Sarcouncil Journal of Multidisciplinary

ISSN(Online): 2945-3445

Volume- 06| Issue- 04| 2026

Research Article **Received:** 01-03-2026 | **Accepted:** 25-03-2026 | **Published:** 12-04-2026

Resilience-by-Design in Machine Learning–Integrated Systems: Architectural Frameworks and Policy Implications

Evans Addo

Northeastern University - College of Engineering, Boston, MA, USA

Abstract: The fast adoption of machine learning (ML) into cyber-physical systems (CPS) is changing the operations of most critical infrastructure in the energy, utilities, and safety-critical industries. On one hand, AI-based predictive analytics and adaptive control can be used to improve efficiency and detect faults; however, it also introduces new weaknesses concerning the integrity of the data, the robustness of the models, adversarial manipulation, and the policy controls. The layered risks can be tackled only with the help of traditional cybersecurity methods. This review contributes to an ML-integrated CPS resilience-by-design framework based on the principles of resilience engineering and in line with the United States (U.S.) cybersecurity and AI governance frameworks. It combines ideas of adaptive systems, adversarial threat models, organizational preparedness variables, and policy frameworks to suggest a layered architectural model incorporating security, robustness, human supervision, and governance integration. The research establishes critical gaps in research and future perspectives on how to operationalize resilience metrics, adversarial stress testing, and maturity-based governance alignment. Resilience-by-design offers a disciplined journey towards achieving machine learning-affiliated critical infrastructure in a shifting threat environment by introducing engineering, policy, and socio-technical viewpoints into the solution.

Keywords: Resilience-by-Design, Cyber-Physical Systems (CPS), Machine Learning Integration, AI Risk Management, Fault-Tolerant Architecture, Adversarial AI.

INTRODUCTION

Information technology (IT) and operational technology (OT) have become closely linked to the digital transformation of the critical infrastructure, making it more efficient and more vulnerable. The recent case studies of Supervisory Control and Data Acquisition (SCADA) attacks in U.S. utilities demonstrate that the exposure of Human Machine Interfaces (HMIs) to the internet, weak passwords, and ineffective IT/OT segmentation are recurring examples of how attackers can disrupt operations (Panful *et al*., 2025a). This is evidenced by the fact that cyberattacks present an immediate threat to physical and safety.

Simultaneously, both artificial intelligence (AI) and machine learning (ML) are becoming increasingly implemented into safety-critical systems. Fault detection and diagnostics involving the use of AI are found in nuclear power plants to enhance anomaly identification and predictive maintenance (Bonsu & Adeoye, 2025). Equivalent AI-based anomaly detection models increase the accuracy of monitoring and responsiveness in compliance-driven settings (Gaye *et al*., 2025a). Yet the National Institute of Standards and Technology (NIST) AI Risk Management Framework also considers AI systems to be technologies that create decisions affecting real or virtual surroundings and emphasizes that their results may have a direct impact on physical processes (NIST, 2023). In the CPS deployment, model failures, drift, or adversarial manipulation may hence result in operational damage.

The conceptual basis of dealing with this complexity is offered by resilience engineering. The transition of Safety-I to Safety-II redefines safety as the ability to enable systems to change and maintain performance in different conditions instead of being able to avoid failure (Hollnagel *et al*., 2015). Woods also focuses on resilience as the ability to foresee, track, act, and learn (Woods, 2015). Resilience in cyber environments goes to the levels of preparation, absorption, recovery, and adaptation on both technical and organizational levels (Linkov & Kott, 2019).

However, the principles of resilience are not often implemented within the platform of ML-based CPSs. The AI systems can also be used as weapons, which reduces the obstacles to the automated cyberattack, as well as to the target infrastructures (Suleiman & Donkor, 2025). Organizational literature also demonstrates that formal compliance with regulations does not necessarily result in a long-term culture of cybersecurity, particularly in essential utilities (Panful *et al*., 2026; Panful *et al*., 2025b). These results point to the fact that resilience-by-design needs to combine architecture, governance, and human factors.

***Corresponding Author:** Evans Addo

DOI- https://doi.org/10.5281/zenodo.19545447

The U.S. policy frameworks offer systematic guidelines that are usually implemented separately. The NIST Cybersecurity Framework 2.0 presents cybersecurity outcomes based on governance (Pascoe, 2023), whereas the AI RMF presents governance, mapping, measurement, and management functions of trustworthy AI (NIST, 2023). Comparative maturity models indicate disproportionate practices in terms of implementing practices across the sectors (Rea-Guaman *et al*., 2017; Garba *et al*., 2020), and regulatory scholarship signals the fragmentation of AI oversight (Agbadamasi *et al*., 2025; Brakye & Adam, 2025).

