Advances in IT and Digital Security
ISSN 3143-4266
DAC Insight Publishers
https://journals.dacinsightpublishers.com/AIDY

SECURING GENERATIVE AI PIPELINES IN CRITICAL INFRASTRUCTURE ENVIRONMENTS: A CONCEPTUAL ARCHITECTURE FOR IDENTITY, DATA INTEGRITY, AND ADVERSARIAL RISK

Mayokun Philips Adegbite[1]*, Mubarak Olayiwola Ahmed[2], Abolaji Adebayo[3]
[1] Oshawa Power and Utilities Corporation, Ontario, Canada
[2] Cisco Systems, Lagos, Nigeria
[3] Computacenter, USA
Corresponding author*: maphilips1989@gmail.com

ABSTRACT

The deployment of generative artificial intelligence in critical infrastructure introduces security and governance challenges that classical machine learning practice does not fully address. Large language models, retrieval-augmented pipelines, and multimodal foundation models are being integrated into energy, water, and emergency response operations for documentation assistance, procedure summarisation, operator decision support, and incident response. The security properties of these systems differ fundamentally: inputs are open-ended natural language, outputs can trigger downstream actions through connected tools and agents, and adversaries can manipulate behaviour through prompt injection, retrieval poisoning, model extraction, and supply chain compromise. This paper presents a three-concern architecture for securing generative AI pipelines in critical infrastructure. Identity addresses authentication and authorisation of users, agents, models, and data sources. Data integrity addresses the provenance, validation, and freshness of training, retrieval, and prompt data. Adversarial risk addresses prompt injection, jailbreaking, output exfiltration, and supply chain attacks against foundation models. Controls are aligned with the NIST AI Risk Management Framework, NIST SP 800-207 Zero Trust Architecture, the OWASP Top Ten for LLM Applications, and the EU AI Regulation. Three reference scenarios in energy, water, and emergency response illustrate the architecture in practice. The principal contribution is a practitioner-ready framework providing a common vocabulary for security architects, operational technology engineers, and governance professionals engaged in responsible generative AI adoption in high-consequence environments.

Keywords: *generative AI; large language models; critical infrastructure; identity; data integrity; adversarial machine learning; pipeline security; conceptual framework.*

1. INTRODUCTION

Generative artificial intelligence systems based on large language models, multimodal foundation models, and retrieval-augmented pipelines are being integrated into critical infrastructure operations for tasks ranging from documentation assistance and procedure summarization to operator decision support and incident response (Ozowara et al., 2025a; Adebayo et al., 2025b; Kumuyi et al., 2024a). The integration offers operational benefits including faster information access, reduced documentation burden, and structured support for complex decisions (Adebayo et al., 2025a; Kumuyi et al., 2023). The integration also introduces security considerations that differ in important ways from those of classical machine learning systems (Kumuyi et al., 2024b).

The security profile of generative artificial intelligence systems is fundamentally different from earlier AI deployments. These systems accept open-ended natural language inputs, retrieve information from external knowledge bases, and may take actions through connected tools. In agentic configurations, multiple specialised agents coordinate work, each capable of independent decision-making within the scope of delegated authority.

This conceptual paper presents an architecture for securing generative artificial intelligence pipelines in critical infrastructure environments. The architecture is structured around three concerns. Identity addresses the authentication and authorization of users, agents, models, and data sources. Data integrity addresses the provenance, validation, and freshness of training, retrieval, and prompt data. Adversarial risk addresses the threat surface introduced by prompt injection, jailbreaking, output exfiltration, and supply chain attacks against foundation models (Adebayo, 2022a).

The paper makes three principal contributions. First, it articulates the three-concern architecture spanning identity, data integrity, and adversarial risk. Second, it aligns the architecture explicitly with the NIST Artificial Intelligence Risk Management Framework, the Zero Trust Architecture specification, and the OWASP Top Ten for Large Language Models. Third, it provides reference scenarios illustrating architecture application to energy, water, and emergency response deployments.

Adoption trajectories for generative artificial intelligence in critical infrastructure indicate growing operational deployment alongside continued caution about high-consequence applications. The trajectory reflects both the operational benefits of the technology and the recognition that critical infrastructure environments require disciplined adoption (Kumuyi, 2024b; NIST, 2024b).

The security challenge introduced by generative artificial intelligence in critical infrastructure is qualitatively distinct from the security challenges associated with earlier machine learning deployments. Classical machine learning systems for fault prediction, anomaly detection, or load forecasting operate on structured operational data with well-defined input and output schemas. Their attack surface is relatively narrow: adversaries must manipulate the input data streams that feed the model or the training data that shaped it. Generative artificial intelligence systems, by contrast, accept open-ended natural language inputs, produce open-ended natural language outputs, and in agentic configurations may interact with external systems, databases, and physical control interfaces (Bommasani et al., 2021; Greshake et al., 2023). The breadth of the generative AI attack surface requires explicit architectural treatment that classical machine learning security practice does not provide.

The National Institute of Standards and Technology has recognized this qualitative shift in its Generative AI Profile, which identifies unique risks associated with large language models that are not addressed by existing AI risk management guidance (NIST, 2024a). These risks include the susceptibility of language models to prompt injection, the potential for training data extraction, the difficulty of reliably detecting harmful or inaccurate outputs, and the challenge of maintaining consistent behavior under adversarial prompting. The CISA Cybersecurity Performance Goals, established for critical infrastructure entities, provide baseline cybersecurity practices against which AI-specific controls must be layered (CISA, 2023a). The intersection of generative AI capabilities with critical infrastructure operational requirements creates a governance challenge that requires coordinated response from security practitioners, operational technology engineers, and organizational risk managers.

2. RELATED WORK

A decade of adversarial machine learning research has established that the vulnerability of learned models to adversarial manipulation is a structural property of gradient-based learning rather than a defect of specific architectures (Biggio & Roli, 2018).

The security of artificial intelligence systems has attracted sustained research attention across multiple communities. The adversarial machine learning community has established a foundational taxonomy of attacks that target AI systems at training time, through data poisoning and backdoor injection, and at inference time, through adversarial examples and model extraction (Goodfellow et al., 2015; Madry et al., 2018). Szegedy et al. (2014) demonstrated that neural networks could be caused to misclassify inputs through small, imperceptible perturbations, founding the study of adversarial examples that has grown into a substantial research program. Carlini and Wagner (2017) developed stronger attack methods that established that defensive distillation, an early proposed defense, was insufficient. The progression of attack and defense research has produced a body of knowledge applicable to the security evaluation of generative AI systems, even though the specific attack modalities for large language models differ from those applicable to classification systems.

The vulnerability of deep neural networks to adversarial examples extends across modalities including image, audio, and text, with significant implications for multimodal AI systems deployed in critical infrastructure (Akhtar & Mian, 2018).

The security of large language models specifically has emerged as a research priority corresponding to the rapid deployment of these systems in high-stakes applications. Greshake et al. (2023) demonstrated indirect prompt injection in real-world LLM integrations, establishing the practical exploitability of an attack class previously discussed primarily in theoretical terms. Wei et al. (2023) characterized the failure modes of LLM safety training, demonstrating that aligned models could be jailbroken through a variety of prompt constructions and analyzing the mechanisms underlying safety training failure. Weidinger et al. (2022) provided a comprehensive taxonomy of risks posed by language models spanning six risk areas, providing a structured framework for risk assessment that complements the security-focused taxonomy of the NIST AI 100-2 document (Vassilev et al., 2024). Yao et al. (2024) surveyed LLM security and privacy from the perspectives of both attack and defense, synthesizing a rapidly growing research literature. Perez and Ribeiro (2022) characterized prompt injection as a category of attack distinct from adversarial examples and cross-site scripting, establishing the conceptual grounding for the growing body of work on LLM-specific attacks.

The security of critical infrastructure has been studied extensively in the operational technology and industrial control system security communities. Igure et al. (2006) provided an early analysis of security issues in SCADA networks that established many of the concerns addressed by subsequent standards and regulatory frameworks. Yadav and Paul (2019) reviewed architecture and security in SCADA systems, providing a contemporary synthesis of the research literature. The MITRE ATT&CK framework for Industrial Control Systems, based on the broader ATT&CK framework documented by Strom et al. (2018), provides a practitioner-oriented taxonomy of adversary tactics, techniques, and procedures applicable to critical infrastructure environments. Stouffer et al. (2023) updated the NIST Guide to Operational Technology Security, providing authoritative guidance on the security of operational technology systems that includes emerging considerations for AI integration. Marchang and Sumathy (2021) surveyed adversarial attacks specifically in critical infrastructure contexts, providing a synthesis relevant to the threat landscape analysis in this work.

Review of architecture and security considerations in SCADA systems identifies the convergence of IT and OT networks as the primary driver of expanded attack surface in critical infrastructure environments that now incorporate AI-assisted monitoring (Yadav & Paul, 2019).

Zero trust architecture has emerged as a foundational security principle for enterprise and government environments, with direct applicability to AI pipeline security. Rose et al. (2020) and Stafford (2020) articulated zero trust architecture principles and implementation patterns in the context of the NIST SP 800-207 publication. Phiayura and Teerakanok (2023) analyzed migration to zero trust architecture in enterprise contexts, addressing implementation challenges relevant to organizations adopting these principles for AI system security. The Cybersecurity and Infrastructure Security Agency Zero Trust Maturity Model provide a staged implementation framework applicable to AI security governance within a broader zero trust program (CISA, 2023b). The application of zero trust principles to AI-specific trust boundaries, including the boundary between trusted system instructions and untrusted retrieved or user-provided content, extends the zero-trust architecture to novel contexts not addressed in foundational documents.

Privacy and data governance in AI contexts builds on a body of research and regulatory development that predates the generative AI era. Carlini et al. (2021) demonstrated that training data could be extracted from large language models through systematic querying, establishing the memorization and extraction risk that informs the data integrity concern of this architecture. The General Data Protection Regulation provides a regulatory framework for personal data protection applicable to AI deployments in European markets or involving European data subjects. ISO/IEC 27005 provides risk management guidance that encompasses data privacy risks in information security management systems applicable to AI contexts (ISO, 2018b). The intersection of data privacy regulation with AI training and retrieval practices is an active area of regulatory development that practitioners must monitor as the regulatory environment matures.

3. BACKGROUND

Throughout this paper, the following organisational abbreviations are used: the National Institute of Standards and Technology (NIST); the Cybersecurity and Infrastructure Security Agency (CISA); the International Organization for Standardization (ISO); the International Electrotechnical Commission (IEC); the Department of Energy (DOE); the North American Electric Reliability Corporation (NERC) and its Critical Infrastructure Protection standards (CIP); the Open Worldwide Application Security Project (OWASP); Supervisory Control and Data

Acquisition systems (SCADA); and the MITRE ATT&CK framework for adversary tactics and techniques.

The adoption of generative artificial intelligence in critical infrastructure is proceeding faster than the development of sector-specific security standards, creating a period of elevated risk in which organizations must rely on general AI security guidance while awaiting sector-specific elaboration. This gap is particularly acute in operational technology environments where existing cybersecurity standards address conventional software systems but do not yet fully address the emergent properties of large language models, retrieval-augmented architectures, and agentic systems. Practitioners operating in this environment must exercise heightened judgment, drawing on multiple sources of guidance and engaging actively with standards development processes to ensure that their operational experience informs the frameworks that will govern AI deployments in critical infrastructure.

Foundation models trained on large corpora of text, code, and other modalities have become widely available through cloud services and open weights releases. Their capabilities span tasks that previously required separate specialized models, including summarization, translation, question answering, and code generation (Adebayo, 2020). Capability evaluation and benchmarking have produced authoritative frameworks for comparing models across multiple dimensions (NIST, 2018b).

Cloud environments introduce security considerations including shared tenancy risks, data residency requirements, and distributed access management that practitioners must evaluate before deploying AI systems in hosted infrastructure (Krutz & Vines, 2010; Ryan, 2013). Enterprise data breach analysis consistently identifies misconfigured cloud access controls as a primary attack vector against cloud-hosted analytical systems (Cheng et al., 2017).

retrieval-augmented generation pipelines combine foundation models with retrieval over organization specific knowledge bases (NIST, 2014; Adebayo, 2025a; Adebayo, 2025b). The pattern addresses the limitations of foundation models in handling current or organization specific information. Practitioners deploy vector databases that support semantic search, structured retrieval over knowledge graphs, and hybrid retrieval that combines multiple approaches.

Agentic systems extend foundation models with the ability to take actions through tool use, plugin invocation, or other side effects. The capability shifts from question answering to action introduces new categories of risk because errors or compromises propagate into downstream systems. Effective agentic security requires explicit tool allowlists, structured input-output specifications, and human approval steps for consequential actions.

Emerging regulatory expectations for artificial intelligence span multiple jurisdictions and topics. The National Institute of Standards and Technology artificial intelligence risk management framework provides voluntary structure for organizational artificial intelligence governance. The European Union artificial intelligence regulation establishes binding requirements with varying intensity by application risk level. Sector specific guidance from the Cybersecurity and Infrastructure Security Agency and other authorities addresses critical infrastructure considerations.

Industry standards activity on generative artificial intelligence security is accelerating, with multiple standards bodies producing relevant guidance. The Open Web Application Security

Project top ten for large language models provides one widely referenced resource. Sector specific guidance from the Department of Energy, the Cybersecurity and Infrastructure Security Agency, and industry consortia complements the broader guidance.

Models trained on web-scale corpora acquired general language capabilities that transfer effectively to narrow task domains (Bommasani et al., 2021; Devlin et al., 2019). This enables rapid deployment for summarisation, question answering, and structured reasoning. Retrieval-augmented generation pipelines extend this capability by grounding model outputs in organisation-specific knowledge bases

Critical infrastructure sectors have adopted artificial intelligence technologies across multiple operational domains. Energy sector applications include predictive maintenance for generation and transmission assets, anomaly detection in operational data, and documentation support for operational procedures (DOE, 2022; NERC, 2023). Water sector applications include process optimization, regulatory compliance monitoring, and incident response support. Emergency response applications include situation summarization, resource coordination, and communications support for field personnel (DOE, 2023). Each domain presents, furthermore, distinct data sensitivity profiles, consequence scales, and regulatory environments that shape security requirements.

