nuclear assurance and regulatory discussions.

Against this programme-based background, the growing use of AI introduces a new dimension to the existing challenges of cyber security in the nuclear sector. AI systems are not only tools that may support cyber security; they may also become cyber-vulnerable assets.

The cyber security of AI systems differs from conventional software cyber security in several ways. AI systems may be vulnerable to a set of threats such as manipulation of training data, poisoning attacks, evasion attacks, model manipulation, data drift, insecure model updates, trojan attacks and dependency vulnerabilities in AI software libraries (Khadka et al., 2024; Ndiaye et al., 2024; Yockey et al., 2023). In industrial control and nuclear environments, these threats are further complicated by physical process constraints. Manipulated data must often remain plausible within the expected range of sensor and actuator behaviour, and the consequences of incorrect model outputs may affect not only information assets but also physical process states.

Resilience is therefore a key concept in this review. In the context of AI systems in the nuclear sector, resilience refers to the ability of an AI-enabled function to maintain, recover, or support safe and secure operation under adverse cyber conditions. This includes the ability to detect manipulation, tolerate degraded or uncertain data, resist adversarial inputs, provide meaningful explanations, support operator decision-making, and avoid unsafe responses.

For these reasons, verification, validation, robustness testing, explainability, and lifecycle assurance are central to the cyber security and resilience of AI systems in the nuclear sector. AI components in nuclear I&C systems cannot be accepted only because they achieve high accuracy on a test dataset. They must also be evaluated for reliability, transparency, robustness, adversarial behaviour, lifecycle control, and compatibility with regulatory

expectations (Park & Koo, 2026). Similarly, secure lifecycle approaches for AI-enabled industrial and nuclear systems should include data and model management, risk assessment, adversarial testing, explainable AI (XAI), and resilience considerations (Ndiaye et al., 2024).

3 Methodology

This paper uses a structured literature review approach guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) to identify and analyse academic publications addressing the cyber security and resilience of AI systems in the nuclear sector. The review focuses on studies that discuss AI systems, adversarial threats, cyber security, resilience, verification and validation (V&V), explainability, or secure lifecycle concerns in nuclear, operational technology, I&C, ICS, or cyber-physical system (CPS) contexts.

Structured searches were conducted in Scopus and Web of Science on 6 June 2026. The same Boolean search expression was used in both databases and entered using the database-specific all-fields search syntax, namely ALL(...) in Scopus and an all-fields search in Web of Science. The search strategy was based on five conceptual blocks: operational technology and industrial control systems; cyber security and resilience; adversarial attacks and data manipulation; AI and machine learning; and the nuclear sector. Synonymous terms within each block were combined with OR, while the five blocks were combined with AND. The exact search string is provided in the footnote[1].

The review was guided by the following research questions:

- RQ1: What types of artificial intelligence systems are discussed in the context of cyber security and resilience in nuclear or nuclear-relevant contexts?
- RQ2: Which cyber threats, adversarial techniques, and attack surfaces affecting artificial intelligence systems are identified in the reviewed literature?
- RQ3: How does the reviewed literature address the resilience, robustness, verification, validation, explainability, and secure lifecycle of artificial intelligence systems in nuclear or nuclear-relevant environments?

Publications were included if they were published in English between 2021 and 2026, were available in full text, and addressed cyber security, resilience, adversarial threats, or secure lifecycle issues related to AI systems in the nuclear sector or in a closely related nuclear-relevant OT, I&C, ICS, or cyber-physical system context. After the Boolean search, records were filtered according to the formal eligibility limits of the review: publication year, language and document type. Eligible Scopus document types were articles, reviews and conference papers; eligible Web of Science document types were articles, proceedings papers and review articles. Additional contextual sources cited in the Introduction and Background were used to frame the review and were not counted as part of the final synthesis unless they met the full inclusion criteria.

The literature review was guided by PRISMA, which structures the review process into identification, screening, eligibility, and inclusion stages. An adapted PRISMA flow diagram was used to document the selection process, including the number of records identified, duplicates removed, records screened, full-text publications assessed, and publications included in the final review, as presented in Fig. 1.