This review summarizes the concepts of resilience engineering and discusses the ways in which the ML components can be systematically incorporated into safe, fault-tolerant CPS architectures. It also assesses the relevance of resilience-by-design to the U.S. cybersecurity and AI governance. This paper contributes to a formalized base of enhancing resilience in machine learning systems by providing a connection between system structure, predictive models, and regulatory measures.

FOUNDATIONS OF RESILIENCE ENGINEERING IN CYBER-PHYSICAL SYSTEMS

Resilience engineering emerged to address the shortcomings of the traditional safety and reliability models. The traditional methods tend to presuppose that the failure of systems is due to the components malfunctioning, and the mitigation of the safety can be enhanced via the removal of the failures. This perspective is also observed in the Safety-I perspective, in which safety is considered as the lack of adverse events (Hollnagel *et al*., 2015). But in more complicated and tightly coupled systems, this is not a sufficient assumption. Contemporary cyber-physical systems (CPS) are used in dynamic conditions in which variability is inevitable.

Safety-II reframes resilience as the capacity of systems to ensure that "as many things as possible go right" by supporting adaptive performance (Hollnagel *et al*., 2015). This view does not consider human operators as liabilities, but as necessary sources of flexibility and readjustment. Resilience in CPS environments, where both automation, sensors, and control loops interact in a continuous manner, relies upon the presence of an adaptive capacity on both technical and human levels.

Woods also conceptualizes resilience in four main capabilities, which are the power to anticipate, monitor, respond, and learn (Woods, 2015). Such capabilities explain the ability of systems to be functional despite uncertainty and disruption. Anticipation is the ability to identify possible threats in the future; monitoring helps in being aware of the current system states, response is timely, and learning helps improve the disturbances. These functions, when applied to CPS, go past cybersecurity detection to operational continuity and recovery.

Preparation, absorption, recovery, and adaptation are parts of system performance that are integrated into resilience in cyber contexts, in particular (Linkov & Kott, 2019). This model highlights the fact that resilience is not the control mechanism but a lifecycle property of the system. In the case of ML-integrated CPS, it implies that resilience needs to consider both cyber intrusions and algorithm failure, e.g., model drift, corrupted training data, or adversarial manipulation.

The U.S. utility SCADA incident history provides some empirical evidence of how the failure is often not due to a single vulnerability but rather caused by multiple layers of vulnerabilities, such as the exposed HMIs, insufficient separation between segments, and the weak ignorance of remote-access controls (Panful *et al*., 2025a). The existence of structured adversary models like Adversarial Tactics, Techniques and Common Knowledge (ATT&CK) on Industrial Control Systems (ICS) goes further to show that attackers operate in phase-based behavior to attack both the IT and OT infrastructure (Alexander *et al*., 2020). These trends prove that CPS resilience should not focus on digital or physical levels of interdependencies, but significant interdependence between them should be considered.

Human and organizational factors are also on the center stage. Energy-related studies indicate that the lack of awareness among the staff and vulnerability to phishing, as well as low compliance rates, constantly lead to the violations of operations (Panful *et al*., 2025b). In addition to that, the adherence to formal regulations does not necessarily result in an effective organizational culture (Panful *et al*., 2026). These results support the fact that resilience engineering in CPS should be based on the incorporation of governance, workforce behavior, and cultural maturity in addition to technical protective measures.

Publisher: SARC Publisher

Cybersecurity maturity checkups also show the inconsistent use of formalized risk management procedures in sectors (Rea-Guaman *et al*., 2017; Garba *et al*., 2020). Such fluctuation means that resilience cannot merely be presumed by the existence of frameworks. As an alternative, it must be intentionally incorporated into system architecture, work processes, and supervision systems.

Combined, these foundations form resilience in CPS as a multi-layered property, which is adaptive capacity, lifecycle management, adversarial awareness, and organizational readiness. In the case of machine learning-integrated systems, resilience engineering should hence go beyond the conventional fault tolerance concept to encompass algorithmic robustness, governance integration, and socio-technical coordination.

INTEGRATION OF MACHINE LEARNING INTO SECURE AND FAULT-TOLERANT CPS ARCHITECTURES

The implementation of machine learning (ML) in cyber-physical systems (CPS) is changing the process of monitoring, diagnostics, and control. AI-based fault detection and diagnostics (FDD) systems are being increasingly applied in safety-critical systems, including nuclear power plants, to detect pernicious anomalies, abnormal system behavior, and predict an impending failure (Bonsu & Adeoye, 2025). The systems are better at detecting quickly and being reliable in operation than the traditional rule-based or physics-based systems. Equally, AI-based self-healing control systems and real-time diagnostics improve the adaptive response of a system when it is operating in dynamic conditions (Adeoye *et al*., 2025).