The regulatory environment for artificial intelligence in critical infrastructure has developed rapidly. The National Institute of Standards and Technology published its Artificial Intelligence Risk Management Framework to provide voluntary structure for organizational AI governance (NIST, 2023). The Generative AI Profile supplements this framework with guidance specific to the security and safety properties of large language models (NIST, 2024a). The European Union Artificial Intelligence Regulation establishes binding requirements that vary by application risk level, placing many critical infrastructure AI applications in the high-risk category subject to conformity assessment. Sector-specific guidance from the Cybersecurity and Infrastructure Security Agency addresses critical infrastructure considerations in detail (CISA, 2023a; CISA, 2023b). The Department of Energy has published sector-specific maturity models and engineering strategies for operational technology cybersecurity (DOE, 2022; DOE, 2023).

Furthermore, standards activity on information security governance provides foundational structure for the architecture. The ISO/IEC 27001 standard establishes requirements for information security management systems that provide a baseline for AI security governance (ISO, 2022a). ISO/IEC 27002 provides a companion code of practice with detailed implementation guidance for information security controls (ISO, 2022b). ISO/IEC 27005 addresses information security risk management, providing process guidance applicable to AI risk assessment (ISO, 2018b). ISO 31000 provides risk management principles at the enterprise level that encompass AI risk as a component of broader organizational risk (ISO, 2018a). IEC 62443 addresses security for industrial automation and control systems in operational technology environments, providing sector-specific control requirements relevant to critical infrastructure AI deployments (IEC, 2018). The NIST Cybersecurity Framework, now in its second major revision, provides a widely adopted structure for cybersecurity program organization that encompasses AI-related risks (NIST, 2018b; NIST, 2024b).

Zero trust architecture principles provide, therefore, a complementary framing for AI security that emphasizes continuous verification rather than perimeter-based trust (Rose et al., 2020; Rose et al., 2020). The principle that no entity should be trusted by default, regardless of network location,

applies with particular force to AI systems whose inputs and outputs cross multiple trust boundaries. The NIST Special Publication 800-207 articulates zero trust architecture principles in the context of enterprise security programs (Rose et al., 2020). The CISA Zero Trust Maturity Model provides a staged progression framework for implementing zero trust across the identity, device, network, application, and data pillars (CISA, 2023b). NIST SP 800-53 Revision 5 provides a comprehensive control catalog that encompasses controls applicable to AI system governance, including supply chain risk management, configuration management, and audit and accountability (NIST, 2020). The older NIST Risk Management Framework establishes a system life cycle approach to security and privacy that informs AI system authorization and ongoing monitoring (NIST, 2018c).

Frontier AI developers have published preparedness frameworks and responsible scaling policies that establish internal governance standards for managing catastrophic risk, providing a reference point for critical infrastructure operators evaluating AI vendors' safety posture (OpenAI, 2024). The broader security implications of pervasive internet connectivity, including vulnerabilities in critical systems and the governance challenges of securing a hyper-connected infrastructure, are examined in foundational practitioner literature that informs the urgency of the architectural response described in this paper (Schneier, 2018).

4. THREAT LANDSCAPE FOR GENERATIVE AI IN CRITICAL INFRASTRUCTURE

Prompt injection attacks craft inputs that cause the model to ignore its intended instructions and follow attacker provided directives instead. The attack surface includes direct user input, retrieved content, tool outputs, and any other input that reaches the model context (Ozowara, 2025b). Mitigation strategies include structured input handling, output validation, and explicit attention to the privilege boundaries between trusted instructions and untrusted content.

Retrieval poisoning attacks compromise the documents or other content that retrieval-augmented pipelines incorporate into model context. The attack can produce both immediate effects through poisoned content reaching the model and persistent effects if poisoned content is incorporated into knowledge bases. Mitigation strategies include content provenance controls, integrity verification, and explicit attention to the trust level of retrieved content.

Output exfiltration attacks induce the model to emit sensitive information embedded in its training data or in retrieved content. The attack may target intellectual property, personal data, or operational information. Mitigation strategies include output filtering, training data sanitization, and explicit attention to the sensitivity of content that the model can access.

Supply chain attacks against foundation models and the broader generative artificial intelligence ecosystem include compromise of training data, model weights, fine-tuning processes (the supervised adaptation of a pre-trained model to domain-specific tasks using curated training data), and the tooling ecosystem (Ozowara, 2025a). The diverse and rapidly evolving ecosystem creates opportunities for compromise that would be difficult in more mature software ecosystems. Mitigation strategies include integrity verification, vendor risk assessment, and explicit attention to the provenance of model artifacts.

Additionally, model extraction attacks attempt to reconstruct deployed models through repeated queries, potentially producing functional clones or revealing training data. The attack surface is

particularly relevant for organizations that expose models through public or semipublic interfaces. Mitigation strategies include rate limiting, query monitoring, and model watermarking.

Cross modal attacks against multimodal foundation models extend traditional prompt injection to inputs that combine text, image, audio, and other modalities. Implementation patterns for mitigation address each modality and the interactions between them.

The threat surface of generative artificial intelligence systems in critical infrastructure environments differs from conventional information technology threat surfaces in ways that require explicit characterization. Understanding the attack classes that apply to these systems is a prerequisite for architectural response. The academic literature on adversarial machine learning provides a foundational taxonomy that informs this characterization (Goodfellow et al., 2015; Szegedy et al., 2014; Chakraborty et al., 2018).

Prompt injection represents the class of attacks most distinctively associated with large language model deployments (Greshake et al., 2023). The attack exploits the unified treatment of instructions and content in language model context windows: an adversary who can inject text into the model context can potentially override system instructions with attacker-controlled directives. Indirect prompt injection extends this attack surface to content retrieved from external sources, enabling attacks that execute when a model processes documents, web pages, or other content from the environment (Greshake et al., 2023). In critical infrastructure contexts where models retrieve content from operational knowledge bases, indirect prompt injection can produce consequences proportional to the actions the model can take on retrieved instruction. Defense strategies must distinguish between the privilege of system instructions and the trust level of retrieved content, structuring the context to make instruction override detectable and preventable (NIST, 2024a; Carlini et al., 2021).

Furthermore, retrieval poisoning represents a persistent form of prompt injection attack directed at the knowledge base from which retrieval-augmented generation systems draw their context. Rather than crafting a single malicious input, retrieval poisoning introduces adversarial content into the index that a model will retrieve in response to particular queries. The adversarial content can carry embedded instructions that execute when retrieved, producing effects that are difficult to attribute and may persist across many model invocations. Mitigation requires content provenance controls that track the source and authorization state of indexed documents, integrity verification mechanisms that detect unauthorized modification of indexed content, and trust level assignment that governs how retrieved content interacts with system instructions (NIST, 2020; ISO, 2022a).

Additionally, output exfiltration encompasses attacks that induce a model to emit sensitive information from its training corpus or from documents in its retrieval context (Carlini et al., 2021). Language models trained on sensitive organizational data may memorize and reproduce that data in response to carefully crafted queries. retrieval-augmented systems that access documents containing personally identifiable information, intellectual property, or operational sensitive information may expose that information through outputs. Mitigation strategies include output filtering that detects sensitive information patterns before outputs reach users or downstream systems, training data curation that minimizes sensitive memorization, and access controls on retrieval that limit the documents a given user or agent can access (NIST, 2020; ISO, 2022b).

Moreover, supply chain attacks against the generative artificial intelligence ecosystem address the provenance and integrity of the artifacts that compose deployed systems. Foundation model weights acquired from public repositories may have been modified to embed backdoors that activate in response to specific inputs. Fine-tuning pipelines may be compromised to introduce adversarial behavior during model customization. Vector databases may be populated with adversarially crafted embeddings (dense numerical vector representations of text or other content used for semantic similarity search) that cause retrieval to preferentially return attacker-controlled content. The diversity and rapid evolution of the generative AI tooling ecosystem create multiple points of potential compromise that do not exist in more mature software categories. Integrity verification, vendor assessment, and provenance tracking for model artifacts are essential mitigations (NERC, 2023; DOE, 2023).

Furthermore, cross-modal attacks against multimodal foundation models extend the prompt injection threat surface to non-text modalities. Adversarially crafted images, audio, or other inputs can cause multimodal models to misinterpret content, ignore instructions, or produce outputs that serve attacker objectives. In critical infrastructure contexts where models process images of equipment, audio from control rooms, or documents containing charts and diagrams, cross-modal attacks represent a meaningful threat vector that requires mitigation at each modality and at the interfaces between modalities.

Additionally, the Cybersecurity and Infrastructure Security Agency have documented the threat landscape for critical infrastructure more broadly in guidance documents that contextualize AI-specific risks within a wider operational environment (CISA, 2021; CISA, 2022). Ransomware attacks against critical infrastructure entities have demonstrated the potential for cyber incidents to produce operational disruptions at scale. The introduction of AI systems into operational workflows creates additional attack surfaces that must be integrated into existing threat models rather than treated in isolation (CISA, 2023a).

A systematic review of adversarial attacks in critical infrastructure contexts identifies energy, water, and transport as the sectors with the highest documented attack surface exposure to AI-specific threats (Marchang & Sumathy, 2021).

The MITRE ATT&CK framework for Industrial Control Systems provides a practitioner-oriented taxonomy of adversary tactics and techniques specific to operational technology environments that complements the AI-specific threat characterisation in this architecture (Alexander et al., 2020). Critical infrastructure cybersecurity at the policy level is shaped by a sequence of executive directives and strategic documents including the 2021 executive order on improving the nation's cybersecurity, the 2023 National Cybersecurity Strategy, and the 2024 implementation plan governing their execution across federal and critical infrastructure contexts (House, 2021; House, 2023b; House, 2024). Industry research on cybersecurity trends and predictions provides forward-looking intelligence on the threat vectors most likely to affect AI-integrated critical infrastructure in the near term (Gartner, 2021).

Cyber security analysis of critical infrastructures across energy, water, transport, and communications sectors identifies common architectural vulnerabilities and defensive principles that apply across sectors when AI systems are introduced into operational workflows (Maglaras et al., 2018). Multi-aspect adversarial learning frameworks for cybersecurity intelligence integrate attack modelling, threat detection, and response orchestration in ways that inform the adversarial risk concern of the three-concern architecture (Sarker, 2023).

5. ARCHITECTURE OVERVIEW

Table 1: Three-Concern Conceptual Architecture for Securing Generative AI Pipelines in Critical Infrastructure

Concern	Core Security Property	Principal Threats	Key Controls	Framework Alignment
Identity	Establish and verify who (human, agent, model, or data source) is permitted to interact with the pipeline at each trust boundary.	Credential theft; agent impersonation; delegation chain abuse; prompt injection via identity spoofing.	Multifactor authentication; role-based access control; scoped agent tokens; model version identity; data source provenance metadata.	NIST SP 800-207 (Zero Trust); ISO/IEC 27001; NIST AI RMF Govern; CISA Zero Trust Maturity Model.
Data Integrity	Ensure that data entering the model context (training, retrieval, or prompt) is authentic, current, and trusted only to the level its provenance warrants.	Retrieval index poisoning; stale or outdated content; provenance spoofing; sensitive data exfiltration through generated outputs.	Content provenance metadata; freshness filtering; embedding integrity verification; output sensitivity classification; tamper-evident audit logging.	ISO/IEC 27001/27002; NIST SP 800-53 Rev 5; IEC 62443; ISO 31000 risk management.
Adversarial Risk	Resist attempts by adversaries to manipulate model behaviour, extract sensitive information, or use the pipeline as an attack vector against connected systems.	Direct and indirect prompt injection; jailbreaking; model extraction; output exfiltration; supply chain compromise of model artifacts.	Input validation and trust boundary separation; output filtering; rate limiting; red teaming; adversarial robustness evaluation; supply chain integrity verification.	NIST AI 100-2e2023 (Adversarial ML Taxonomy); NIST AI 600-1 (Generative AI Profile); OWASP LLM Top 10; MITRE ATT&CK for ICS.

Note: Framework alignment references are indicative; specific control mapping should be conducted against current versions of cited frameworks. ATT&CK refers to MITRE ATT&CK for Industrial Control Systems (Alexander et al., 2020; Strom et al., 2018).

Table 1 summarises the three-concern framework, mapping each concern to its core security property, principal threat categories, key controls, and alignment with authoritative governance frameworks.

The three concerns overview presents identity, data integrity, and adversarial risk as the organizing structure for the architecture (Kumuyi, 2023). Each concern is addressed through multiple architectural elements that collectively span the lifecycle of generative artificial intelligence deployment (Badmus, 2023). The concerns are not strictly separable because architectural elements often address multiple concerns simultaneously.

Design principles guiding the architecture include defense in depth across the three concerns, explicit trust boundaries between trusted instructions and untrusted content, audit ability of model invocations, and bounded autonomy for agentic systems in operational environments. The principles reflect the maturation of generative artificial intelligence security practice from experimental to production grade capability.

Scope assumptions include the existence of established cybersecurity program functions, the availability of capabilities to authenticate and authorize users and systems, and the organizational maturity to operate generative artificial intelligence systems responsibly (Dagodzo et al., 2022).

The architecture is intended for critical infrastructure organizations adopting generative artificial intelligence at meaningful scale (Taiwo, 2022; Ekechi et al., 2022).

Architecture composition with broader cybersecurity infrastructure includes integration with identity providers, telemetry pipelines, incident response procedures, and governance forums. The integration should support consistent practice across generative artificial intelligence and other cybersecurity domains.