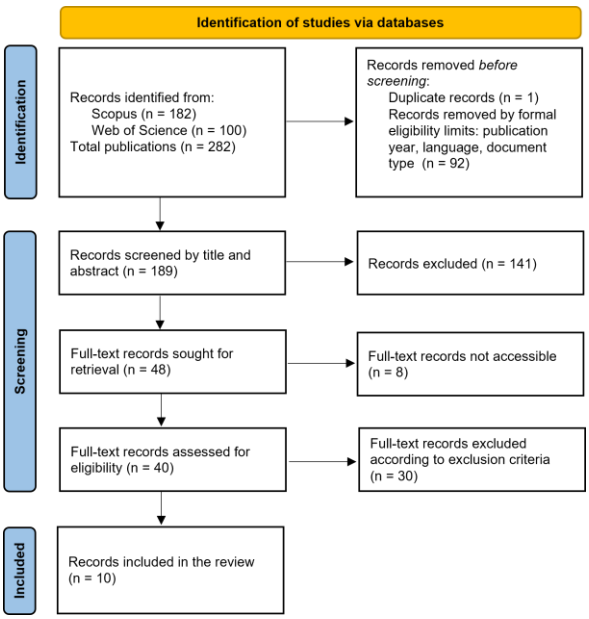

Figure 1. PRISMA flow diagram of the literature selection process

The initial search returned 182 records in Scopus and 100 records in Web of Science, resulting in 282 records in total. Before title and abstract screening, 92

[1]("OT" OR "operational technology" OR "I&C" OR "instrumentation and control" OR "instrumentation and control systems" OR "I&C systems" OR "industrial control system" OR "industrial control systems" OR "ICS" OR "SCADA" OR "distributed control system" OR "DCS" OR "cyber-physical" OR "cyber physical" OR "cyber-physical systems") AND ("cybersecurity" OR "cyber security" OR "computer security" OR "information security" OR "AI security" OR "ML security" OR "security testing" OR "adversarial testing" OR "secure lifecycle" OR "verification and validation" OR "V&V" OR "robustness" OR "resilience" OR "safety and security" OR "functional safety") AND ("adversarial attack" OR "adversarial attacks" OR "adversarial machine learning" OR "aML" OR "data poisoning" OR "poisoning attack" OR "feature poisoning" OR "evasion attack" OR "exploratory attack") AND ("AI" OR "artificial intelligence" OR "machine learning" OR "ML" OR "resilient machine learning" OR "rML" OR "AI model" OR "ML model" OR "large language model" OR "LLM") AND ("nuclear industry" OR "nuclear sector" OR "nuclear power plant" OR "NPP" OR "nuclear facility" OR "nuclear facilities" OR "nuclear reactor" OR "nuclear safety" OR "nuclear security")

records were removed by applying the formal eligibility limits of the review, namely publication year, language and document type. Duplicate checking was performed by comparing DOI where available, and by title, authors and publication year where DOI information was missing or incomplete. One duplicate record was identified. The low number of duplicates was manually validated using DOI and title-author-year checks and appears to reflect database-specific indexing and the narrow combination of AI, adversarial, OT/ICS and nuclear search terms. The remaining 189 records were screened by title and abstract, of which 141 were excluded because they did not address the combined focus of the review.

Title and abstract screening were used as the first relevance filter after the application of formal eligibility limits and duplicate removal. At this stage, publications were excluded when the title and abstract clearly showed that they did not address the combined focus of the review. This included records that focused on general cyber security without AI or ML, general AI without cyber security or resilience relevance, non-nuclear or non-nuclear-relevant domains, purely conventional IT security, general critical infrastructure without OT, I&C, ICS, CPS, or nuclear relevance, or topics unrelated to the research questions. When the title and abstract did not provide enough information to make a reliable exclusion decision, the publication was retained for full-text assessment.

The remaining 48 publications were selected for full-text retrieval. Eight publications were excluded because the full text was not accessible. The remaining 40 full-text publications were assessed for eligibility. After full-text assessment, 30 publications were excluded, resulting in ten publications included in the final synthesis.

Full-text publications were assessed using four linked relevance dimensions:

1. AI/ML relevance,
2. AI system cyber security and resilience relevance,
3. nuclear or nuclear-relevant context, and
4. contribution to the research questions.

AI/ML relevance required that a publication addressed an AI/ML model, AI component, AI-enabled function, autonomous or ML-enabled system, or AI-based detection, diagnosis, decision-support, or control function.

AI system cyber security and resilience relevance required that the cyber security or resilience issue was linked to the AI/ML system itself, to an AI component, or to an AI-enabled function. This included adversarial threats against AI models or data, manipulation of training or real-time process data used by AI systems, model tampering, evasion, poisoning, model robustness, explainability, V&V of AI components, secure AI lifecycle, AI security testing, deployment assurance of AI-enabled systems, or resilience mechanisms that protect or compensate for AI-enabled functions. General nuclear cyber security, cyber risk assessment, incident response, or ICS/SCADA security was not included unless it also provided evidence relevant to the assurance, robustness, validation, lifecycle security, or operational resilience of AI systems in nuclear or nuclear-relevant environments.

Nuclear relevance required either a direct nuclear-sector context, such as nuclear power plants, nuclear I&C systems, nuclear testbeds, reactor simulators, or nuclear cyber-physical systems, or a clearly nuclear-relevant OT, I&C, ICS, or CPS context with explicit transferability to nuclear AI assurance.