AI-based models of anomaly detection, in regulatory and compliance settings, provide better detection, shorter response, and less human interaction in the monitoring process (Gaye *et al*., 2025a). These applications demonstrate the essence of the value proposition of ML in CPS: the possibility to process big amounts of sensor data, identify non-linear trends, and assist in making predictive decisions.

Nevertheless, the implementation of ML into the CPS architecture creates additional structural risks. The NIST AI Risk Management Framework underlines the fact that AI systems are socio-technical systems and the outputs of which affect real-world environments (NIST, 2023). As opposed to deterministic control algorithms, the ML models rely on training data, statistical generalization, and changing operational settings. This exposes them to model drift, adversarial input, data poisoning, and loss of interpretability. Such weaknesses can be propagated by other related subsystems in complex environments.

This concern is enhanced by threat modeling in ICS environments. ATT&CK framework of ICS records organized adversarial strategies that assault both the enterprise IT and operational control domains (Alexander *et al*., 2020). Additional case-based assessments of U.S. SCADA attacks indicate that attackers take advantage of improperly configured remote access, lax authentication, and IT/OT interdependencies to attain operational effect (Panful *et al*., 2025a). In these architectures, the predictive analytics and automated decision systems will be compromised by an upstream data pipeline or control interface compromise when incorporated by ML components.

The threat space may also be increased by AI systems themselves. The surveys of the AI-supported cyber threats outline the growing application of automated attack generation, adaptive phishing, and intelligent malware (Kaloudi & Li, 2020). Generative AI devices decrease the technical obstacles of the opponents of the vital infrastructure, such as energy and safety systems of the population (Suleiman & Donkor, 2025). Such a dual-use capability of AI implies that resilience-by-design should consider adversarial ML risks alongside conventional cyber threats.

On the architectural level, the ML integration based on resilience needs to include layered controls. To avoid the manipulation of sensor streams and training datasets, first, there is a need to ensure data integrity and secure ingestion mechanisms. Second, the model must be validated, explained, and performance monitoring must be in place to identify degradation or aberrant outputs. Third, automation decision systems should have redundancy and fallback mechanisms that guarantee that one can override or isolate such systems without interfering with fundamental physical processes. These design principles correlate with the wider recommendations on cybersecurity governance in the NIST Cybersecurity Framework 2.0, which focuses on governance, risk management strategy, and supply chain issues (Pascoe, 2023).

Publisher: SARC Publisher

The effectiveness of the ML integration is also informed by organizational maturity. The comparative studies on cybersecurity capability maturity models show that structural governance and risk management practices are not uniformly adopted in different sectors (Rea-Guaman *et al*., 2017; Garba *et al*., 2020). The sophisticated ML systems can be implemented in the environment without proper monitoring, professional experience of the staff, and control without adequate maturity. Research in the utility and energy industry attests that technical controls are still compromised by human-factor vulnerabilities and the lack of security culture (Panful *et al*., 2025b; Panful *et al*., 2026).

So, incorporating ML into CPS could not be considered a pure performance improvement. It is an architectural change that changes system dependencies, attack surfaces, and governance requirements. Resilience-by-design requires that ML components be integrated into secure, fault-tolerant, and policy-congruent architectures that provide triple bottom line coverage of technical robustness, adversarial exposure, and organizational readiness.

AI-DRIVEN PREDICTIVE ANALYTICS AND ADAPTIVE CONTROL IN RESILIENT CPS

One of the most important sources of machine learning (ML) applications to cyber-physical systems (CPS) is predictive analytics. AI-powered fault detection and diagnostics (FDD) models are applied in nuclear power plant control systems to detect the slightest anomalies in sensor behavior and forecast unstable situations before they lead to system failures (Bonsu & Adeoye, 2025). These models contribute to continuity in the operations of the organization as it decreases the response time and promote preventative maintenance efforts. Equally, real-time diagnostics and self-corrective control schemes that are based on AI make adaptive changes to the system, enabling the control mechanism to react dynamically to changing working conditions (Adeoye *et al*., 2025).

In compliance-intensive industries, the models of anomaly detection have been shown to perform with quantifiable increases in detection rates and monitoring efficiency over the old methods of manual and rule-based anomaly detection (Gaye *et al*., 2025a). These systems automate the massively analyzed data and give warning signals of abnormal pattern occurrences. Nevertheless, they can perform well only with the help of solid data pipelines, model validation, and interpretation mechanisms.

In terms of resilience, predictive analytics should be incorporated in a larger adaptive control architecture. Resilience engineering focuses on the ability to detect and track potential disruptions and stop them at an early stage (Woods, 2015). Predictive models have a direct relationship with the anticipation role in that they forecast any arising risk. But resilience needs response and learning abilities as well. The principles of Safety-II emphasize that adaptive changes are necessary to maintain performance when the conditions are changing (Hollnagel *et al*., 2015). Thus, predictive analytics cannot and should not exist in isolation but must exist in feedback loops allowing timely human supervision, remedial action, and ex post facto learning.