Figure 1. Three Concern Architecture for Generative AI Security

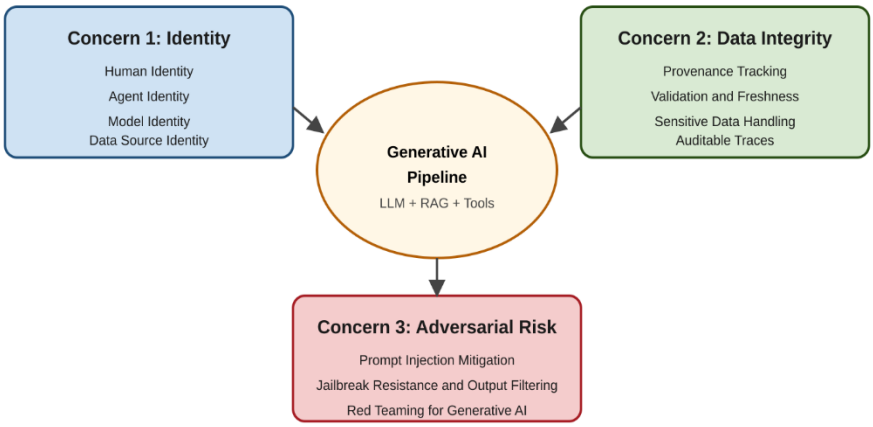

Figure 1: Three Concern Architecture for Generative AI Security

The three-concern architecture integrates identity, data integrity, and adversarial risk into a coherent security framework for generative artificial intelligence deployments in critical infrastructure. The architecture is designed to compose with existing cybersecurity program structures rather than replace them, building AI-specific controls on foundations of established security practice. The architecture is aligned with the NIST Cybersecurity Framework functions of Identify, Protect, Detect, Respond, and Recover, mapping AI-specific controls to each function as appropriate (NIST, 2018b; NIST, 2024b).

The design principles guiding the architecture reflect, therefore, both general security engineering principles and considerations specific to generative artificial intelligence systems. Defense in depth applies across all three concerns, recognizing that no single control is sufficient against the breadth of applicable threat scenarios. Explicit trust boundaries between content of different privilege levels are essential, given the way language models process mixed contexts of system instructions and user inputs. Audit ability of model invocations is a foundational principle, enabling post-incident analysis, compliance demonstration, and governance oversight of model behavior. Bounded autonomy for agentic systems limits the scope of damage from compromised or misbehaving agents (NIST, 2023; NIST, 2024a).

The architecture recognizes that generative AI security does not begin when a model is deployed. Security considerations apply throughout the development lifecycle: during model selection and acquisition, during fine-tuning or adaptation, during retrieval index construction, during system integration, and during production operation. The NIST AI Risk Management Framework provides

a lifecycle perspective on AI risk that encompasses these stages (NIST, 2023). The DOE Cybersecurity Capability Maturity Model provides a complementary maturity framework for operational technology security programs that can be extended to encompass AI system security (DOE, 2023). Integration of AI security into these established lifecycle and maturity frameworks supports consistent governance rather than creating isolated AI-specific program silos.

The architecture maps controls to three authoritative frameworks: the NIST AI Risk Management Framework and its Generative AI Profile, the NIST SP 800-207 Zero Trust Architecture, and the OWASP Top Ten for Large Language Model Applications. The NIST AI RMF organizes AI risk management into four functions: Govern, Map, Measure, and Manage (NIST, 2023). The Generative AI Profile extends this structure with specific guidance on the risks associated with large language models, including prompt injection, data poisoning, and output quality (NIST, 2024a). The Zero Trust Architecture provides an identity-centric security model that aligns with the identity concern of the architecture (Rose et al., 2020; CISA, 2023b). The OWASP Top Ten for LLMs provides a practitioner-oriented taxonomy of common vulnerabilities that complements the more abstract framework structures.

6. IDENTITY CONCERN

Human identity in generative artificial intelligence contexts extends traditional identity practice to encompass the additional sensitivity of model interactions. Technical controls include multifactor authentication for model access, role-based authorization tied to operational responsibilities, and structured logging of identity bound model interactions. The integration with broader identity infrastructure should support consistent governance.

Agent identity addresses the authentication and authorization of automated systems that interact with models. Agent identity governance requires service identity management, scoped tokens that limit agent capability, and structured boundaries between agent action and human approval. Agent identity should support both functional needs and the audit ability of agent activity.

Model identity addresses the auditable identity of each model invocation, supporting the ability to trace decisions back to specific model versions and configurations. Practitioners implement immutable model identifiers, version-controlled model deployment, and structured logging that captures the model used in each interaction (Adebayo, 2025d). Model identity supports both incident response and ongoing governance.

Data source identity addresses the provenance of training data, reference content, and other inputs to model behavior. Effective data source governance incorporates structured provenance metadata, content signing where applicable, and explicit attention to the trust level of each source. Data source identity supports both security analysis and broader compliance considerations (CISA, 2023a; CISA, 2023b).

Delegation patterns for generative artificial intelligence agents address the scenarios in which an agent acts on behalf of a human user (Greshake et al., 2023; DHS, 2023). Recommended controls include scoped tokens, explicit delegation chains, and audit trails that link agent action to the delegating identity (Adebayo, 2023b).

Implementation maturity progresses through recognizable stages that practitioners and assessment frameworks have characterized. Early maturity programs focus on baseline capability development and compliance demonstration (Adebayo, 2021a). Intermediate maturity programs develop

integrated capability across multiple domains and begin to address proactive risk management (Rose et al., 2020). Advanced maturity programs demonstrate sustained capability, continuous improvement, and integration with broader business processes (Apruzzese et al., 2019). The progression is not strictly linear, and entities may advance more rapidly in some domains than others.

Investment decisions in critical infrastructure cybersecurity benefit from explicit cost benefit analysis even where complete quantification is not feasible. Governance decisions should incorporate structured analysis of expected risk reduction, consideration of operational benefits, and explicit attention to the dependencies between investments. The analysis should support decision making rather than provide false precision in the face of substantial uncertainty. Mature programs treat investment analysis as an ongoing discipline rather than a one-time exercise (Kevin, 2023a).

Vendor and supplier relationships affect implementation outcomes substantially. Implementation patterns that strengthen vendor relationships include explicit cybersecurity requirements in procurement, structured vendor risk assessment, ongoing engagement with vendor security practices, and participation in industry consortium activity that influences vendor product direction (Ekechi, 2022). The relationships are long-term and benefit from sustained investment rather than transactional engagement (CISA, 2021).

Organizational change management is a recurring practical challenge in cybersecurity program development (Topol, 2019). Implementation patterns that support successful change include executive sponsorship, structured communication, training and capability development, and explicit attention to the cultural dimensions of change. Technical capability without organizational adoption produces limited operational value.

Human identity management in generative AI contexts builds on established identity and access management practice but introduces additional dimensions specific to model interaction. The sensitivity of model interactions extends beyond the sensitivity of the data they access: the model itself is a powerful reasoning engine that may be induced to take consequential actions, produce harmful content, or reveal sensitive training data in response to crafted inputs. Role-based authorization for model access should therefore be informed by both the data sensitivity of accessible retrieval context and the action scope of connected tools, not only by the functional need to interact with the model. The NIST SP 800-53 access enforcement control family provides a framework for implementing role-based access controls applicable to AI system interactions (NIST, 2020). ISO/IEC 27001 requires organizations to establish access control policies that encompass AI systems within their information security management scope (ISO, 2022a).

Multifactor authentication for model access reduces the risk that stolen credentials enable unauthorized access to sensitive AI capabilities. In operational technology environments, the challenge of implementing multifactor authentication for systems that must respond to time-critical operator actions requires authentication designs that do not introduce unacceptable latency. OpenID Connect provides a widely implemented protocol for federated identity that supports multifactor authentication integration and is applicable to AI system access. The integration of AI system authentication with operational identity providers enables consistent governance of both conventional and AI-mediated access to sensitive information and systems.

Agent identity in generative AI pipelines represents a novel challenge for identity governance. Software agents that invoke language models on behalf of human users, retrieve documents from knowledge bases, or execute tool calls must be assigned identities that support auditing, authorization, and revocation. Service account hygiene practices, including credential rotation, scope minimization, and activity monitoring, apply to AI agents but must be adapted to the operational characteristics of automated systems that may invoke models at high frequency. The principle of least privilege requires that agent credentials carry only the permissions needed for the specific tasks the agent is authorized to perform, not the full permissions of the human user on whose behalf the agent acts. Scoped tokens that expire after the completion of a task sequence and structured delegation chains that link agent actions to authorizing human identities are implementation patterns that support these principles (Rose et al., 2020; CISA, 2023b).

Model identity addresses the auditable tracking of which model version produced which output. In operational environments where model versions are updated, fine-tuned, or substituted, the ability to trace a specific output to a specific model configuration is essential for incident investigation, regulatory compliance, and quality assurance. Immutable model identifiers, version-controlled deployment registries, and structured logging that captures the model identifier for each invocation support model identity. The NIST SP 800-37 Risk Management Framework provides a system authorization process that can be applied to model versions as information system components, establishing explicit authorization for production use and documenting the security controls implemented for each version (NIST, 2018c). The DOE National Cyber Informed Engineering Strategy emphasizes the integration of security considerations into engineering decisions throughout the technology lifecycle, a principle applicable to the selection and authorization of AI model versions (DOE, 2022).

Data source identity in retrieval-augmented pipelines governs which external sources contribute to model context and at what trust level. Sources of high integrity, such as internally curated procedure documents, vendor manuals reviewed by subject matter experts, and regulatory guidance published by authoritative bodies, warrant higher trust than public web content or user-submitted documents. Structured provenance metadata records the origin, authorization state, and last review date of each indexed document. Content signing, where applicable, provides cryptographic assurance of document integrity. The NERC Critical Infrastructure Protection standards impose specific requirements on the integrity of data used in bulk electric system operations that inform data source identity requirements for AI systems in that sector (NERC, 2023). ISO/IEC 27002 provides guidance on supplier relationships and supply chain security applicable to the management of external data sources in retrieval systems (ISO, 2022b).

Furthermore, the concept of delegation chains in generative AI agent systems addresses scenarios in which multiple agents interact in sequence, each acting on behalf of a human user or an upstream agent. A human user may delegate to a planning agent, which in turn delegates to a retrieval agent, a code execution agent, and a communication agent in sequence. Each step in the chain introduces an opportunity for privilege escalation or for adversarial content to redirect agent behavior. Explicit delegation chains that record the originating human identity, the scope of delegation at each step, and the authorization basis for each agent action provide the audit trail necessary for post-incident analysis and compliance demonstration. Scoped tokens that carry only the permissions needed for each step prevent privilege escalation across delegation boundaries. The zero trust architecture principle of verifying explicitly applies to each delegation step, requiring that trust is not assumed

from the existence of a delegation chain but is verified at each boundary (Rose et al., 2020; CISA, 2023b).

Consequently, implementation maturity for AI identity governance follows recognizable stages analogous to maturity models for conventional identity and access management programs. Early-maturity programs focus on baseline authentication for model access and structured logging of model invocations. Intermediate-maturity programs develop integrated identity governance that incorporates AI system identities alongside human and service identities in the same governance infrastructure, supporting consistent policy enforcement and audit. Advanced-maturity programs demonstrate continuous monitoring of model invocation patterns, anomaly detection that identifies deviation from baseline behavior, and adaptive access controls that respond to detected anomalies. The NIST Risk Management Framework provides a system authorization process applicable to AI system identity governance that supports progression through these maturity stages (NIST, 2018c; NIST, 2020). The DOE Cybersecurity Capability Maturity Model provides a domain-specific maturity model for operational technology security programs that encompasses AI system identity governance in sectors where DOE guidance applies (DOE, 2023).

Insider threat remains one of the most consequential attack vectors in critical infrastructure, with authorised users posing risks that perimeter controls do not address (Sgandurra et al., 2017). Identity and access management frameworks that integrate role-based controls, least-privilege enforcement, and continuous monitoring provide the organisational foundation for managing both insider risk and credential-based external attacks (Mbonu et al., 2020a; Oshoba et al., 2019). Multi-factor authentication adoption models for enterprise environments address the implementation challenges of deploying stronger authentication without disrupting operational workflows (Oshoba et al., 2021). Cloud identity and access governance optimisation extends these principles to hybrid and multi-cloud deployments where AI pipelines commonly reside (Mbonu et al., 2022c).

7. DATA INTEGRITY CONCERN

Validation and freshness controls ensure that data used in retrieval reflects current and authoritative content. Practitioners enforce explicit content lifecycle management, structured update procedures, and integration with broader content management infrastructure. Stale or invalid content can produce inaccurate model outputs with operational consequences (Kumuyi, 2024a; Sanni et al., 2024b).

Sensitive data handling addresses the protection of confidential information that enters training, retrieval, or prompt contexts. Core controls include data classification schemes, structured controls on sensitive data ingestion, and explicit attention to data minimization in retrieval contexts (Sanni et al., 2024a). Sensitive data in model contexts presents distinct risks from sensitive data in conventional systems.

Auditable traces of model invocations support both security analysis and broader compliance considerations. Audit-ready deployments incorporate structured logging of inputs, outputs, retrieved content, and model versions, with retention sufficient to support investigation. The traces should be protected against tampering and should integrate with broader audit infrastructure.

Vector database security (controlling access to the indexed embedding store that supplies context to the language model) in retrieval-augmented pipelines addresses the storage, retrieval, and integrity of the embedding representations that support semantic search. Recommended controls

encompass access restrictions on the vector database, integrity protection of stored embeddings, and structured attention to the privacy implications of embedding representations.

Provenance tracking for training, retrieval, and prompt data addresses the fundamental question of where data originates and whether it should be trusted at the trust level implied by its position in the model context. Data that reaches a model as part of the system prompt is typically treated as highly trusted, while data in the human turn may receive lower trust. Retrieval-augmented data occupies an intermediate position that requires explicit trust assignment. The ISO/IEC 27001 information security management framework provides a structure for data classification that can be extended to AI data provenance requirements (ISO, 2022a). ISO/IEC 27005 provides risk assessment guidance applicable to the evaluation of data provenance risks in AI contexts (ISO, 2018b).

Furthermore, validation and freshness controls address the temporal dimension of data integrity. Stale data in retrieval indexes can produce outdated model outputs with operational consequences, particularly in critical infrastructure contexts where procedures, configurations, and regulatory requirements change over time. Freshness assurance requires content lifecycle management that enforces review cycles for indexed documents, structured update procedures that re-validate content when source documents change, and freshness metadata that enables filtering of outdated retrieval results (NIST, 2020; DOE, 2022). The National Institute of Standards and Technology Security and Privacy Controls provide a baseline for information integrity controls that apply to AI data management (NIST, 2020).