Contribution to the research questions required that the publication provided evidence relevant to at least one of the three research questions. The included publications were classified according to the contribution across the four relevance dimensions into one of three groups, namely as: core studies, conceptual core-supporting studies, or supporting studies. This classification was not used as a quality ranking, but as a relevance and contribution classification indicating how directly each publication supported the review focus and research questions.

Core studies were those that met all four dimensions at a high level and addressed AI/ML systems or AI components, cyber security or resilience of such systems, a direct nuclear-sector context, and a direct contribution to at least one research question. These studies treated AI systems, AI components, or ML-enabled nuclear applications as cyber-vulnerable, resilience-critical, or assurance-relevant assets.

Conceptual core-supporting studies had strong AI/ML and cyber security or resilience relevance, but their nuclear contribution was either conceptual, based on a nuclear-relevant use case, or transferable to nuclear AI assurance rather than a direct empirical assessment of AI systems in NPPs.

Supporting studies did not primarily examine the cyber security of AI systems themselves but were retained when they provided nuclear-specific evidence on AI-enabled detection, testbed-based validation, response, recovery, operational resilience, or deployment constraints needed to interpret how AI systems may be assured in nuclear cyber-physical environments, thereby contributing to at least one research question.

Publications were excluded from the final synthesis if they failed to meet the combined focus of the review. This included publications on general AI applications in nuclear engineering without a cyber security or resilience perspective; nuclear cyber security without AI or intelligent systems; general ICS, SCADA, CPS, or critical-infrastructure security without nuclear-specific contribution; AI-based intrusion detection or anomaly detection without implications for AI robustness, validation, explainability, lifecycle assurance, or nuclear deployment; general adversarial machine learning without nuclear or nuclear-relevant applicability; and AI governance, safety, or ethics

without both a cyber security and nuclear-sector connection.

The included publications were analysed using structured data extraction aligned with the three research questions. The extracted fields corresponded to the comparison dimensions in Table 1 and covered each publication's role, AI system or function, nuclear or nuclear-relevant context, threat or assurance focus, relevant attack surface or adversarial technique where applicable, and contribution to the research questions.

Because the final corpus included heterogeneous publication types, including surveys, conceptual frameworks, testbeds, simulation studies, and technical case studies, this assessment was used as a relevance and contribution appraisal rather than as a formal statistical quality or risk-of-bias assessment.

The relatively small final corpus reflects the narrow focus of the review rather than a broad lack of literature on either AI or nuclear cyber security. Many retrieved publications addressed AI for cyber security, general AI applications in the nuclear sector, general ICS or CPS cyber security, or AI assurance without a nuclear cyber security perspective. They were therefore excluded because this review specifically focuses on the cyber security and resilience of AI systems themselves in nuclear or nuclear-relevant environments.

4 Results

The final synthesis included ten publications: four core studies, two conceptual core-supporting studies, and four supporting studies. The included publications were not treated as a homogeneous body of evidence, but were compared according to their role, AI system or function, nuclear or nuclear-relevant context, threat or assurance focus, and contribution to the research questions. Table 1 shows three partially overlapping perspectives: core studies addressing AI systems as cyber-vulnerable or assurance-relevant assets; conceptual core-supporting studies providing transferable threat-modelling and trustworthiness concepts; and supporting studies showing how AI-enabled detection, response, testbeds, and operational resilience are constrained by nuclear-specific environments.

This distinction is important because only the core studies are used as direct evidence on the cyber security or assurance of AI systems themselves, whereas conceptual core-supporting and supporting studies are used more narrowly to interpret transferable threat concepts, deployment constraints, validation environments, response mechanisms, and nuclear-specific assurance needs.

Table 1. Summary of publications included in the final synthesis

Publication	Role	AI system / function	Context	Threat / assurance focus	Review contribution
Yockey et al. (2023)	C	ML digital twins, AutoML, ACS	Advanced reactor / NPP ACS	Data and model attack surfaces	Shows ACS ML components as cyber-relevant assets.
Khadka et al. (2024)	C	ML-enabled nuclear applications	Nuclear industry	Adversarial attacks, resilient ML	Maps attack surfaces, defences, barriers and opportunities for resilient ML.
Park and Koo (2026)	C	AI components in nuclear I&C systems	NPP I&C	V&V, robustness, transparency, software dependencies	Provides assurance considerations for AI components in nuclear I&C systems.
Ndiaye et al. (2024)	C	ML/AI fault detector, XAI	Industrial I&C with explicit nuclear relevance	Secure lifecycle, adversarial testing, XAI	Links adversarial testing and explainability to AI deployment in I&C systems.
Sundaram and Abdel-Khalik (2021)	CCS	Predictive model, covert cognizance	Nuclear reactor modelling	Trustworthiness, manipulation detection	Connects predictive model integrity with nuclear cyber-physical trustworthiness.
Al-Hawawreh et al. (2025)	CCS	ML-based IDS	Closed-loop CPS/ICS with NPP use case	Adversarial examples, evasion	Supports threat modelling for AI-enabled detection in nuclear-relevant CPS.
Yoo et al. (2025)	S	Autoencoder and RL	NPP under FDI attack	Detection, reconstruction, recovery	Shows AI-enabled resilience and recovery under false data injection.
Thiyagarajan and Hammad (2024)	S	Transformer-LSTM detector	Air-gapped NPP ICS	DoS, MITM, white-box detection	Supports deployment-related resilience in constrained nuclear ICS.
Zebrowitz et al. (2026)	S	ML cyber testbed	NPP network testbed	Generalisability, validation, XAI	Demonstrates the need for testbed validation and cross-environment evaluation.
Lee et al. (2021)	S	Hidden Markov model	NPP cyber response	Security state estimation, functional impact	Links cyber state estimation with response-oriented resilience.