The necessity of resilient predictive architecture is further explained by cyber threat intelligence. Adversarial models of ICS, including ATT&CK, show that attackers can target monitoring and control systems at various phases or alter their data (Alexander *et al*., 2020). SCADA analyses based on cases demonstrate that vulnerabilities in access control and system configuration are frequently the source of operational effects (Panful *et al*., 2025a). When predictive analytics is based on tainted data streams or open ports, its results can increase risk instead of reducing it.

Also, the predictive defense strategies are complicated by the AI-enabled adversarial capabilities. The surveys of AI-based cyber threats report on how automatic malware response and intelligent generation of attacks evolve in response to detection (Kaloudi & Li, 2020). Generative AI also reduces the obstacle to those who attack critical infrastructure (Suleiman & Donkor, 2025). These developments imply that predictive resilience should include adversarial awareness, model robustness tests, and secure data governance.

These requirements are supported by policy direction. The NIST AI Risk Management Framework emphasizes the fact that credible AI systems should be secure, reliable, and resilient during the lifecycle (NIST, 2023). Similarly, the NIST Cybersecurity Framework 2.0 focuses on the functions of governance, ongoing monitoring, and improvement as fundamental cybersecurity functions (Pascoe, 2023). Lifecycle monitoring, performance evaluation, and integration of the

Publisher: SARC Publisher

structured incident response should thus be part of predictive analytics that will be in tandem with these frameworks.

AI-based predictive analytics must be an element of layered defense and adaptation control systems in the case of resilient CPS architectures. The data validation, redundancy, explainability, and human-in-the-loop control are required to make sure that predictive outputs reinforce the operational integrity instead of destabilizing it. Predictive analytics placed in governance-consistent and fault-tolerant designs can be used to improve the ability to anticipate and respond to events and maintain the stability of the system in adversarial and uncertain environments.

FAILURE MODES AND RISK AMPLIFICATION IN ML-INTEGRATED CPS

Whereas machine learning (ML) can improve predictive accuracy and automate cyber-physical systems (CPS), it also introduces new failure modes that increase systemic risk. Compared to deterministic control algorithms, ML models rely on data quality, statistical inference, and dynamic operating environments. According to the NIST AI Risk Management Framework, AI systems work in socio-technical settings and can lead to unpredictable results in case of any data or context alteration (NIST, 2023). This renders the issues of model drift, biased training data, and underperforming generalizations very important resilience issues.

In safety-related industries like nuclear power, AI-based fault detection enhances the detection of anomalies, yet there are issues with data constraints, explainability, and insecurity (Bonsu & Adeoye, 2025). Correspondingly, the AI-based compliance monitoring systems also indicate the problems of data availability, regulatory restrictions, and lower transparency in automated decisions (Gaye *et al*., 2025a). Such limitations point to the fact that predictive performance will not assure operational resiliency.

The model of cyber threats also shows how the ML-enabled architecture can be abused. The ATT&CK ICS framework identifies adversarial strategies that travel horizontally to IT settings to the technology systems in operation (Alexander *et al*., 2020). The U.S. SCADA case studies reveal that misconfigurations of remote access, combined with poor credential hygiene, can be further escalated into a direct process manipulation (Panful *et al*., 2025a). Unless ML systems use secure data flows or unauthenticated user interfaces, attackers can use model input to mitigate predictions or cause unsafe control processes.

AI technologies are also increasing the attack surface. Studies on AI-powered cyber threat report on the use of intelligent automation by offenders to modify phishing activities and malware to avoid detection systems (Kaloudi & Li, 2020). Generative AI also reduces the technical aspect of coordinated cyberattacks against critical infrastructure (Suleiman & Donkor, 2025). These advancements make more intrusions, which are more advanced and enhanced with AI, possible in the case of ML-incorporated CPSs.

On top of the technical weaknesses, organizational weaknesses add to the risk. The literature in the energy industry reveals that human-fact-related weaknesses, such as inadequate training, the lack of policy compliance, and low awareness, are some of the significant causes of cybersecurity breaches (Panful *et al*., 2025b). Comparative studies between the two utilities show that regulatory compliance is not always a guarantee of the effectiveness of the security culture in the long term (Panful *et al*., 2026). Lack of knowledge or control in an ML-integrated system can result in model degradation or misconfiguration existing without being noticed.

The scope of capability maturity analysis also displays a mixed distribution of organized practices of cybersecurity in domains (Rea-Guaman *et al*., 2017; Garba *et al*., 2020). Advanced ML implementation can run in environments where appropriate resilience safeguards are not in place, without the presence of mature governance and monitoring processes. Fragmentation of AI regulation is also pointed out by regulatory scholarship, which makes accountability and risk management difficult (Agbadamasi *et al*., 2025; Brakye & Adam, 2025).