However, sensitive data handling in AI contexts requires attention to several dynamics that do not arise in conventional data management. Language models may memorize and reproduce training data, creating risks that persist even when access controls prevent direct retrieval of source documents (Carlini et al., 2021). retrieval-augmented systems that access sensitive documents create risks of inadvertent disclosure through generated outputs. Embedding representations in vector databases encode semantic content in ways that may reveal source document characteristics to adversaries who can query the embedding space. The NIST SP 800-53 Revision 5 system and communications protection and access enforcement controls address technical mechanisms for sensitive data protection applicable to AI systems (NIST, 2020). ISO/IEC 27002 provides implementation guidance for sensitive data controls that can be applied to AI data contexts (ISO, 2022b).

Auditable traces of model invocations serve, moreover, multiple governance purposes. For security analysis, traces enable detection of anomalous patterns that may indicate adversarial activity, including systematic probing that suggests extraction attempts or unusual output patterns that may indicate prompt injection success. For compliance, traces document that AI-assisted decisions were made under appropriate identity and authorization controls. For incident response, traces provide the forensic record needed to reconstruct the context of a compromised invocation and assess the scope of impact. The IEC 62443 standards for industrial automation and control system security provide audit logging requirements applicable to operational technology contexts that can inform AI invocation logging design (IEC, 2018). The NERC CIP standards impose specific logging and record retention requirements for bulk electric system operations that AI systems operating in that sector must satisfy (NERC, 2023).

Vector database security in retrieval-augmented pipelines requires application of security controls to a relatively novel infrastructure category. Vector databases store embedding representations that

support semantic similarity search, enabling retrieval of content semantically related to query embeddings. Access controls on the vector database govern which users and agents can retrieve which document segments, implementing a form of information retrieval access control that complements document-level access controls. Integrity protection of stored embeddings detects adversarial modification that could redirect retrieval toward attacker-controlled content. Privacy considerations arise from the semantic content encoded in embeddings, which may be partially recoverable through query analysis.

Cloud computing introduces a distinct set of security challenges encompassing data residency, multi-tenancy isolation, shared infrastructure risks, and the evolving threat landscape facing cloud-hosted applications and data stores (Coppolino et al., 2017; Singh & Chatterjee, 2017; Tabrizchi & Kuchaki Rafsanjani, 2020). Mobile edge and fog computing architectures extend these challenges to distributed processing nodes closer to operational technology environments, requiring security controls adapted to resource-constrained edge platforms (Roman et al., 2018). Analysis of security issues specific to cloud-based architectures provides a structured basis for evaluating the cloud security properties of retrieval-augmented generation deployments (Hashizume et al., 2013).

Freshness assurance for retrieval-augmented systems requires content lifecycle management that extends beyond initial indexing to encompass the full operational life of each document in the retrieval store. Operational procedures in critical infrastructure environments are subject to revision as regulatory requirements change, as equipment configurations are updated, and as incident reports generate new guidance. A model that retrieves an outdated version of a procedure in a time-critical operational context may produce advice that was accurate at time of indexing but is now incorrect or unsafe. Consequently, lifecycle controls should include mandatory review triggers whenever source documents change, automated staleness detection that flags segments whose underlying source has been updated, and retrieval-time filtering that prevents the model from presenting outdated content without explicit staleness warnings (NIST, 2020; ISO, 2022a).

Vector database security encompasses controls on the embedding index itself, not only on the documents from which embeddings are derived. An adversary who can inject embeddings directly into the index, bypassing the ingestion pipeline, can cause the retrieval system to preferentially return attacker-controlled content without that content having passed through document-level integrity checks. Furthermore, an adversary who can query the embedding space systematically may recover approximate reconstructions of indexed document content through embedding inversion techniques, creating a data exfiltration risk independent of direct document access. Mitigations include cryptographic integrity protection of stored embeddings, access controls on the embedding insertion API that restrict which systems can modify the index, and anomaly detection on embedding query patterns that may signal extraction attempts (ISO, 2022b; NIST, 2020; Carlini et al., 2021).

8. ADVERSARIAL RISK CONCERN

Prompt injection mitigation requires explicit architectural attention to the boundaries between trusted instructions and untrusted content. Defence-in-depth requires structured input handling that

separates system instructions from user input, output validation that constrains acceptable model behavior, and explicit attention to the privilege boundaries between content of different trust levels.

Output filtering applies structured controls to model outputs before they reach downstream systems or users. Output-layer controls encompass content classification, sensitive information detection, and policy-based controls on actions that the output could trigger. Output filtering complements rather than replaces upstream controls.

Red teaming for generative artificial intelligence systems extends traditional red team practice to include model specific attacks. Red team exercises include structured assessment of prompt injection robustness, evaluation of jailbreak resistance, and end to end scenario testing of agentic capabilities. Red teaming should be a continuous activity rather than a one-time assessment.

Defense in depth across the adversarial risk concern combines multiple complementary controls. No single control is sufficient against the breadth of adversarial techniques, but the combination produces meaningful defense .

Technical depth in critical infrastructure cybersecurity continues to advance through both academic research and operational experience. New methodologies, new tooling, and new threat tradecraft emerge continuously. Practitioner programs must balance investment in advanced capability against the foundational practices that comprise the bulk of effective defense. The proper balance varies across entities and across time as the threat environment and the available capability evolve.

Integration patterns across the architectural and procedural elements discussed in this work require explicit organizational attention. Technical capability without organizational integration produces limited operational value. Organizational patterns that support integration include cross functional governance forums, structured information sharing across teams, and explicit attention to the cultural dimensions of cybersecurity practice (Patrick et al., 2020). Mature programs demonstrate sustained organizational integration alongside technical capability.

Measurement of program effectiveness across the dimensions discussed in this work supports both internal management and external communication. Measurement practices have matured during the corpus period, with growing consensus on which measures provide meaningful signal. Common pitfalls include the use of activity measures rather than outcome measures, the conflation of compliance demonstration with substantive effectiveness, and the absence of context that supports interpretation of specific measures.

Consequently, prompt injection mitigation requires layered architectural controls that address the attack at multiple points in the processing pipeline. Input-level controls structure the model context to separate system instructions from user inputs, making the instruction-content boundary explicit and resistant to manipulation through user-controlled content. Structured input schemas that constrain the format and content of user inputs reduce the attack surface available to prompt injection attempts. Privilege separation in the model context assigns different trust levels to system instructions, retrieved content, and user inputs, preventing lower-privileged content from asserting the authority of higher-privileged content (Greshake et al., 2023). The OWASP Top Ten for Large Language Models addresses prompt injection as the most critical vulnerability in current LLM deployments, providing implementation guidance that complements the architectural controls described here.

Output-level controls validate model outputs before they reach users or downstream systems. Content classification detects outputs that deviate from expected patterns, including outputs that appear to execute injected instructions rather than respond to the intended query. Policy-based controls constrain the actions that outputs can trigger, limiting the downstream consequences of successful prompt injection. Regular expression and semantic filtering detect sensitive information patterns that should not appear in outputs, providing a last-resort control for output exfiltration attempts. The defense-in-depth principle requires that output-level controls complement rather than replace input-level and context-level controls, recognizing that no single control is sufficient against the range of prompt injection variants (NIST, 2024a; Carlini et al., 2021).

Furthermore, jailbreak resistance addresses adversarial attempts to bypass safety measures by inducing models to generate harmful, policy-violating, or unauthorized content (Wei et al., 2023; Liu et al., 2023). Jailbreak attacks exploit inconsistencies in model safety training, edge cases in the representation of safety policies, and creative prompt constructions that cause models to adopt personas or framings that bypass safety constraints. The research literature documents a persistent competition between jailbreak techniques and model safety improvements, with new jailbreak patterns emerging continuously as models evolve (Wei et al., 2023; Weidinger et al., 2022). Architectural controls that do not rely on model behavior alone are essential, given the demonstrated fragility of purely model-based safety approaches. Input filtering that detects known jailbreak patterns, output monitoring that detects safety-violating content, and human review of flagged outputs provide defense that does not depend solely on model safety training.

Additionally, red teaming for generative AI systems integrates prompt injection testing, jailbreak testing, retrieval poisoning simulation, and end-to-end scenario testing of agentic capabilities (Apruzzese et al., 2019). Red team exercises should include both automated scanning using known attack patterns and manual adversarial testing by human red teamers with knowledge of the specific deployment context. The Cyber Incident Reporting for Critical Infrastructure Act creates legal obligations for reporting significant cyber incidents that should inform red team exercise scope and documentation practices. The CISA Shields Up guidance provides direction on heightened defensive posture during elevated threat periods that encompasses AI system security hardening (CISA, 2022). Red team findings should be integrated into a continuous improvement cycle that tracks remediation of identified weaknesses and documents residual risk for risk acceptance by appropriate governance bodies.

Defense in depth across the adversarial risk concern requires explicit documentation of the control set and the attack scenarios it addresses. No individual control addresses the full range of adversarial techniques applicable to generative AI systems. The combination of input validation, trust boundary enforcement, output filtering, rate limiting, anomaly detection, and red team validation produces meaningful defense that degrades adversary capability across the attack spectrum. The NIST Cybersecurity Framework Protect function provides a structure for organizing these controls within a broader cybersecurity program (NIST, 2024b). The IEC 62443 defense-in-depth principle for industrial automation and control system security provides a sector-specific expression of this general principle applicable to AI-assisted operational technology (IEC, 2018).

Thus, defence in depth for adversarial risk management combines controls at the input, model, and output layers of the generative AI pipeline with organizational controls including red teaming, incident response, and governance oversight. The principle that no single control is sufficient against the breadth of adversarial techniques is well established in the cybersecurity literature.

Applied to generative AI, this principle implies that organizations should not rely solely on model safety training to prevent jailbreaks, solely on input filtering to prevent prompt injection, or solely on access controls to prevent output exfiltration. The combination of controls across layers produces defense that degrades adversary capability even when individual controls are partially circumvented.

Technical depth in generative AI security continues to advance through academic research on both attack techniques and defenses. The NIST AI 100-2 taxonomy of adversarial machine learning attacks provides a structured framework for understanding the attack landscape against AI systems at training time and inference time (Vassilev et al., 2024). The OWASP Top Ten for Large Language Models translates academic research on LLM vulnerabilities into practitioner-accessible guidance. The intersection of classical adversarial machine learning research with the novel properties of large language models is producing a growing body of work on prompt injection, jailbreaking, and data extraction that practitioners must monitor as the field evolves (Greshake et al., 2023; Wei et al., 2023; Carlini et al., 2021; Weidinger et al., 2022). Integration of academic research findings into operational security practice requires structured processes for monitoring the research literature, evaluating the applicability of new findings to the deployment context, and updating controls in response to new attack patterns.

Physical-world adversarial attacks demonstrate that the vulnerability of neural networks extends beyond the digital domain to manipulable physical objects such as traffic signs and road markings, posing concrete risks for autonomous systems operating in critical infrastructure environments (Eykholt et al., 2018; Kurakin et al., 2017). Internet of Things deployments in critical infrastructure expand the attack surface by connecting field devices with limited security capabilities to supervisory networks, requiring layered security controls spanning device authentication, communication encryption, and network segmentation (Bertino & Islam, 2017; Sicari et al., 2015).

9. GOVERNANCE AND EVALUATION

Furthermore, policy oversight of generative artificial intelligence deployment addresses authorization, monitoring, and lifecycle management. Governance programmes establish explicit authorization gates for production deployment, ongoing monitoring of model behavior and incident reports, and structured retirement procedures for models no longer in production use (NIST, 2018a). The oversight should integrate with broader governance practice rather than operating in isolation.

Evaluation approaches for generative artificial intelligence systems include capability evaluation, safety evaluation, and security evaluation. Evaluation practice includes structured benchmark testing, adversarial evaluation, and operational pilot testing. Evaluation should address both expected and adversarial conditions (Genge et al., 2017).

Incident response procedures for generative artificial intelligence systems extend traditional incident response to include model specific scenarios (Schinagl et al., 2015; Hahn et al., 2015; Adebayo, 2025c). Incident response programmes include playbooks for prompt injection incidents, response procedures for output exfiltration discoveries, and structured coordination with broader incident response (Idika et al., 2025).

Governance integration with broader artificial intelligence governance supports consistent practice across the organization. Mature programmes develop unified artificial intelligence policy, structured authorization processes, and integration with broader risk management.

Policy oversight of generative artificial intelligence deployment requires explicit governance structures that authorize AI systems for production use, monitor their behavior during operation, and manage their retirement at end of useful life. Authorization gates for production deployment should verify that AI systems have completed security evaluation, that controls for each of the three concerns are in place and tested, and that the operational team has received appropriate training. The NIST AI Risk Management Framework Govern function addresses organizational accountability for AI system performance and risk (NIST, 2023). The DOE Cybersecurity Capability Maturity Model provides a maturity framework for operational technology security programs that encompasses AI system governance (DOE, 2023). The CISA Cybersecurity Performance Goals provide a baseline of cybersecurity practices applicable to critical infrastructure entities that should inform AI deployment authorization criteria (CISA, 2023a).

Additionally, evaluation approaches for generative AI systems span capability, safety, and security dimensions. Capability evaluation assesses whether the system performs the intended task with acceptable quality. Safety evaluation assesses whether the system produces harmful or unreliable outputs under adversarial or edge-case conditions. Security evaluation assesses the resistance of the system to the attack classes documented in the threat landscape analysis, including prompt injection, retrieval poisoning, output exfiltration, and supply chain compromise. The NIST Adversarial Machine Learning taxonomy provides a structured approach to security evaluation that covers attacks against AI systems at training time, inference time, and supply chain (Vassilev et al., 2024). The NIST AI 600-1 Generative AI Profile provides specific evaluation guidance for large language model security and safety properties (NIST, 2024a).