Note. C = core study; CCS = conceptual core-supporting study; S = supporting study; ACS = autonomous control system; DoS = denial-of-service; FDI = false data injection; IDS = intrusion detection system; MITM = man-in-the-middle; NPP = nuclear power plant; RL = reinforcement learning; XAI = explainable AI.

4.1 Types of artificial intelligence systems addressed in the literature

With regard to RQ1, the reviewed publications show that AI systems in the nuclear cyber security context are discussed in several different forms. One direct example is the use of ML-based digital twins within autonomous control systems (ACS) for advanced reactors. In this setting, AI systems are not merely auxiliary analytical tools, but part of the control logic that supports malfunction detection, component state classification, forecasting, and control strategy selection (Yockey et al., 2023). This makes the AI component itself a cyber-relevant asset because manipulation of the data pipeline or model behaviour may affect the operation of the autonomous control system.

A second group of publications addresses AI components in nuclear I&C systems. These studies focus less on a specific attack scenario and more on whether AI components can be specified, verified, validated, tested, and accepted within the nuclear software assurance framework. This is important because AI components differ from traditional software due to their complexity, dependence on data, limited transparency, and less predictable behaviour (Park & Koo, 2026).

A third group focuses on AI-enabled cyber security functions, especially anomaly detection, intrusion detection, and cyber attack detection in nuclear or nuclear-relevant cyber-physical environments. Examples include transformer-LSTM models for Modbus TCP traffic in air-gapped nuclear ICS environments, ML-based intrusion detection system (IDS) models for closed-loop CPS environments, and machine-learning-based attack detection models evaluated across virtual and hybrid NPP network testbeds (Al-Hawawreh et al., 2025; Thiyagarajan & Hammad, 2024; Zebrowitz et al., 2026). These studies show how AI-enabled detection may be deployed in constrained nuclear or nuclear-relevant environments, even when their primary focus is AI for cyber security rather than the cyber security of AI itself.

A fourth group addresses intelligent and probabilistic support for cyber response and resilience. This includes self-healing control using autoencoders and reinforcement learning (RL), as well as security state estimation using hidden Markov models. These studies extend the role of AI-enabled and probabilistic methods from detection to response, recovery, and operational decision support (Lee et al., 2021; Yoo et al., 2025). They are therefore relevant to resilience, although they do not primarily analyse AI systems as attack targets.

4.2 Cyber threats, adversarial techniques and attack surfaces

With regard to RQ2, the reviewed literature shows that threats to AI systems in nuclear or nuclear-relevant environments should not be limited to attacks on the final model. The attack surface spans the whole AI lifecycle, including data collection, training and validation data, real-time process data, model artefacts, AI outputs, software dependencies, and the operational deployment environment.

The clearest attack surface is the data pipeline. Yockey et al. (2023) examine attacks against training data and real-time process data in ML-based autonomous control systems. This is particularly important because nuclear AI systems may rely on sensor data, process variables, historical operational data, and real-time communication between digital twins and control systems. In such settings, manipulation of data may be more feasible or more consequential than direct manipulation of the model architecture.

The second attack surface is the AI model and its software environment. This includes attacks against model architectures, model files, model parameters, model behaviour, software libraries, and dependencies. In their ACS case study, Yockey et al. (2023) assessed attacks against real-time data and training data as higher risk than direct attacks against ML model files or architectures, because data manipulation had more direct effects on model performance and ACS behaviour. Nevertheless, model-level attacks remain relevant because model substitution or manipulation could affect malfunction detection, forecasting, classification, or control-strategy selection.

The third attack surface is adversarial input manipulation. Several publications discuss adversarial examples, feature poisoning, evasion attacks, and false data injection (Al-Hawawreh et al., 2025; Khadka et al., 2024; Ndiaye et al., 2024; Yockey et al., 2023). In industrial and nuclear cyber-physical systems, such attacks differ from classical adversarial examples in image classification because the manipulated values must remain plausible within the physical process. Adversarial samples must therefore deceive the AI model while preserving the physical effect of the underlying attack.