According to the resilience engineering approach, such failure modes must be resolved by anticipatory design. Safety-II focuses on being able to adapt to changes instead of responding to failure (Hollnagel *et al*., 2015). The 4 functions of resilience described by Woods are anticipate, monitor, respond, and learn, which offer a systematic view of identifying how the failures of ML can spread and how the buildup of safeguards can be broken (Woods, 2015).

In ML-integrated CPS, failure is rarely isolated. Model errors can cascade across control loops,

Publisher: SARC Publisher

influence automated decisions, and interact with human operators under time pressure. Thus, resilience-by-design needs the continuous validation of the model, adversarial testing, the proper control of the data, a clear human override mechanism, and the organization of post-incident learning. In their absence, ML integration can end up increasing the systemic risks that it is supposed to decrease.

A RESILIENCE-BY-DESIGN ARCHITECTURAL FRAMEWORK FOR ML-INTEGRATED CPS

The above paragraphs indicate that the resilience of machine learning-inspired cyber-physical systems (CPS) should go beyond the controls. It needs an architectural style that incorporates adaptive capacity, antagonistic awareness, and governance harmony into system design. Technical robustness, human oversight, and policy compliance are thus interdependent layers that are incorporated in resilience-by-design.

On the conceptual level, resilience engineering focuses on the continuity of the system functions on anticipation, monitoring, response, and learning (Woods, 2015). Safety-II also emphasizes the need to ensure that systems maintain adaptive performance within variability and not just catastrophe prevention (Hollnagel *et al*., 2015). Resilience in the cyber domain encompasses preparation, absorption, recovery, and adaptation at the lifecycle levels (Linkov & Kott, 2019). It is these principles that offer the structural basis of architectural integration.

Layered Architectural Model

The resilience-by-design architecture of ML-integrated CPS can be structured in five layers of interacting layers:

Secure Sensing and Data Integrity Layer

This layer provides a reliable sensor and control device, and enterprise system data ingestion. The case studies of SCADA reveal that operational compromise is possible due to the exposure of HMIs, weak credentials, and deficient segmentation (Panful *et al*., 2025a). The use of structured adversary models like ATT&CK in ICS proves that attackers use vulnerabilities in IT and beyond wireless boundaries (Alexander *et al*., 2020). Accordingly, encrypted communication, access control measures, network segmentation, and supply chain assurance mechanisms are underpinned.

Robust ML and Model Governance Layer

ML parts need to be regularly checked and reviewed. The use of AI in FDD systems to enhance predictive behaviors in nuclear control settings proves to be more advantageous but also brings up issues about the quality of the data and cybersecurity risk (Bonsu & Adeoye, 2025). NIST AI Risk Management Framework insists that AI systems should be secure, reliable, and resilient at all stages of the lifecycle (NIST, 2023). In this layer, there are model validation, adversarial testing, drift detection, explainability mechanisms, and secure model update processes.

Adaptive Control and Fault-Tolerant Execution Layer

The control systems should be linked to predictive analytics with secure feedback mechanisms. The explanations of how AI can facilitate adaptive operations are based on real-time diagnostics and self-correcting control schemes (Adeoye *et al*., 2025). Nevertheless, the fallback mechanisms and the human override controls are needed to avoid unsafe automated actions. Resilience demands that automated decisions may be isolated and that they won't interfere with fundamental physical processes.

Human Supervision and Organizational Resilience Layer

CPS stability continues to lie through human operators. Studies in the energy industry affirm that employee ignorance and lax adherence patterns are the leading contributors to the violation (Panful *et al*., 2025b). Besides, regulatory compliance is not a sufficient factor of cybersecurity culture maturity (Panful *et al*., 2026). It is a layer where workforce training, role clarity, escalation procedures, and continuous improvement are developed.

Governance and Policy Alignment Layer

Architectural resilience should conform to the national standards and the regulatory frameworks. The NIST Cybersecurity Framework 2.0 highlights the governance and risk management strategy as well as supply chain considerations (Pascoe, 2023). The AI RMF divides the AI risk management into Govern, Map, Measure, and Manage functions (NIST, 2023). Capability maturity assessments indicate an unbalanced use of defined cybersecurity practices (Rea-Guaman *et al*., 2017; Garba *et al*., 2020), which supports the importance of quantifiable governance integration.

Publisher: SARC Publisher

Addressing Adversarial and Emerging Risks

AI technologies raise dual-use issues. AI-based threat surveys provide insights into dynamic malware and automated attack plans (Kaloudi & Li, 2020). Generative AI also reduces the obstacles to enemies of important infrastructure (Suleiman & Donkor, 2025). Resilience-by-design architecture should hence include adversarial ML security, secure data provenance, and formal red-teaming.