Evaluating the robustness of neural networks against adversarial attacks requires test suites that go beyond accuracy metrics to include attack success rates under realistic adversary models (Carlini & Wagner, 2017).

Red teaming for generative AI systems extends traditional red team practice to include model-specific attacks (Apruzzese et al., 2019; Pitropakis et al., 2019). Structured adversarial assessment of prompt injection robustness evaluates whether the system's trust boundary controls withstand realistic injection attempts. Jailbreak testing evaluates whether safety measures embedded in the model or surrounding controls can be bypassed through adversarial prompting. Supply chain assessment evaluates the provenance and integrity of model artifacts and tooling dependencies. Red teaming should be integrated into deployment authorization as a pre-deployment gate and continued as an ongoing activity, given the continuous evolution of adversarial techniques (NIST, 2023).

Incident response procedures for generative AI systems must address attack scenarios that have no precedent in conventional security operations. Prompt injection incidents require identification of the malicious content, assessment of what instructions the model executed, and evaluation of what outputs were produced and where they were delivered. Retrieval poisoning incidents require identification of the adversarial content in the index, assessment of how many invocations were affected, and remediation of the poisoned index segments. Training data extraction incidents require assessment of what sensitive data may have been exposed and to whom. The Cyber Incident Reporting for Critical Infrastructure Act establishes reporting obligations for significant

cyber incidents affecting critical infrastructure that apply to AI-related incidents. The National Cybersecurity Strategy and Implementation Plan provide strategic context for incident response coordination across critical infrastructure sectors. CISA Shields Up guidance provides emergency measures that critical infrastructure entities should implement during periods of elevated cyber threat that encompass AI system risks (CISA, 2022).

Consequently, governance integration connects AI security to broader organizational risk management, information security governance, and regulatory compliance programs. Organizations subject to the NIST Cybersecurity Framework should incorporate AI system management into their framework implementation, mapping AI-specific controls to the Identify, Protect, Detect, Respond, and Recover functions (NIST, 2018b; NIST, 2024b). Organizations operating under NERC CIP standards should integrate AI system security into their CIP compliance programs, addressing the requirements of standards including CIP-007 (System Security Management) and CIP-011 (Information Protection) as they apply to AI-assisted operations (NERC, 2023). ISO/IEC 27001 certification programs should incorporate AI system controls into their information security management system scope (ISO, 2022a). The Office of Management and Budget memorandum on zero trust principles provides policy direction applicable to federal critical infrastructure operators.

The governance lifecycle for generative AI systems in critical infrastructure does not end at deployment authorization. Continuous monitoring of model behavior, periodic review of architectural decisions against the evolving threat landscape, and incident-triggered assessments that evaluate whether controls would have prevented or detected recent adversarial events are essential elements of sustained AI security governance. Organizations that treat AI security as a one-time deployment gate rather than a continuous program discipline will find their controls degrading as threat actor capabilities advance and as model behavior evolves through updates, fine-tuning, and changes in retrieval content (NIST, 2023; CISA, 2023a; DOE, 2023).

Enterprise log analytics and automated incident response capabilities provide the operational infrastructure for detecting and responding to security events in AI-integrated environments (Mbonu et al., 2019b). Proactive threat intelligence models using cloud-native tooling enable earlier detection of adversarial activity targeting AI systems and the operational technology networks they support (Oshoba et al., 2023b). Blockchain-enabled compliance and audit trail models for cloud configuration changes provide tamper-evident records applicable to AI system governance (Oshoba et al., 2020b). Privacy-by-design principles embedded in system architecture reduce the risk of sensitive data exposure through AI-generated outputs (Okoruwa et al., 2020). Metadata-driven access controls and role-based design patterns directly inform the identity concern architecture for AI pipeline access management (Olatunde-Thorpe et al., 2020).

Governance programmes that address the intersection of artificial intelligence with data privacy obligations in healthcare and digital health platforms provide models transferable to other high-consequence sectors where personal and sensitive data are processed by AI systems (Ekechi, 2025b).

10. REFERENCE SCENARIOS

Each scenario is analysed with reference to the three concerns summarised in Table 1, demonstrating how the framework applies to the distinct operational contexts, consequence profiles, and regulatory environments characteristic of each sector.

Scenario one addresses energy sector deployment of a retrieval-augmented generation system for operational documentation and procedure assistance. The scenario walks through the application of the three-concern architecture to a representative deployment, including identity controls for operator access, data integrity controls for procedure content, and adversarial risk controls for the model interaction (Vassilev et al., 2024; NIST, 2024a; Sarker, 2024). The scenario demonstrates the architecture in a context that exercises all three concerns.

Scenario two addresses water utility deployment of a generative assistant for incident response support. The scenario examines the use of the architecture in a higher stakes context where model output may inform decisions affecting public health (Liadi, 2024b). Implementation considerations include enhanced output validation, explicit human approval for consequential decisions, and structured logging for post incident review (Stouffer et al., 2023; NERC, 2023).

Scenario three addresses emergency response deployment of a generative system for situation summarization and resource coordination (DOE, 2023; Wei et al., 2023; Chen et al., 2023). The scenario examines the use of the architecture in a context with significant time-pressure and consequence sensitivity (Liu et al., 2023; NIST, 2023). Implementation considerations include rapid identity verification, integrity controls for the information sources feeding the system, and explicit attention to the limits of model output reliability in unprecedented situations.

Cross cutting lessons from the scenarios include the importance of structured authorization, the value of human oversight for consequential actions, and the operational considerations that distinguish critical infrastructure deployment from generic enterprise practice.

The convergence of these findings across multiple analytical dimensions strengthens confidence in the synthesis presented in this work. Each dimension provides independent evidence that supports the core conclusions (CISA, 2022). The convergence is particularly meaningful because the dimensions draw on distinct evidence sources, distinct methodologies, and distinct stakeholder perspectives (DOE, 2022; ISO, 2022a; ISO, 2022b). Cross dimensional validation of this kind is recognized as a strength of mixed methods research and supports the practical applicability of the conclusions (Weidinger et al., 2022).

Counter examples and edge cases warrant explicit acknowledgment alongside the central findings. The literature includes examples of entities, vendors, and methodologies that do not fit the central patterns identified in the synthesis. These counter examples may reflect early-stage practice, unusual operational contexts, or methodological choices that differ from the dominant patterns. Their existence does not undermine the central findings but rather illustrates the diversity of practice that any synthesis must acknowledge (Bommasani et al., 2021).

Methodological reflections on the development of this work include both the strengths and limitations of the chosen approach (Carlini et al., 2021). Strengths include the breadth of source material, the integration across multiple authoritative frameworks, and the orientation toward both research and practitioner audiences. Limitations include the conceptual rather than empirical character of much of the analysis, the primary orientation toward North American context, and the

dependence on the availability and quality of published evidence. Acknowledging these methodological characteristics supports informed use of the contributions.

The energy sector reference scenario examines deployment of a retrieval-augmented generation system to support operations personnel at a bulk electric system facility. Operators query the system for procedure guidance, regulatory interpretation, and operational history relevant to current system conditions. The system retrieves content from a curated knowledge base of internal procedures, vendor documentation, and regulatory guidance, augmenting this retrieved content with the general knowledge of the foundation model. Identity controls authenticate operator access using multifactor authentication integrated with the facility identity provider, establish role-based authorization that limits retrieval scope to documents relevant to each operator role, and log all model invocations with identity binding (NIST, 2020; NERC, 2023). Data integrity controls enforce document review cycles for all indexed procedures, verify the provenance of regulatory guidance against authoritative sources, and apply freshness filters that exclude outdated content from retrieval. Adversarial risk controls implement structured input handling that isolates operator queries from retrieved content in the model context, apply output filtering that detects potentially injected instructions, and conduct regular red team exercises targeting the retrieval knowledge base for poisoning vulnerabilities (NIST, 2024a; CISA, 2023a).

The water utility scenario examines deployment of a generative AI assistant for incident response support during operational anomalies. The scenario involves higher consequence sensitivity than the energy scenario because model outputs may inform decisions that affect public health. Identity controls are supplemented by mandatory secondary confirmation of operator identity for queries tagged as high-consequence. Data integrity controls require that all content retrieved in support of high-consequence decisions carry provenance metadata documenting the authorizing review and the date of last verification. Adversarial risk controls include enhanced output validation that requires human review of outputs that exceed a consequence threshold before those outputs are acted upon. Structured logging captures the complete context of each invocation for post-incident analysis (NIST, 2023; DOE, 2023). The CISA Zero Trust Maturity Model informs the implementation of identity controls that adapt dynamically to the consequence level of each operator query (CISA, 2023b).

The emergency response scenario examines deployment of a generative AI system for situation summarization and resource coordination during multi-agency emergency response to a critical infrastructure incident. The scenario is characterized by extreme time pressure, novel and rapidly evolving operational context, and the involvement of multiple organizations with distinct identity governance structures. Rapid identity verification protocols that can authenticate personnel from multiple agencies against federated identity infrastructure enable cross-organization access without compromising audit trail completeness. Integrity controls for information sources feeding the system recognize that emergency response information flows may be less curated than routine operational data, applying explicit uncertainty signaling when retrieved content lacks the provenance metadata of verified organizational sources (Stouffer et al., 2023; NIST, 2024a). Output reliability controls provide explicit confidence indicators on model outputs and require human validation of time-critical decisions before operational action.

11. DISCUSSION

11.1. Architectural Contribution and Strengths

Discussion of the architecture indicates that its three-concern structure provides a coherent organizing framework for the diverse security considerations of generative artificial intelligence in critical infrastructure (NIST, 2020; Souppaya et al., 2020; Rose et al., 2020). The architecture is compatible with diverse organizational contexts and deployment patterns provided that the underlying program elements are present in proportionate strength (Da Veiga et al., 2020). The architecture is forward looking, anticipating the maturation of generative artificial intelligence practice over the coming years.

Therefore, future directions include empirical validation through case study, integration with emerging governance frameworks, and the development of more detailed implementation guidance for specific architectural elements (Yuan et al., 2019). The intersection of generative artificial intelligence security with operational technology security represents a particularly active frontier (Pitropakis et al., 2019; Lewis, 2019; Devlin et al., 2019). Continued evolution of adversary capability against generative systems will require continued evolution of frameworks such as the one presented in this paper (NIST, 2018c).

This conceptual paper has presented an architecture for securing generative artificial intelligence pipelines in critical infrastructure environments, organized around three concerns of identity, data integrity, and adversarial risk (IEC, 2018; ISO, 2018b; ISO, 2018a; Stellios et al., 2018). The architecture aligns with multiple authoritative frameworks and is illustrated through three reference scenarios (Madry et al., 2018). It supports the responsible adoption of generative artificial intelligence in critical infrastructure environments (Chakraborty et al., 2018).

Stakeholder feedback through the development of this work has consistently emphasized the practical considerations that distinguish operational technology cybersecurity from generic information technology practice. These considerations include long asset lifecycles, deterministic performance requirements, safety integration, and the operational and business pressures that affect security investment decisions (Lin et al., 2017). The frameworks and architectures developed without explicit attention to these considerations tend to encounter implementation friction that limits their practical adoption (Voigt et al., 2017; Papernot et al., 2016; Almorsy et al., 2016). Sustained engagement with operational stakeholders, beginning early in framework development, supports the development of frameworks that align with the realities of the operational environment (Stouffer et al., 2015; Goodfellow et al., 2015).

Workforce considerations represent a recurring topic in practitioner discussions of operational technology cybersecurity. The talent shortage in operational technology cybersecurity is well documented in industry surveys and government reports (Szegedy et al., 2014; Mo et al., 2012; Nicholson et al., 2012). Training programs, certification frameworks, and academic curricula are adapting to address the talent gap, but the pace of adaptation has lagged the growth in demand (NIST, 2011). The combination of cybersecurity expertise with engineering expertise produces a profile that is rare in the current workforce, and developing this profile through targeted training and structured career pathways remains an industry priority (Ten et al., 2010; Khurana et al., 2010).

International coordination on critical infrastructure cybersecurity expectations has grown through the corpus period (Igure et al., 2006). Bilateral and multilateral exchanges support shared learning about regulatory practice, threat intelligence, and incident response coordination. Harmonization

of expectations across jurisdictions remains incomplete, but the trend toward convergence on principles such as risk-based assessment, segmentation, and continuous monitoring is evident. These developments inform the broader context within which the frameworks and architectures developed for North American operations must operate.

The relationship between cybersecurity considerations and broader resilience considerations requires explicit attention. Cybersecurity is one input to operational resilience, alongside physical security, supply chain resilience, workforce resilience, and operational risk management. Effective programs integrate cybersecurity into broader resilience frameworks rather than treating it in isolation. The integration produces both stronger overall resilience and more efficient use of organizational resources.

The three-concern architecture presented in this paper provides a foundation for responsible generative AI adoption in critical infrastructure, but several research and practice gaps warrant further attention. The empirical evidence base for the effectiveness of specific architectural controls remains limited. Published research has demonstrated the feasibility of prompt injection, retrieval poisoning, and training data extraction attacks, but the effectiveness of specific countermeasures has been evaluated less systematically (Greshake et al., 2023; Carlini et al., 2021; Wei et al., 2023). Controlled studies comparing architectural variants, red team exercises that benchmark control effectiveness, and operational pilots with structured evaluation designs would strengthen the evidence base for practitioner adoption.

11.2. Integration with Operational Technology Security

However, the integration of generative AI security with operational technology security remains an active challenge. The operational technology security community has developed mature practices for industrial control system security through standards including NERC CIP and IEC 62443 (NERC, 2023; IEC, 2018). The integration of AI systems into operational workflows creates hybrid security environments where generative AI security controls must compose with operational technology controls designed for deterministic systems with known behavior. The National Cyber Informed Engineering Strategy provides a framework for integrating security into engineering decisions throughout the technology lifecycle that is applicable to this integration challenge (DOE, 2022). Research on the composition of generative AI security controls with operational technology security controls would support practitioners navigating this integration.