The fourth attack surface is the interface between cyber events and physical process behaviour. False data injection (FDI), man-in-the-middle attacks (MITM), denial-of-service attacks, and Modbus function-code attacks are especially relevant in the reviewed literature (Thiyagarajan & Hammad, 2024; Yoo et al., 2025; Zebrowitz et al., 2026). These studies show that cyber attacks against AI-enabled systems in the nuclear sector cannot be understood only as information security events; they must also be evaluated according to their effects on physical process states, operator interpretation, and safety or security functions.

The fifth attack surface is deployment. In restricted or air-gapped nuclear ICS, concerns over third-party black-box AI tools favour white-box approaches, explainability, controlled deployment, and system-owner visibility into the model (Ndiaye et al., 2024; Park & Koo, 2026; Thiyagarajan & Hammad, 2024).

Table 2. Synthesis of attack surfaces and assurance mechanisms

Attack surface	Threat examples	Nuclear-specific concern	Assurance / resilience response
Training and validation data	Poisoning, label manipulation, biased or unrepresentative data	Real NPP data may be limited, sensitive, simulated or difficult to reproduce.	Data provenance, dataset validation, adversarial data testing, configuration control.
Real-time process data	FDI, sensor manipulation, MITM, replay	Manipulated values may remain physically plausible while misleading AI models or operators.	Process-aware validation, redundancy, plausibility checks, state estimation, testbed/HIL validation.
AI model and software artefacts	Model tampering, backdoors, model substitution, vulnerable libraries	AI components may support operationally significant or safety-relevant decisions.	Model integrity checks, dependency management, secure configuration, adversarial robustness testing.
AI outputs and decision support	False negatives, false positives, misleading confidence, opaque outputs	Incorrect or unexplained outputs may affect operator response and plant-state assessment.	XAI, confidence information, human oversight, alarm validation, independent confirmation.
Deployment and validation environment	Domain shift, cross-network degradation, black-box tools, testbed-to-plant transfer gaps	A model validated in one simulator, network or testbed may not generalise to another NPP environment.	Nuclear-specific testbeds, cross-environment evaluation, white-box deployment, lifecycle assurance.

Note. FDI = false data injection; HIL = hardware-in-the-loop; MITM = man-in-the-middle; XAI = explainable AI.

4.3 Resilience, robustness, verification, validation and secure lifecycle

With regard to RQ3, the reviewed literature addresses resilience and assurance through several mechanisms.

One prominent theme is adversarial robustness. Khadka et al. (2024) identify resilient machine learning in the nuclear sector as an emerging field and emphasise that both models and data must be tested against adversarial threats. Al-Hawawreh et al. (2025) similarly stress that ML-based IDS models in closed-loop CPS environments must be resilient to adversarial examples, especially where misclassification may affect physical processes.

A second theme is verification and validation. Park and Koo (2026) propose aligning AI model development with existing nuclear software V&V stages, while adding AI-specific verification activities. Their approach supports the view that model accuracy should be complemented by evidence on robustness, transparency, and development-process reliability.

A third theme is explainability. Ndiaye et al. (2024) discuss XAI in the context of adversarial testing and secure lifecycle considerations, while the FANCY-X testbed study uses explainability-oriented analysis to examine cross-environment degradation of ML-based attack detection models (Zebrowitz et al., 2026). Explainability is important because operators, assessors, and system owners need to understand why an AI model produced an alert and whether its behaviour remains meaningful when transferred across environments.

Sundaram and Abdel-Khalik (2021) add a complementary perspective by focusing on software cognizance and trustworthiness of predictive models. Their covert cognizance approach suggests that resilience may also depend on whether predictive software can maintain an internal record of its execution and detect manipulation without relying only on external logging, perimeter defences, or conventional intrusion detection mechanisms.

A fourth theme is resilience through response and recovery. Yoo et al. (2025) move beyond cyber attack detection by proposing a self-healing strategy using an autoencoder for detection and inferred value reconstruction, and reinforcement learning for system-level control. Lee et al. (2021) similarly support response by estimating the current security state and linking cyber alarm sequences to security and functional impact analysis. These studies suggest that future nuclear AI systems may need to support not only detection but also interpretation, response prioritisation, and recovery.

A fifth theme is nuclear-specific validation. Several studies rely on simulators, hardware-in-the-loop (HIL) testbeds, or nuclear cyber-physical testbeds rather than real NPP operational data. This is understandable because real NPP data are difficult to access, but it also limits generalisability. Studies using nuclear-specific environments, such as NPP network testbeds, nuclear safety shutdown testbeds, or nuclear I&C cyber-physical simulation environments, are particularly relevant because robust AI assurance requires validation environments that capture both cyber behaviour and physical process dynamics.