The problem of regulatory fragmentation in AI oversight makes the process of accountability and enforcement complex (Agbadamasi *et al*., 2025; Brakye & Adam, 2025). By incorporating the governance mechanisms directly into the layers of architecture, one can eliminate the usage of external compliance audits and incorporate accountability during operations.

Integrative Perspective

The concept of resilience-by-design in ML-integrated CPS is not one of the controls or frameworks. It is an integrated framework that coordinates data security, model resilience, dynamic control, human supervisory and policy governance. With resilience functions introduced into these layers, organizations will be likely to minimize the probability of ML failures, adversarial manipulation, or organizational vulnerabilities growing into systemic disruption.

This architectural model forms the basis of the following section of the work that assesses resilience-by-design in connection with the U.S. cybersecurity and AI policy frameworks.

POLICY AND STANDARDS ALIGNMENT FOR RESILIENCE-BY-DESIGN

The concept of resilience-by-design in machine learning-based cyber-physical systems (CPS) would have to meet the existing U.S. cybersecurity and AI governance frameworks. Issues of technical robustness can never be obtained in the absence of architectural design being dislocated with regulatory expectations, risk governance and accountability mechanisms.

In NIST Cybersecurity Framework (CSF) 2.0, a structured taxonomy of cybersecurity outcomes is structured around six core functions Govern, Identify, Protect, Detect, Respond and Recover (Pascoe, 2023). The introduction of the Govern function is the symptom of a transition to integrating cybersecurity into enterprise risk management and supply chain management. This is specifically useful in a resilience-by-design architecture where this facilitates governance-layer integration, where the ML deployment decisions are aligned with established risk tolerance, roles and oversight responsibilities.

Similarly to CSF 2.0, the NIST AI Risk Management Framework (AI RMF 1.0) divides AI governance into four functions, namely Govern, Map, Measure, and Manage (NIST, 2023). The AI RMF also lays stress on the fact that AI systems should be valid, reliable, safe, secure, and resilient during their lifecycle. In the case of ML-integrated CPS, this would mean constant monitoring of model performance, bias control, explainability and safe model updates. The integration of AI RMF functions with CPS architecture is the key to the predictive analytics and adaptive control systems being responsible and transparent.

Nevertheless, the presence of frameworks does not mean that there is standardized implementation. Comparative studies of cybersecurity maturity models demonstrate that there are differences in the degree of governance, risk measurement practices, and consideration of continuous improvement activities across sectors (Rea-Guaman *et al*., 2017; Garba *et al*., 2020). Such differences indicate that resilience-by-design should include quantifiable maturity standards as opposed to policy documentation only.

Regulatory scholarship also brings out fragmentation in the U.S. AI governance. The existing control systems are spread out, and there is no central authority on AI regulation (Agbadamasi *et al*., 2025). Such a decentralized system will raise questions of accountability, especially when AI systems affect critical infrastructure functions. Moreover, the changing disclosure policies on cybersecurity are forcing organizations to bear the burden of proving organized risk governance (Brakye & Adam, 2025). Integrating resilience into the system architecture can enable transparent reporting and defensible compliance.

Intelligence of threats enhances the requirement for coordinated policy harmony. The AI-based cyber threats, such as automated phishing and intelligent malware, keep developing (Kaloudi and Li, 2020). Another threat brought by generative AI is the possible rise of infrastructure-targeted attacks, which makes the preparation and response coordination to be prioritized by federal agencies (Suleiman & Donkor, 2025). Such policy alignment should, therefore, deal with traditional

Publisher: SARC Publisher

cybersecurity controls, as well as adversarial AI risk mitigation strategies.

The example of sector-specific evidence of the U.S. utilities demonstrates that, without governance integration, technical safeguards prove to be not enough. Remote-access management and segmentation are recurring weaknesses in the SCADA incident analysis (Panful *et al*., 2025a). Organizational literature attests to the fact that compliance-based methods cannot necessarily result in a long-term security culture (Panful *et al*., 2026; Panful *et al*., 2025b). Bridging workforce development and cultural maturity with architectural controls is therefore also needed in policy alignment in resilience-by-design.

Overall, resilience-by-design is consistent with the spirit of CSF 2.0 and AI RMF 1.0 but needs more architectural interconnections. The governance functions should guide the deployment decisions of the ML models, the measurement functionalities should monitor the level of performance of both the models, and the system and the management functions should entail adaptive improvement. With these principles of policy embedded in the design of CPS, organizations will be able to transition to proactive, formalized resilience instead of being reactionary to compliance.

RESEARCH GAPS

Although there is increased research on the field of resilience engineering, AI governance, and CPS security, there are still significant gaps in how the various fields can be combined into a unified resilience-by-design system.