11.3. Governance, Regulatory, and Compliance Landscape

Furthermore, the regulatory environment for generative AI in critical infrastructure continues to evolve. The NIST AI Risk Management Framework and Generative AI Profile provide voluntary guidance that may inform future regulation (NIST, 2023; NIST, 2024a). The European Union AI Regulation establishes binding requirements that apply to organizations operating in European markets. The Cyber Incident Reporting for Critical Infrastructure Act creates mandatory incident reporting obligations that apply to significant AI-related incidents. The National Cybersecurity Strategy directs federal agencies to work with critical infrastructure sectors to develop sector-specific AI security guidance. Practitioners benefit from sustained engagement with the regulatory evolution process, as the architecture presented in this paper will require adaptation as regulatory requirements mature.

11.4. Workforce, Education, and Organisational Readiness

Moreover, the workforce dimension of generative AI security in critical infrastructure presents challenges that technical architecture alone cannot address. The combination of cybersecurity expertise, operational technology domain knowledge, and generative AI system understanding required to implement and operate the architecture described here is rare in the current workforce (NIST, 2011). Training programs that develop this combined competency, certification frameworks that recognize it, and organizational structures that create career pathways for AI security practitioners in critical infrastructure will be essential to scaling the adoption of the architecture. The operational technology cybersecurity talent shortage documented in industry surveys (Nicholson et al., 2012; Mo et al., 2012) will deepen as generative AI creates additional demand for this specialized expertise. Academic programs that integrate AI security with operational technology security, and continuing education programs that upskill existing practitioners, represent high-priority investments for the field.

The intersection of generative AI security with organizational change management deserves explicit acknowledgment in any practical adoption framework. Security architectures that are technically sound but organizationally difficult to implement will be implemented inconsistently, producing uneven protection that adversaries can exploit. The human factors of AI security include operator trust in AI outputs, which may be calibrated at inappropriate levels in either direction, organizational cultures that treat AI security as an impediment to adoption rather than an enabler of responsible deployment, and incentive structures that reward AI capability demonstration over AI security rigor. Security architectures that integrate with existing operational workflows rather than competing with them, that provide security assurance without adding friction to routine operations, and that produce evidence of security effectiveness that organizations can communicate to regulators and stakeholders, will achieve higher adoption rates and more consistent implementation than architectures that impose significant operational overhead.

Additionally, the international dimension of generative AI security in critical infrastructure reflects the global nature of both the technology supply chain and the adversary landscape. Foundation models are developed and deployed by a small number of organizations, with the geopolitical implications of this concentration receiving increasing regulatory attention. Critical infrastructure operators in multiple jurisdictions may use the same foundation models, creating a common attack surface that adversaries can study and exploit across multiple targets. Threat intelligence sharing across jurisdictions, coordinated regulatory approaches that do not create exploitable divergences in security requirements, and international standards development that produces globally applicable guidance rather than jurisdiction-specific frameworks, are dimensions of critical infrastructure AI security that extend beyond the organizational scope of this paper but that shape the environment in which organizational architectures must operate. The Williams (1994) Purdue Enterprise Reference Architecture, the Cardenas et al. (2008) characterization of security research challenges for control systems, and the Igure et al. (2006) analysis of SCADA security issues all reflect a foundational international consensus on critical infrastructure security principles that the generative AI era must extend rather than discard.

Furthermore, the vendor ecosystem for generative AI systems presents challenges for critical infrastructure security governance analogous to those presented by the operational technology vendor ecosystem. Foundation model providers, retrieval infrastructure vendors, orchestration framework developers, and fine-tuning service providers each introduce supply chain risks that the architecture must address. Vendor security assessment processes that evaluate the security

practices of AI technology providers, contractual requirements that establish security obligations and audit rights, and participation in industry consortium activities that develop sector-specific vendor security expectations, are organizational practices that support supply chain risk management for AI systems (NIST, 2020). The NERC CIP supply chain risk management standards, which address the security of vendors to the bulk electric system, provide a model for sector-specific AI vendor security requirements that other critical infrastructure sectors may adapt (NERC, 2023). Sustained organizational commitment to AI vendor security management, beginning in the procurement process and continuing through the operational lifecycle, produces supply chain security posture that one-time assessments cannot achieve.

However, the scalability of the three-concern architecture across different organizational sizes and capability levels deserves explicit consideration. Large critical infrastructure asset owners with mature cybersecurity programs and dedicated security teams will implement the full architecture, integrating AI security governance into established program structures with dedicated engineering and operational resources. Medium-sized operators with developing cybersecurity programs may prioritize the controls most directly relevant to their deployment scenarios, building toward full architecture implementation over a defined roadmap. Smaller operators with limited security resources may begin with foundational controls, particularly strong identity governance and basic audit trail capabilities, and expand toward more sophisticated adversarial risk controls as capability matures. The architecture's alignment with the NIST Cybersecurity Framework supports this tiered adoption by providing a common reference that practitioners at all maturity levels can use to orient their programs (NIST, 2024b).

The temporal dynamics of generative AI security require that the architecture be treated as a living governance artifact rather than a static compliance document. Model versions change, retrieval indexes are updated, tooling dependencies evolve, and threat actor capabilities advance. Governance structures must provide for regular review of architectural decisions against the current threat landscape, evaluation of new controls as they become available, and retirement of controls that have become ineffective or have been superseded by more effective alternatives. Scheduled architectural reviews, triggered by significant changes in the model or deployment environment, and incident-driven reviews that evaluate whether existing controls would have detected or prevented recent incidents, are the principal mechanisms through which architecture currency is maintained (NIST, 2023; Vassilev et al., 2024).

Practitioner guidance derived from the architecture should be expressed in formats accessible to the diverse roles involved in critical infrastructure AI deployment. Security architects require framework-aligned control specifications that integrate with existing security program documentation. Operational technology engineers require implementation guidance expressed in engineering terms, addressing the specific integration patterns of AI systems with supervisory control and data acquisition environments. Legal and compliance teams require regulatory mapping that connects architectural controls to applicable requirements. Executive sponsors require risk-oriented summaries that articulate the residual risk accepted when specific controls are or are not implemented. The NIST AI Risk Management Framework Playbooks provide a model for multi-audience guidance documentation that the architecture described here should emulate in practitioner-facing outputs (NIST, 2023).

The contribution of this paper to the practitioner community is intended to be cumulative with, rather than competitive with, existing guidance from authoritative bodies. The NIST AI Risk

Management Framework, the CISA Cybersecurity Performance Goals, the DOE Cybersecurity Capability Maturity Model, and the NERC Critical Infrastructure Protection standards all provide valuable guidance that practitioners must integrate into coherent programs (NIST, 2023; CISA, 2023a; DOE, 2023; NERC, 2023). The three-concern architecture provides a structural integration layer that helps practitioners understand how AI-specific controls relate to existing program elements, how the new standards and guidance documents relate to each other, and how to prioritize implementation given resource constraints. The goal is not to add to the burden of compliance but to provide a conceptual map that makes the existing landscape of requirements and guidance more navigable for practitioners responsible for critical infrastructure AI security.

11.5. Cross-Sector and International Dimensions

The relationship between security and safety in generative AI systems deployed in critical infrastructure warrants explicit acknowledgment. Safety in the operational technology sense concerns the protection of physical processes, personnel, and the public from harm arising from system malfunction or misuse. Security in the cybersecurity sense concerns the protection of systems from adversarial interference. These two concerns are related but distinct: a prompt injection attack that causes a model to provide incorrect procedure guidance may produce safety consequences through the downstream actions of operators who act on that guidance, even if no direct control system action is involved. The architecture presented here addresses security, but practitioners implementing AI systems in critical infrastructure must also address safety through the application of safety engineering principles, formal hazard analysis, and safety-critical system design practices appropriate to their specific deployment contexts. The integration of AI security and AI safety is an emerging research and practice challenge that the broader field must address.

The educational dimension of generative AI security in critical infrastructure extends beyond the workforce development challenges discussed above to encompass the need for new curricula, professional certification frameworks, and continuing education programs specifically targeting AI security in operational technology environments. University programs in cybersecurity have generally not yet integrated operational technology security or AI security to the degree required to prepare graduates for the combined competency that critical infrastructure AI security demands. Professional certifications in operational technology security, while growing in availability, similarly have not yet addressed generative AI security in depth. The development of educational pathways that integrate generative AI security with operational technology cybersecurity represents a foundational investment for the field that will take sustained effort from academic institutions, professional bodies, and industry to realize.

11.6. Research Agenda and Future Directions

The open research questions identified by this work suggest a productive agenda for academic-practitioner collaboration. Empirical studies of the effectiveness of specific prompt injection mitigations, controlled by threat model and deployment context, would provide evidence-based guidance that is currently unavailable. Formal models of retrieval-augmented generation security properties would support more rigorous security evaluation than current informal approaches permit. Human factors research on operator calibration of AI output trust, and the conditions under which operators over-rely on or under-utilize AI decision support, would inform the design of human-AI interaction patterns that support safe and secure operations. Longitudinal studies of AI security incident patterns in critical infrastructure, enabled by improved incident reporting and sharing, would provide the operational experience base that architecture and control development

requires. This paper invites engagement from researchers across these domains to advance the evidence base for critical infrastructure AI security.

A comprehensive framework for migrating to zero trust architecture documents the organizational and technical challenges of implementation, providing practical guidance applicable to AI system integration within zero trust programs (Phiayura & Teerakanok, 2023).

The foundational characterization of prompt injection as an attack class demonstrated that language models can be caused to ignore system instructions through adversarial user inputs, establishing the core threat that identity and context separation controls must address (Perez & Ribeiro, 2022).

Information security practice in critical infrastructure organizations depends on a body of foundational knowledge encompassing cryptographic controls, network security architecture, access management, and risk governance (Stallings, 2017; Stamp, 2011; Pfleeger et al., 2015). Practitioner-oriented texts and certification reference guides codify this foundational knowledge in formats accessible to security teams responsible for AI governance (Kim & Solomon, 2018; Ciampa, 2018; Andress, 2014; Diogenes & Ozkaya, 2018). The emergence of artificial intelligence as a security concern has prompted academic analysis of the relationship between AI capability and cyber risk, including recognition that AI systems may be leveraged both offensively and defensively (Khisamova et al., 2019).

Internet of Things security encompasses the full stack of device, network, and application controls required to protect the sensor endpoints and edge devices that feed data into AI-assisted operational systems in critical infrastructure (Mosenia & Jha, 2017).

Securing cyber-physical systems in the electric power grid requires coordinated controls across information technology and operational technology layers, a challenge made more complex by the introduction of AI-assisted monitoring and control (Sridhar et al., 2012). The 2015 Ukraine grid attack demonstrated that coordinated adversarial operations targeting smart grid infrastructure can produce physical outages at scale (Liang et al., 2017).

Smart grid security requires authentication and integrity controls across the full communication stack linking distributed energy resources, substations, and energy management systems to supervisory control platforms. Effective implementation draws on both power systems engineering practice and information security frameworks to address the unique properties of energy sector operational technology (Anwar & Soltesz, 2016).

Artificial intelligence applications in high-consequence domains including healthcare and critical infrastructure share foundational governance challenges: ensuring system outputs are reliable, explainable, and subject to meaningful human oversight (Beam & Kohane, 2018). The intersection of AI capability and clinical or operational consequence has motivated governance frameworks that balance innovation with safety, providing transferable lessons for critical infrastructure AI governance programs.

Pattern recognition methods developed for fraud detection in financial transaction systems, including anomaly detection, supervised classification, and ensemble approaches, provide methodological parallels that informed the development of AI-based intrusion detection for operational technology environments (West & Bhattacharya, 2016). Cross-domain transfer of analytical methods is a recurring pattern in the development of AI security capabilities.

Healthcare organisations deploying AI in clinical and administrative contexts face governance requirements analogous to those facing critical infrastructure operators: high-consequence of system error, stringent regulatory oversight, and the need for auditable and explainable decision support (Aminu-Ibrahim & Ogbete, 2023; Anioke & Atima, 2023a; Anioke & Atima, 2023b; Ogbete & Aminu-Ibrahim, 2024; Kuponiyi & Akomolafe, 2024; Kuponiyi & Akomolafe, 2025; Nnaji & Akinlolu, 2022; Nnaji & Akinlolu, 2023; Nnaji & Akinlolu, 2024a). Comparative analysis of AI governance frameworks across high-consequence sectors reveals both common structural elements and domain-specific adaptations that practitioners can leverage.

IT service management, financial analytics, and organisational data governance frameworks developed in adjacent professional contexts offer implementation patterns for AI governance program design in critical infrastructure, particularly in areas such as change management, performance monitoring, and risk-informed decision support (Edivri & Oteri, 2021; Edivri & Oteri, 2022; Edivri & Oteri, 2023; Lawal & Oduleye, 2021b; Oduleye & Medon, 2021; Sanni & Atima, 2021a; Sanni & Atima, 2021b; Sanni & Atima, 2021c; Sanni & Attah, 2023c; Wedraogo & Sanni, 2024; Yeboah & Nnabueze, 2021; Yeboah & Nnabueze, 2024; Michael & Ogunsola, 2022a; Michael & Ogunsola, 2024a). Adaptation of these frameworks requires attention to the distinctive operational rhythms, regulatory contexts, and consequence profiles of critical infrastructure environments.

Autonomous and semi-autonomous platforms including unmanned aerial systems that interface with infrastructure monitoring networks, perform right-of-way surveillance, and support inspection operations introduce additional identity, data integrity, and adversarial risk considerations that AI security governance frameworks must address as these platforms proliferate (Dagodzo & Ahiaeke Patrick, 2022; Dagodzo & Ahiaeke Patrick, 2025). The integration of autonomous platforms with AI-assisted data analysis creates compound attack surfaces that require coordinated security treatment.

Foundational texts on computer security and network defense provide the conceptual vocabulary that AI security practitioners bring to generative AI governance (Bishop, 2018; Stallings, 2017). Research on cyber-physical systems security in smart grid and IoT environments documents the expanded attack surface created when networked sensing, control, and analytical systems are integrated into operational technology (Humayed et al., 2017; Lopez et al., 2018). Policy and behavioral frameworks for promoting compliance with security and regulatory requirements in technology-intensive organizations offer implementation insights relevant to AI security program adoption.