4.4 Cross-cutting observations

Across the ten included publications, three cross-cutting observations emerge. First, the literature increasingly recognises that AI systems in the nuclear sector create new attack surfaces. These include training datasets, real-time process data, model architectures, model files, AI libraries, deployment pipelines, and AI-enabled decision outputs.

Second, much of the literature still treats AI primarily as a cyber security tool rather than as a cyber-vulnerable asset. Anomaly detection, intrusion detection, and self-healing control are well represented, but fewer publications directly test the robustness of the AI system itself against adaptive adversaries. This distinction is important because an AI-based detector or controller that improves cyber security under normal assumptions may introduce additional risk if its own data, model, or outputs can be manipulated.

Third, nuclear-specific constraints substantially affect how AI cyber security and resilience should be evaluated. These constraints include safety relevance, digital I&C architecture, air-gapping, limited availability of representative datasets, regulatory expectations, V&V requirements, explainability, and the need to distinguish cyber attacks from equipment faults or normal transients. Therefore, cyber security and resilience of AI systems in the nuclear sector should not be evaluated only through conventional ML performance metrics. It should also consider lifecycle assurance, adversarial robustness, physical process validity, explainability, and the potential impact on nuclear safety and security functions.

5 Discussion

The reviewed literature shows that the cyber security and resilience of AI systems in the nuclear sector is an emerging but still fragmented research area. The selected studies approach the topic from different angles, including autonomous control, nuclear I&C assurance, resilient machine learning, adversarial testing, intrusion detection, security state estimation, and cyber-physical testbeds. This diversity reflects the early stage of the field, but it also shows that AI systems are becoming part of the digital ecosystem that must be considered in nuclear cyber security.

A central implication is the need to distinguish between AI used for cyber security and the cyber security of AI systems themselves. Several included studies use AI as a tool for detecting attacks, identifying anomalies, supporting response, or improving resilience. However, an AI system that supports cyber security is not necessarily secure, robust, explainable, or suitable for deployment in a nuclear environment. If an AI-based detector, digital twin, predictive model, or autonomous control component can be manipulated through training data, real-time data, model architecture, adversarial inputs, or deployment conditions, it may introduce additional risk instead of reducing it.

Building on the attack-surface synthesis presented in Table 2, the findings suggest that the cyber security and resilience of AI systems in the nuclear sector should be understood as an assurance problem spanning the AI lifecycle and the nuclear operational environment. As the main analytical contribution of this review, this paper proposes a synthesis-derived assurance agenda structured around five interrelated layers: data and lifecycle assurance, model robustness and integrity, OT and I&C deployment assurance, cyber-physical validation, and operational and regulatory assurance.

The first layer is data and lifecycle assurance. The reviewed literature shows that AI security risks are not limited to the trained model, but also include training data, validation data, real-time input data, data provenance, dataset representativeness, and lifecycle control. This is particularly important in the nuclear sector because real operational data may be limited, sensitive, simulated, or difficult to reproduce. Assurance activities should therefore address the integrity, traceability, representativeness, and protection of data across development, validation, deployment, operation, and maintenance.

The second layer is model robustness and integrity. AI components in nuclear I&C systems cannot be accepted only because they perform well on a test dataset. They must also be evaluated for adversarial robustness, model integrity, transparency, explainability, vulnerable software dependencies, and secure configuration management. This includes threats such as poisoning, evasion, adversarial perturbations, model tampering, backdoors, vulnerable open-source libraries, and model substitution. The implication is that AI models should be treated as cyber-relevant assets whose behaviour, parameters, dependencies, and updates require assurance.

The third layer is OT and I&C deployment assurance. Nuclear systems are characterised by restricted connectivity, long lifecycles, legacy systems, strict configuration control, limited tolerance for unverified change, and strong safety and security expectations. These characteristics may reduce some external attack paths, but they do not eliminate risks arising from portable media and mobile devices, internal networks, maintenance activities, supply chains, test environments, software dependencies, or poorly controlled updates. AI assurance must therefore be integrated with existing nuclear software assurance, cyber security, configuration management, and change-control practices.

The fourth layer is cyber-physical validation. Adversarial threats in nuclear cyber-physical systems differ from adversarial threats in conventional information technology because manipulated values must often remain physically plausible, respect process constraints, and preserve the intended physical effect of the underlying attack. Cyber security evaluation of AI systems in the nuclear sector must therefore account for process dynamics, sensor and actuator behaviour, control logic, safety functions, and the interaction between cyber events and physical plant states. Simulators, HIL environments, cyber-physical testbeds, and NPP network testbeds are therefore valuable and often necessary. However, they also create uncertainty regarding generalisability. The

FANCY-X study is particularly relevant in this respect, because it demonstrates that cyber attack detection performance may degrade when models are transferred between virtual and hybrid NPP network environments. This supports the need to evaluate AI cyber security methods not only for internal accuracy, but also for cross-environment robustness and deployment validity.