First, literature in resilience engineering highlights adaptive capacity and socio-technical coordination, yet fails to operationalize machine learning (ML)-particular risks including model drift, adversarial manipulation, or black box decision-making (Woods, 2015; Hollnagel *et al*., 2015). Even though resilience is contextualized with respect to anticipation, monitoring, response, and learning, information is scarce regarding embedding these functions into ML lifecycle management in the CPS settings. On the same note, cyber resilience models are used to explain preparation, absorption, recovery, and adaptation (Linkov & Kott, 2019), but its architectural implementation to ML-integrated control systems is not well-established.

Second, AI-related research tends to emphasize the precision of prediction and the effectiveness of automation too much as opposed to systemic resilience. The opinions on AI-based fault detection and anomaly detection systems indicate the improvement of the performance; however, they also note the limitations in terms of interpretability, data quality, and vulnerability to cybercrime (Bonsu & Adeoye, 2025; Gaye *et al*., 2025a). Little empirical analysis exists of the ways in which the failures of ML can spread across interconnected CPS architectures.

Third, the risk of adversarial AI is still to be inadequately incorporated into the CPS resiliency frameworks. The literature records the increasing popularity of AI-based phishing, intelligent malware, and creation of automated attacks (Kaloudi & Li, 2020; Suleiman & Donkor, 2025). Detailed ICS threat modeling also demonstrates the path that attackers follow on IT and OT layers to corrupt operation processes (Alexander *et al*., 2020). Nevertheless, there are few studies which associate these adversarial situations with resilience engineering metrics or architectural stress-testing frameworks.

Fourth, there is disparity in governance alignment in sectors. Comparative maturity studies show that there is a difference in the practice of cybersecurity governance (Rea-Guaman *et al*., 2017; Garba *et al*., 2020). Fragmentation of AI regulation is another issue identified by regulatory scholarship and the emerging requirement of disclosure (Agbadamasi *et al*., 2025; Brakye & Adam, 2025). Although NIST CSF 2.0 and AI RMF 1.0 offer systematic guidance (Pascoe, 2023; NIST, 2023), there is still little integration in the application to ML-enabled CPS.

Lastly, technical safeguards are still compromised by organizational vulnerabilities. Research in the energy sector reveals the ongoing failures in the level of awareness and compliance behavior among the employees (Panful *et al*., 2025b). Comparative research of the public and the private utilities reveals that formal compliance is not necessarily enough to establish a resilient cybersecurity culture (Panful *et al*., 2026). Further empirical studies must be carried out that directly correlate organizational maturity with architecture resilience results.

FUTURE DIRECTIONS

To move towards resilience-by-design of ML-integrated CPS, research and implementation in technical, organizational, and policy spheres are required to be coordinated.

To begin with, the following studies ought to institutionalize quantifiable resilience measures of

Publisher: SARC Publisher

ML-enabled CPS. This involves connecting model robustness, explainability, and drift detection means to system-level recovery and adaptation features (Woods, 2015; NIST, 2023). There should be clear metrics that facilitate an organized assessment other than predictive accuracy.

Second, unified adversarial stress-testing models on ML components in critical infrastructures should be created. These frameworks must include the ICS-specific threat models (Alexander *et al*., 2020) and incorporate the changing patterns of attacks that are enabled by AI (Kaloudi & Li, 2020; Suleiman & Donkor, 2025).

Third, AI RMF lifecycle functions, Govern, Map, Measure, and Manage, should be clearly mapped to CPS architectural layers (NIST, 2023). Corresponding to these functions with the CSF 2.0 governance and risk management strategies (Pascoe, 2023) would assist in the process of integrating technical design and regulatory compliance measurably.

Fourth, architectural resilience benchmarks should be used together with capability maturity assessments. The integration of the maturity model knowledge (Rea-Guaman *et al*., 2017; Garba *et al*., 2020) with the system-level resilience assessment would assist in making sure that the level of governance is at par with the complexity of ML deployment.

Fifth, human-in-the-loop oversight models should be enhanced in future work. Adaptive capacity can be supported by improving training of operators, explainable AI interfaces, and well-managed escalation, which are in line with Safety-II principles (Hollnagel *et al*., 2015). The effectiveness of resilience-by-design can be practically demonstrated by the empirical validation of the high-risk fields like energy and utilities (Panful *et al*., 2025a).

By considering these aspects, resilience-by-design can transform into a conceptual compilation of approaches into an empirically established strategy of architecture and governance of machine-learning-integrated cyber-physical systems.

CONCLUSION

Machine learning applied to the architecture of cyber-physical systems has greatly enhanced predictive monitoring, anomaly detection, and automated control of key areas of infrastructure. Nevertheless, there are also new technical, organizational, and governance risks brought by these advancements. With increased data-driven and interrelated CPS environments, models, data pipelines, or oversight malfunctions can quickly spread into operational failure.