The technical foundations of cybersecurity governance for critical infrastructure systems draw on a body of research spanning network security, cloud security, and threat intelligence (Baykara & Daimi, 2018; Cherdantseva & Hilton, 2013; Enisa, 2022). IoT-enabled energy systems introduce specific protocol-level vulnerabilities that require layered authentication and integrity controls extending from field devices to supervisory platforms (Sani et al., 2019; Sajid et al., 2016). Phishing and social engineering attacks targeting critical infrastructure personnel exploit human factors in ways that complement technical exploit techniques, requiring awareness training alongside technical controls as part of a comprehensive defence posture (Gupta et al., 2018; Genge et al., 2017).

The evolving threat landscape for AI-assisted critical infrastructure systems spans multiple research streams. Foundational work addresses control system security (Cardenas et al., 2008;

Cardenas et al., 2009), adversarial machine learning (Apruzzese et al., 2022; Yao et al., 2024), and adversary tactics taxonomies (Strom et al., 2018), and systematic analyses of cybersecurity investment effectiveness (Ozowara et al., 2022; Adebayo, 2022b; Adebayo, 2023a; Adebayo, 2024). AI governance frameworks for data privacy compliance (Ekechi et al., 2025) and deep learning-based malware classification for cloud environments (Idika et al., 2021) address the intersection of AI capability and security that defines the risk landscape this architecture is designed to address. Mandatory incident reporting obligations for critical infrastructure (Act, 2022), zero trust implementation policy (Budget, 2022), and practitioner vulnerability taxonomies for large language model deployments (Project, 2023) collectively constitute the governance and compliance environment within which the architecture must operate.

The classification of adversarial machine learning attacks provides essential structure for threat modeling exercises targeting AI-integrated critical infrastructure. Training-time attacks, including data poisoning and backdoor injection, corrupt model behaviour at the source before deployment; inference-time attacks, including adversarial examples and model extraction, exploit deployed models through crafted inputs (Biggio & Roli, 2018; Goodfellow et al., 2015; Madry et al., 2018). The intersection of these classical adversarial machine learning categories with the specific properties of large language models produces attack variants that differ in important respects from those applicable to classification systems. Language models are trained on text rather than structured feature vectors, making gradient-based adversarial example generation less directly applicable; however, the semantic flexibility of natural language creates attack surfaces through prompt crafting that have no analogue in classification settings (Greshake et al., 2023; Perez & Ribeiro, 2022; Wei et al., 2023). Realistic adversarial attack modeling for machine learning security systems requires evaluation under threat models that account for the practical capabilities of adversaries rather than worst-case theoretical bounds (Apruzzese et al., 2022).

The governance of agent identity in multi-agent generative AI systems presents challenges that exceed those of conventional software service account management. When an orchestrating agent invokes subordinate agents, each with access to tools and data sources, the chain of delegated authority must be traceable from the originating human principal through each agent hop to the specific action taken. Service account governance practices developed for microservices architectures provide a starting framework: each agent should have a dedicated service identity, credentials scoped to minimum required permissions, rotation schedules aligned with task durations, and activity logs associated with each credential (Oshoba et al., 2019; Oshoba et al., 2021; Mbonu et al., 2020a). The zero trust principle of never implicitly trusting any entity based on its position in the network applies with particular force to agent identities, which may be compromised through supply chain attacks on agent software or through prompt injection that redirects agent behaviour while preserving the agent identity used for authentication (Rose et al., 2020; CISA, 2023b).

The data integrity concern extends beyond the retrieval index to encompass the complete data supply chain from source document creation to model context assembly. Documents that enter the retrieval pipeline may originate from internal authoring systems, external vendor documentation, regulatory publications, or user-submitted content; each source carries distinct trust properties that must be reflected in the provenance metadata attached to indexed segments (ISO, 2022a; ISO, 2018b). Content signing using cryptographic mechanisms provides tamper-evidence for high-value documents where provenance assurance is critical to operational safety. Blockchain-enabled audit trail models applied to cloud configuration management demonstrate the feasibility of

tamper-evident record-keeping for AI governance events, including model deployment authorizations, index update approvals, and policy configuration changes (Oshoba et al., 2020b; Omoegun et al., 2022). Enterprise log analytics capabilities that aggregate telemetry from AI system components, retrieval infrastructure, and underlying cloud services provide the detection surface needed to identify data integrity violations in production deployments (Mbonu et al., 2019b).

The adversarial risk concern encompasses both the technical controls applied to the AI system and the broader security posture of the networks and systems through which AI-generated outputs flow. Industrial control systems and SCADA environments that receive guidance or control inputs informed by AI outputs represent the ultimate downstream consequence target of successful adversarial attacks against AI pipelines (Cardenas et al., 2008; Cardenas et al., 2009; Yadav & Paul, 2019). Adversarial examples crafted to persist under physical-world conditions, including sensor noise, viewing angle variation, and environmental perturbation, represent a threat class particularly relevant to AI systems that process imagery or sensor data from operational environments (Eykholt et al., 2018; Kurakin et al., 2017). Deep learning-based malware classification systems deployed to protect AI infrastructure face adversarial evasion attempts that exploit the same gradient-based vulnerabilities present in the AI systems they protect (Idika et al., 2021; Pitropakis et al., 2019). Modeling of realistic adversarial attacks against machine learning security systems demonstrates that threat models must account for adaptive adversaries who observe defences and evolve their techniques accordingly (Apruzzese et al., 2022).

Cybersecurity workforce development represents a governance prerequisite for sustained AI security programme implementation. Organisations that lack personnel with the combined competencies in generative AI systems, cybersecurity engineering, and operational technology security will struggle to implement the architectural controls described in this paper consistently over time (Adebayo, 2023a; NIST, 2011). Structured workforce development programmes that integrate applied research curricula, certification pathways, and practical laboratory experience in AI system security build the organisational capability needed for sustained programme execution. Threat detection capability assessments for resource-constrained organisations provide a maturity model framework applicable to smaller critical infrastructure operators who may not have dedicated AI security teams (Adebayo, 2024). Fintech and financial infrastructure organisations have developed security frameworks for high-volume transaction processing platforms that address threat modeling, access control, and audit trail requirements in ways transferable to AI pipeline security governance (Adebayo, 2022b). The MITRE ATT&CK framework and its ICS-specific extension provide standardised vocabulary for threat modeling exercises that enables consistent communication of AI security risks across technical, operational, and governance stakeholders (Strom et al., 2018; Alexander et al., 2020).

The governance and technical landscape for generative AI security in critical infrastructure is shaped by a dense and evolving body of standards, regulations, executive directives, and practitioner guidance. Organisations implementing this architecture opwithin a layered policy environment that includes mandatory incident reporting obligations (Act, 2022), federal zero trust implementation requirements (Budget, 2022), executive orders addressing both cybersecurity (House, 2021) and artificial intelligence governance (House, 2023a), and a rapidly developing body of sector-specific guidance from CISA (CISA, 2023a; CISA, 2023b), DOE (DOE, 2022; DOE, 2023), and NERC (NERC, 2023). The OWASP Top Ten for Large Language Model Applications provides a practitioner taxonomy of common vulnerabilities that complements the

more abstract framework structures of the NIST AI Risk Management Framework (Project, 2023; NIST, 2023). Cybersecurity predictions from industry analysts document the near-term trajectory of the threat landscape and the defensive investments that organisations should prioritise (Gartner, 2021). A comprehensive survey of large language model security and privacy documents the attack surface exposed by current-generation models in terms that directly inform the adversarial risk concern of the three-concern architecture (Yao et al., 2024).

Systematic reviews of cybersecurity investments and their relationship to organisational performance provide empirical grounding for governance decisions about AI security programme resourcing (Ozowara et al., 2022). AI governance frameworks addressing data privacy compliance across multiple regulatory jurisdictions simultaneously are increasingly necessary as critical infrastructure operators deploy AI systems that process personal data subject to GDPR, HIPAA, and sector-specific privacy requirements (Ekechi et al., 2025). The intersection of AI capability with insider threat detection, financial crime compliance, and real-time risk monitoring has generated a body of applied research in fintech and financial services that demonstrates the transferability of AI security governance frameworks across high-consequence domains (Ozowara, 2022; Adebayo, 2022b; Adebayo, 2023a).

The scalability of AI security governance across the diversity of critical infrastructure operator sizes, capability levels, and sector-specific regulatory environments requires architectural guidance that is adaptable rather than prescriptive at the control level. Small water utilities with limited security staffing face a different implementation reality from large energy utilities with dedicated operational technology security teams; yet both must address the same fundamental concerns of identity verification, data integrity assurance, and adversarial risk management when deploying generative AI in operational contexts (Maglaras et al., 2018; CISA, 2023a). Capability maturity models that define staged progression from initial to optimised security posture, such as the DOE Cybersecurity Capability Maturity Model, provide a framework for organisations to plan incremental implementation of architectural controls aligned with their current capability and resource constraints (DOE, 2023). The Cybersecurity Framework Profile concept enables sector-specific customisation of control priorities that can be applied to produce AI security implementation profiles relevant to energy, water, transportation, and emergency response sectors (NIST, 2024b). Adoption of these governance instruments supports the development of sector-wide benchmarks for AI security practice that facilitate information sharing, peer comparison, and collective improvement across the critical infrastructure community. The NERC Critical Infrastructure Protection standards demonstrate how sector-level governance frameworks can establish enforceable baseline requirements that drive consistent security practice across a large and diverse population of infrastructure operators, providing a model that AI security governance frameworks for other sectors may adapt (NERC, 2023). Investment in AI security governance at the sector level, including development of sector-specific guidance, training programmes, and information sharing mechanisms, produces collective security benefits that exceed what any individual organisation could achieve through independent programme development (Sarker, 2023; Schneier, 2018).

However, the relationship between cybersecurity frameworks and AI governance frameworks is one of complementarity rather than substitution. Organisations that implement the NIST Cybersecurity Framework as the organising structure for their security programme should integrate AI system management within that structure, mapping AI-specific controls to the Identify, Protect, Detect, Respond, and Recover functions (NIST, 2024b; NIST, 2018b). The ISO 27001 information

security management system provides a certification-backed governance structure that encompasses AI system controls within its scope when organisations choose to include AI pipelines in their management system boundary (ISO, 2022a; ISO, 2022b). The IEC 62443 series for industrial automation and control system security provides sector-specific control requirements for operational technology environments where AI systems are being integrated (IEC, 2018). Regulatory frameworks that impose mandatory security requirements on critical infrastructure operators, including NERC CIP for the bulk electric system and sector-specific frameworks in water, aviation, and financial services, establish the compliance baseline above which AI security governance must build (NERC, 2023).

The vendor ecosystem for generative AI systems presents governance challenges analogous to those addressed by supply chain risk management frameworks developed for conventional software and operational technology (Strom et al., 2018). Foundation model providers, retrieval infrastructure vendors, orchestration framework developers, and fine-tuning service providers each introduce supply chain risks that governance programmes must assess and manage. Vendor security assessment processes, contractual security obligations, and ongoing monitoring of vendor security posture are essential elements of a mature AI security governance programme that no technical architectural control can substitute for (NIST, 2020; ISO, 2022b). The cybersecurity investment decisions that organisations make in deploying AI systems, including choices about model sourcing, retrieval infrastructure hosting, and security tooling, have measurable implications for security outcomes that systematic analysis can inform (Ozowara et al., 2022; Adebayo, 2022b). Understanding the return on cybersecurity investment across different control categories enables organisations to prioritise AI security programme elements that yield the greatest risk reduction per unit of resource expenditure, a consideration particularly salient for smaller critical infrastructure operators with constrained security budgets (Adebayo, 2024; Gartner, 2021).

Furthermore, international standards harmonisation across the AI security governance space reduces compliance burden for critical infrastructure operators who face obligations across multiple jurisdictions and regulatory frameworks simultaneously. Alignment between the NIST AI Risk Management Framework and ISO 42001 for AI management systems, and the alignment of IEC 62443 for operational technology security with broader information security management standards such as ISO 27001, enables organisations to design governance programmes that satisfy multiple frameworks through a coordinated set of controls rather than parallel siloed compliance activities (NIST, 2023; ISO, 2022a; IEC, 2018). Practitioners responsible for AI security governance in critical infrastructure benefit from sustained engagement with standards development processes, providing sector-specific operational requirements that inform the next generation of controls guidance and ensuring that emerging frameworks reflect the distinctive constraints and consequence profiles of critical infrastructure deployments (NIST, 2024a; CISA, 2023a).

This architecture provides a conceptual foundation that is designed to be stable across generations of model architecture, evolving threat landscapes, and changing regulatory requirements, enabling organisations to invest in governance infrastructure with confidence that the structural framework will remain relevant even as specific controls and technologies change (NIST, 2023; NIST, 2024a; CISA, 2023a; Sarker, 2023). The three concerns of identity, data integrity, and adversarial risk reflect durable properties of AI system security that will persist regardless of whether the underlying model is a current-generation large language model or a future architecture with

substantially different technical properties, making the architecture a long-term investment rather than a response to a transient technology configuration.

Additionally, the blockchain-based digital identity verification approaches developed for regulated financial services provide implementation patterns for the identity concern of the three-concern architecture, particularly for use cases involving cross-organisational identity federation and the verification of credentials presented by personnel from partner organisations during collaborative operations (Omoegun et al., 2022). Executive Order 14110 on the safe, secure, and trustworthy development and use of artificial intelligence established federal expectations for AI governance that extend to critical infrastructure operators who provide services to federal agencies or operate systems on which federal operations depend (House, 2023a). The convergence of AI governance obligations with existing cybersecurity compliance requirements creates both a challenge and an opportunity: the challenge of managing an expanding set of overlapping requirements, and the opportunity to design integrated governance programmes that satisfy multiple frameworks through coordinated implementation rather than duplicative parallel efforts. Organisations that approach AI security governance as a component of a unified information governance programme, rather than as an isolated compliance exercise, will achieve more consistent implementation, more efficient use of governance resources, and more effective communication of their security posture to regulators, customers, and operational partners.