The fifth layer is operational and regulatory assurance. If AI is used to support detection, response, recovery, anomaly interpretation, operational decision-making, or autonomous control, then its reliability, failure modes, explainability, and potential for manipulation become relevant to operators, assessors, and regulators. AI outputs may affect operator understanding, response prioritisation, or plant-state assessment. Assurance should therefore include confidence information, XAI, human oversight, alarm validation, procedural integration, and evidence suitable for independent or regulatory review. In this sense, some AI systems may need to be assessed as cyber-relevant digital assets, and, depending on their function and potential consequences of compromise, some may also need to be treated as critical digital assets.

The proposed assurance agenda also clarifies the main gap in the current literature. There is comparatively more work on AI-enabled detection and response support than on the systematic protection of AI systems as cyber-vulnerable assets. Fewer studies directly evaluate poisoning, evasion, model manipulation, secure AI lifecycle, AI model supply chain, adaptive adversaries, or adversarial robustness in nuclear-specific settings. This indicates a gap between the expected future use of AI in advanced reactors, digital I&C, autonomous functions, and cyber-physical monitoring, and the maturity of methods for assuring the cyber security of such systems. Future studies should therefore place greater emphasis on reproducible nuclear-relevant testbeds, realistic process-aware attack scenarios, cross-environment validation, explainable and auditable AI, secure AI lifecycle management, and the relationship between AI model behaviour and nuclear safety and security functions.

These findings should also be interpreted in light of several limitations. The review is limited by its focus on Scopus and Web of Science, English-language publications, selected publication types, and the period from 2021 to 2026. As a result, relevant grey literature, regulatory guidance, technical reports, or earlier work may not have been included. The final corpus is also relatively small, because the review deliberately focused on the cyber security and resilience of AI systems in nuclear or nuclear-relevant environments rather than on the broader use of AI for cyber security. In addition, the classification of publications into core, conceptual core-supporting, and supporting studies involved interpretive judgement, particularly for borderline AI-enabled detection and response studies. The proposed assurance agenda should therefore be understood as an analytical structure derived from the reviewed literature, not as a validated maturity framework.

6 Conclusion

This literature review examined recent academic work on the cyber security and resilience of AI systems in the nuclear sector. Based on the final synthesis of ten publications, the review shows that the field is still emerging and fragmented, but increasingly relevant as AI is considered for cyber security, monitoring, decision support, digital twins, and control-related functions in nuclear environments. The main conclusion is that AI systems should not be treated only as analytical or cyber-defence tools. As they become part of nuclear digital systems, they may also become cyber-relevant assets with their own data, model, software, deployment, and operational attack surfaces.

The contribution of the review is twofold. First, it distinguishes the cyber security of AI systems from the broader literature on AI for cyber security in nuclear environments. Second, it synthesizes the reviewed literature into a nuclear-specific AI security assurance agenda structured around data and lifecycle assurance, model robustness and integrity, OT and I&C deployment assurance, cyber-physical validation, and operational and regulatory assurance. This synthesis indicates that conventional machine-learning performance metrics are not sufficient on their own. AI systems should also be assessed through assurance evidence related to integrity, robustness, explainability, deployment validity, human oversight, and regulatory review.

Important gaps remain in relation to poisoning, evasion, model manipulation, secure AI lifecycle management, AI model supply chains, adaptive adversaries, and validation in realistic nuclear I&C or cyber-physical environments. Future research should therefore focus on nuclear-specific adversarial testing, reproducible testbeds, cross-environment validation, explainable and auditable AI, and the relationship between AI model behaviour, cyber attack scenarios, and nuclear safety and security functions. As AI becomes more closely connected to nuclear digital systems, the cyber security and resilience of AI systems should become an explicit part of nuclear computer security programmes, V&V activities, and regulatory discussions.

References

Al-Hawawreh, M., Baig, Z., & Zeadally, S. (2025). Resilient Intrusion Detection Models for Closed Control-Loop in Cyber-Physical Systems: Combating Adversarial Examples. *IEEE Internet of*

Things Magazine, *8*(1), 73–80. https://doi.org/10.1109/IOTM.001.2400068

Arhouni, F. E., Abdo, M. A. S., Ouakkas, S., Bouhssa, M. L., & Boukhair, A. (2025). Artificial intelligence-driven advances in nuclear technology: Exploring innovations, applications, and future prospects. *Annals of Nuclear Energy*, *213*, 111151. https://doi.org/10.1016/j.anucene.2024.111151