One way of responding to this complexity in a structured manner is through resilience-by-design. Instead of considering resilience as a responsive recovery measure, it builds an adaptive capacity, layered security, human control, and governance alignment into system design. It is a strategy that combines technical strength, organizational maturity, and adherence to policy.

To ensure long-term safety of machine learning-based infrastructure, it is important to have a joint design of the sensing, analytics, control, and governance layers. Resilience-by-design can provide a feasible way to achieve AI-integrated systems security in a shifting threat environment by integrating engineering practice with structured risk management and adaptive principles.

REFERENCE

1. Adeoye, M. B., Annankra, J. A., & Yakin, Z. "AI-driven real-time diagnostics and self-correcting control schemes for next-generation nuclear energy systems." *World Journal of Advanced Research and Reviews* 27.3 (2025): 1550-1557.
2. Agbadamasi, T. O., Opoku, L. K., Adukpo, T. K., & Mensah, N. "Navigating the intersection of US regulatory frameworks and artificial intelligence: Strategies for ethical compliance." *World Journal of Advanced Research and Reviews* 25.3 (2025): 969-979.
3. Alexander, O., Belisle, M., & Steele, J. "MITRE ATT&CK for industrial control systems: Design and philosophy." *The MITRE Corporation: Bedford, MA, USA* 29 (2020): 21-85.
4. Bonsu, J. O., & Adeoye, M. B. "AI-Driven Fault Detection and Diagnostics in Nuclear Power Plant Control Systems: A Review." *Journal Of Engineering And Computer Sciences* 4.9 (2025): 552-559.
5. Brakye, K., & Adam, A. B. K. "CYBERSECURITY RISK DISCLOSURE AND REGULATORY COMPLIANCE: EVALUATING MARKET SENSITIVITY AND DISCLOSURE EFFECTIVENESS IN US PUBLIC COMPANIES."
6. Garba, A. A., Siraj, M. M., & Othman, S. H. "An explanatory review on cybersecurity capability maturity models." *Adv. sci. technol. eng. syst. j* 5.4 (2020): 762-769.
7. Gaye, A., Apaflo, B. N., & Donkor, A. A. "Evaluating the effectiveness of AI-driven

Publisher: SARC Publisher

anomaly detection in food safety compliance monitoring." (2025).
8. Hollnagel, E., Wears, R. L., & Braithwaite, J. "From Safety-I to Safety-II: a white paper." *The resilient health care net: published simultaneously by the University of Southern Denmark, University of Florida, USA, and Macquarie University, Australia* (2015).
9. Kaloudi, N., & Li, J. "The ai-based cyber threat landscape: A survey." *ACM Computing Surveys (CSUR)* 53.1 (2020): 1-34.
10. Linkov, I., & Kott, A. "Fundamental concepts of cyber resilience: Introduction and overview." *Cyber resilience of systems and networks*. Springer, Cham, 2019. 1-25.
11. AI, N. "Artificial intelligence risk management framework (AI RMF 1.0)." *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai* (2023): 100-1.
12. Panful, B., Apaflo, B., & Hutchful, N. "Cyber-Physical Systems Under Threat: A Case-Study Review of Recent SCADA Attacks in the US Utility Sector." *Journal Of Engineering And Computer Sciences* 4.12 (2025): 104-117.
13. Nnadi, K., Okrah, Y. A., & Opoku, J. A. "Advancing Zero Trust for SMEs: A Review of Short-Lived Certificates, MFA, and Lightweight Identity Solutions."
14. Panful, B., Apaflo, B., Filani, A., Nnadi, K., & Hutchful, N. "Human factor vulnerabilities in energy industry cybersecurity: Assessing employee awareness and behavior in breach prevention." *International Journal for Multidisciplinary Research (IJFMR)* 7.6 (2025): 1-18.
15. Pascoe, C. E. "Public draft: The NIST cybersecurity framework 2.0." *National Institute of Standards and Technology* (2023).
16. Rea-Guaman, A. M., San Feliu, T., Calvo-Manzano, J. A., & Sanchez-Garcia, I. D. "Comparative study of cybersecurity capability maturity models." *International conference on software process improvement and capability determination*. Cham: Springer International Publishing, (2017).
17. Suleiman, A., & Donkor, A. "Strengthening Homeland Security Preparedness against Adversarial Use of Generative AI in the United States: A Scoping Review." *World Journal of Advanced Research and Reviews.* (2025).
18. Woods, D. D. "Four concepts for resilience and the implications for the future of resilience engineering." *Reliability engineering & system safety* 141 (2015): 5-9.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Addo, E. "Resilience-by-Design in Machine Learning–Integrated Systems: Architectural Frameworks and Policy Implications" *Sarcouncil Journal of Multidisciplinary* 6.4 (2026): pp 1-10.

Publisher: SARC Publisher