Ayodeji, A., Mohamed, M., Li, L., Di Buono, A., Pierce, I., & Ahmed, H. (2023). Cyber security in the nuclear industry: A closer look at digital control systems, networks and human factors. *Progress in Nuclear Energy*, *161*, 104738. https://doi.org/10.1016/j.pnucene.2023.104738

Blakely, B., Li, Y., Dave, A., Kultgen, D., & Vilim, R. (2025). *AI-Driven Cybersecurity Testbed for Nuclear Infrastructure: Comprehensive Evaluation Using METL Operational Data* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2512.01727

Cancila, D., Daniel, G., Sirven, J.-B., Chihani, Z., Chersi, F., & Vinciguerra, R. (2024). Research Directions on AI and Nuclear. *EPJ Web of Conferences*, *302*, 17005. https://doi.org/10.1051/epjconf/202430217005

Eggers, S. (2021). A novel approach for analyzing the nuclear supply chain cyber-attack surface. *Nuclear Engineering and Technology*, *53*(3), 879–887. https://doi.org/10.1016/j.net.2020.08.021

Huang, Q., Peng, S., Deng, J., Zeng, H., Zhang, Z., Liu, Y., & Yuan, P. (2023). A review of the application of artificial intelligence to nuclear reactors: Where we are and what's next. *Heliyon*, *9*(3), e13883. https://doi.org/10.1016/j.heliyon.2023.e13883

IAEA. (2021a). Computer Security for Nuclear Security (IAEA Nuclear Security Series No. 42-G). International Atomic Energy Agency. https://www-pub.iaea.org/MTCD/Publications/PDF/PUB1918_web.pdf

IAEA. (2021b). Computer Security Techniques for Nuclear Facilities (IAEA Nuclear Security Series No. 17-T (Rev. 1)). International Atomic Energy Agency. https://www-pub.iaea.org/MTCD/Publications/PDF/PUB1921_web.pdf

Khadka, A., Sthapit, S., Epiphaniou, G., & Maple, C. (2024). Resilient Machine Learning: Advancement, Barriers, and Opportunities in the Nuclear Industry. *ACM Computing Surveys*, *56*(9), 1–29. https://doi.org/10.1145/3648608

Lee, C., Ho Chae, Y., & Hyun Seong, P. (2021). Development of a method for estimating security state: Supporting integrated response to cyber-attacks in NPPs. *Annals of Nuclear Energy*, *158*, 108287. https://doi.org/10.1016/j.anucene.2021.108287

Lezzi, M., Martino, L., Damiani, E., & Yeun, C. Y. (2025). A Systematic Literature Review on AI-Based Cybersecurity in Nuclear Power Plants. *Journal of Cybersecurity and Privacy*, *5*(4), 79. https://doi.org/10.3390/jcp5040079

Mendoza, H., Yadav, V., Alam, S. B., Eskins, D., & Iyengar, R. (2026). Advances in digital twins and AI/ML for condition monitoring in nuclear applications. *Frontiers in Energy Research*, *14*, 1716514. https://doi.org/10.3389/fenrg.2026.1716514

Ndiaye, N. G., Kirdan, E., & Waedt, K. (2024). *Cybersecurity Testing for Industry 4.0: Enhancing Deployments in operational I&C systems Through Adversarial Testing and Explainable AI*. https://doi.org/10.18420/INF2024_169

Park, J., & Koo, S. (2026). Reliability verification methods for artificial intelligence components in instrumentation and control systems at nuclear power plants. *Nuclear Engineering and Technology*, *58*(4), 104096. https://doi.org/10.1016/j.net.2025.104096

Sundaram, A., & Abdel-Khalik, H. (2021). Covert Cognizance: A Novel Predictive Modeling Paradigm. *Nuclear Technology*, *207*(8), 1163–1181. https://doi.org/10.1080/00295450.2020.1812349

Thiyagarajan, K., & Hammad, I. (2024). Anomaly Detection in Air-Gapped Industrial Control Systems of Nuclear Power Plants. *2024 Cyber Awareness and Research Symposium (CARS)*, 1–6. https://doi.org/10.1109/CARS61786.2024.10778886

Yockey, P., Erickson, A., & Spirito, C. (2023). Cyber threat assessment of machine learning driven autonomous control systems of nuclear power plants. *Progress in Nuclear Energy*, *166*, 104960. https://doi.org/10.1016/j.pnucene.2023.104960

Yoo, S., Mohler, G., & Zhang, F. (2025). Self-Healing Control of Nuclear Power Plants Under False Data Injection Attacks. *Nuclear Science and Engineering*, *199*(1), 162–175. https://doi.org/10.1080/00295639.2024.2372520

Zebrowitz, M., Doran, M. A., & Zhang, F. (2026). FANCY-X: A nuclear power plant network testbed for cross-network cybersecurity evaluation and generalizability. *Progress in Nuclear Energy*, *198*, 106444. https://doi.org/10.1016/j.pnucene.2026.106444