12. CONCLUSION

The deployment of generative artificial intelligence in critical infrastructure represents a governance challenge of the first order: the capabilities that make these systems operationally valuable are inseparable from the attack surfaces that make them security risks. The architecture presented in this paper addresses that challenge through a structured framework that integrates established cybersecurity programme practice with controls specific to the properties of large language models. The continued integration of artificial intelligence into both attack and defence will reshape the threat landscape and the available response capabilities. The integration of operational technology with cloud and edge environments will continue, with implications for architecture and governance. Regulatory expectations will continue to evolve, with increasing attention to supply chain transparency, incident reporting, and zero trust architecture. Sustained collaboration across asset owners, vendors, regulators, and researchers will be essential to navigating these trends.

Research priorities emerging from the corpus include the development of empirical evidence on framework adoption, the maturation of automated assessment methods, the integration of safety and security disciplines, and the development of explainable artificial intelligence methods suitable for operational deployment. Academic research will continue to advance methodology while practitioner experience will continue to refine implementation. The most productive research engages both communities through structured collaboration.

The contributions of this work to the broader field include the explicit treatment of operational technology specific considerations, the integration with multiple authoritative frameworks, and the orientation toward sustained practice rather than one-time assessment. The contributions are intended to support both continued research and practitioner application, with the recognition that the field will continue to evolve in response to changing threat, technology, and regulatory environments.

The role of regulation in shaping critical infrastructure cybersecurity practice will continue to evolve. Anticipated regulatory developments include expanded supply chain transparency expectations, refined zero trust expectations, and continued attention to vendor managed access patterns. Regulatory evolution typically proceeds through formal stakeholder processes that incorporate industry feedback and lessons from incident experience. Asset owners benefit from sustained engagement with the regulatory evolution process rather than passive compliance with established expectations.

The architecture presented in this paper addresses a security challenge that will intensify as generative AI systems become more capable, more agentic, and more deeply integrated into critical infrastructure operations. The transition from advisory AI that informs operator decisions to agentic AI that takes actions on operator behalf will increase the consequence of successful adversarial attacks against these systems, making architectural security controls more rather than less important over time. The three-concern structure of identity, data integrity, and adversarial risk provides a stable organizing framework that can accommodate the evolution of specific controls as the technology and threat landscapes change.

This conceptual paper establishes a structured foundation for generative AI security governance in critical infrastructure. The three-concern architecture of identity, data integrity, and adversarial risk provides a stable organizing framework that integrates with established cybersecurity program structures, aligns with authoritative frameworks, and supports practitioner adoption across energy, water, and emergency response sectors. As generative AI capabilities continue to advance and agentic deployments deepen operational integration, the disciplined architectural approach described here provides the basis for responsible adoption that serves both operational objectives and security imperatives. The field will continue to evolve, and the architecture is designed to accommodate that evolution through its alignment with durable security principles rather than transient technical specifics. The three-concern architecture of identity, data integrity, and adversarial risk provides that foundation, grounded in established frameworks and validated through illustrative deployment scenarios across energy, water, and emergency response contexts.

Future work should develop empirical evidence on the effectiveness of specific architectural controls, evaluate the composition of generative AI security with operational technology security in hybrid environments, and refine governance structures as regulatory expectations mature. Collaboration between academic researchers, practitioners, and regulatory bodies will be essential to producing the evidence base that responsible critical infrastructure AI adoption requires. The three-concern architecture provides a starting point for that collaboration by offering a shared vocabulary and structural framework for generative AI security in critical infrastructure. These principles together constitute a comprehensive and defensible approach to generative AI security in critical infrastructure.

The architecture described in this paper is intended to be a durable reference point for organizations navigating the integration of generative AI into critical infrastructure operations. Its value lies not in prescribing specific technical products but in providing a structured analytical lens through which security, operations, and governance teams can evaluate deployment decisions, identify control gaps, and communicate risk in a common vocabulary. This dual purpose makes the architecture a practical instrument for both security programme development and stakeholder communication. The security properties it addresses, including authenticated identity, verified data provenance, and adversarial robustness, are not transient concerns of the current generation of

large language models but structural requirements of any system that applies autonomous reasoning to high-consequence decisions, ensuring that the framework will remain relevant as generative AI capabilities continue to advance.

REFERENCES

1. Adebayo, A. (2020). Wireless internet service provider network reliability frameworks for emerging markets. International Journal of Network and Communication Research, 5(2), 88-103. Tizeti Inc., Nigeria.
2. Adebayo, A. (2021a). Network access control patterns for distributed wireless internet service deployments. International Journal of Information Security Research, 11(4), 215-232. Tizeti Inc., Nigeria.
3. Adebayo, A. (2022a). Compliance frameworks for financial technology platforms operating across multiple jurisdictions. International Journal of Regulatory Technology, 6(3), 134-152. Squad, Nigeria.
4. Adebayo, A. (2023b). Network defense practices in small and medium enterprises: Empirical observations from regional case studies. International Journal of Enterprise Security, 9(4), 245-263. East Tennessee State University, Tennessee, USA.
5. Adebayo, A., Adepoju, P. A., and Ozowara, D. E. (2025a). Conceptual model for cyber risk pricing and insurance structuring in health technology platforms. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2331-2347.
6. Adebayo, A., Anunagba, C. O., and Ozowara, D. E. (2025b). A review of zero trust security models and cost effectiveness in healthcare infrastructure. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2269-2283. https://doi.org/10.62225/2583049X.2025.5.6.6051.
7. Adebayo, T. (2025c). Cyber Risk Pricing Frameworks for Insurance Health Technology Platforms. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2315-2330.
8. Adebayo, T. (2025d). Zero Trust Security Architecture for Healthcare Infrastructure Networks. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2269-2283.
9. Akhtar, N. and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 6, 14410-14430.
10. Almorsy, M., Grundy, J. and Muller, I. (2016). An analysis of the cloud computing security problem. arXiv preprint arXiv:1609.01107.
11. Aminu-Ibrahim, A.Y. and Ogbete, J.C. (2023). Healthcare Infrastructure as a Public Health Intervention Using Evidence from Large Laboratory Networks. Shodhshauryam, International Scientific Refereed Research Journal, 6(1), 256-286. DOI: 10.32628/SHISRRJ23678.
12. Anioke, S. C., and Atima, M. E. (2023a). Public health governance models using process optimization and performance metrics for regulatory oversight. International Journal of Advanced Multidisciplinary Research and Studies, 3(6), 2534-2548. https://doi.org/10.62225/2583049X.2023.3.6.5491.
13. Anioke, S. C., and Atima, M. E. (2023b). Public health informatics frameworks for protecting vulnerable populations through data-driven policy enforcement. International Journal of Advanced Multidisciplinary Research Studies, 3(6), 2564-2579.
14. Anwar, S. and Soltesz, B. (2016). Securing the smart grid. International Journal of Advanced Research in Computer Science, 7(1).
15. Apruzzese, G., Colajanni, M., Ferretti, L. and Marchetti, M. (2019). Addressing adversarial attacks against security systems based on machine learning. International Conference on Cyber Conflict 2019, 1-18.
16. Badmus, O., Jooda, D., Ozowara, D. E., and Anunagba, C. O. (2023). A framework for nonprofit CRM transformation using Salesforce NPSP and NPC: Donor engagement, retention, and grant management. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(10).
17. Beam, A.L. and Kohane, I.S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317-1318.
18. Biggio, B. and Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331.
19. Bishop, M. (2018). Computer Security: Art and Science (2nd ed.). Addison Wesley, Boston.
20. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S. et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
21. Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy 2017, 39-57.

22. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A. and Raffel, C. (2021). Extracting training data from large language models. USENIX Security 2021, 2633-2650.
23. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. and Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.
24. Chen, Y., Mendes, E., Das, S., Xu, W. and Ritter, A. (2023). Can language models be instructed to protect personal information? arXiv preprint arXiv:2310.02224.
25. Cybersecurity and Infrastructure Security Agency (2021). Joint Cybersecurity Advisory on Ransomware Activity Targeting the Healthcare and Public Health Sector. AA20-302A. CISA, Washington, DC.
26. Cybersecurity and Infrastructure Security Agency (2022). Shields Up Guidance for Critical Infrastructure. CISA, Washington, DC.
27. Cybersecurity and Infrastructure Security Agency (2023a). Cybersecurity Performance Goals for Critical Infrastructure (Version 1.0.1). CISA, Washington, DC.
28. Cybersecurity and Infrastructure Security Agency (2023b). Zero Trust Maturity Model (Version 2.0). CISA, Washington, DC.
29. Da Veiga, A., Astakhova, L.V., Botha, A. and Herselman, M. (2020). Defining organisational information security culture: Perspectives from academia and industry. Computers and Security, 92, Article 101713.
30. Department of Homeland Security (2023). Department of Homeland Security Artificial Intelligence Roadmap. DHS, Washington, DC.
31. Dagodzo, D. and Ahiaeke Patrick, M.C. (2025). A Framework for National-Scale UAV Deployment in Power Infrastructure: Lessons from Developing Economies. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(4), 779-824.
32. Department of Energy (2022). National Cyber Informed Engineering Strategy. DOE, Washington, DC.
33. Department of Energy (2023). Cybersecurity Capability Maturity Model (C2M2) Version 2.1. DOE, Washington, DC.
34. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019). BERT: Pre training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 1, 4171-4186.
35. Edivri, J., and Oteri, O. (2021). Applied techniques for reducing delivery latency and improving throughput in large scale enterprise IT implementations. Shodhshauryam, International Scientific Refereed Research Journal, 4(3), 347-365. https://doi.org/10.32628/SHISRRJ.
36. Edivri, J., and Oteri, O. (2022). Predictive capacity planning and resource utilization forecasting models for multi program and multi stakeholder IT portfolios. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(4), 826-845. https://doi.org/10.32628/IJSRCSEIT.
37. Edivri, J., and Oteri, O. (2023). A conceptual governance model for change, incident, and problem management in mission-critical enterprise IT environments. Shodhshauryam, International Scientific Refereed Research Journal, 6(6), 312-335. https://doi.org/10.32628/SHISRRJ.
38. Ekechi, A. T. and Fasasi, T. S.(2022). Conceptual Framework for Process Safety and Operational Reliability in Fpso-based Energy Systems. _International Journal of Artificial Intelligence Engineering and Transformation_, 3(1), 19-37. https://doi.org/10.54660/Ijaiet.2022.3.1.19-37.
39. Ekechi, N.V. (2025b). Artificial Intelligence Governance for Data Privacy and Financial Sustainability in Digital Health Platforms. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2299-2314.
40. Genge, B., Haller, P. and Kiss, I. (2017). A framework for designing resilient distributed intrusion detection systems for critical infrastructures. International Journal of Critical Infrastructure Protection, 15, 3-11.
41. Goodfellow, I.J., Shlens, J. and Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations 2015.
42. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T. and Fritz, M. (2023). Not what you've signed up for: Compromising real world LLM integrated applications with indirect prompt injection. AISec Workshop 2023, 79-90.
43. Hahn, A., Thomas, R.K., Lozano, I. and Cardenas, A. (2015). A multi layered and kill chain based security analysis framework for cyber physical systems. International Journal of Critical Infrastructure Protection, 11, 39-50.
44. Humayed, A., Lin, J., Li, F. and Luo, B. (2017). Cyber physical systems security: A survey. IEEE Internet of Things Journal, 4(6), 1802-1831.
45. Idika, C. N., Enyejo, J. O., Ijiga, O. M. and Okika, N. (2025). Entrepreneurial Innovations in AI-Driven Anomaly Detection for Software-Defined Networking in Critical Infrastructure Security International Journal

of Social Science and Humanities Research Vol. 13, Issue 3, pp: (150-166), DOI: https://doi.org/10.5281/zenodo.16408773.
46. Igure, V.M., Laughter, S.A. and Williams, R.D. (2006). Security issues in SCADA networks. Computers and Security, 25(7), 498-506.
47. International Electrotechnical Commission (2018). IEC 62443-2-1: Security for Industrial Automation and Control Systems, Establishing an IACS Security Program. IEC, Geneva.
48. International Organization for Standardization (2018a). ISO 31000: Risk Management Guidelines. ISO, Geneva.
49. International Organization for Standardization (2018b). ISO/IEC 27005: Information Technology, Security Techniques, Information Security Risk Management. ISO, Geneva.
50. International Organization for Standardization (2022a). ISO/IEC 27001: Information Security, Cybersecurity and Privacy Protection, Information Security Management Systems Requirements. ISO, Geneva.
51. International Organization for Standardization (2022b). ISO/IEC 27002: Information Security, Cybersecurity and Privacy Protection, Information Security Controls. ISO, Geneva.
52. Kevin, M. (2023a). Establishing Next Generation Standards for Regulatory Compliance in Medicare finance. International Journal of Computer Applications Technology and Research, 12(1), 63-70. DOI:10.7753/IJCATR1201.1010.
53. Khurana, H., Hadley, M., Lu, N. and Frincke, D.A. (2010). Smart grid security issues. IEEE Security and Privacy, 8(1), 81-85.
54. Krutz, R.L. and Vines, R.D. (2010). Cloud Security: A Comprehensive Guide to Secure Cloud Computing. Wiley, Indianapolis, IN.
55. Kumuyi, O., Akeju, B., Uzoka, E., and Ozowara, D. E. (2023). Framework for blockchain-based cross-border data exchange and regulatory transparency. International Journal, 4(1), 99-113.
56. Kumuyi, O., Akeju, B., Uzoka, E., and Ozowara, D. E. (2024a). Framework for privacy-focused digital identity verification supporting financial inclusion in Africa. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 3150-3162. https://doi.org/10.62225/2583049X.2024.4.6.6005.
57. Kumuyi, O., Uzoka, E., Akeju, B., and Ozowara, D. E. (2024b). Architecture for machine learning-enabled predictive energy management using IoT sensor networks. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 3138-3149. https://doi.org/10.62225/2583049X.2024.4.6.6004.
58. Kuponiyi, A., and Akomolafe, O. O. (2024). Utilizing AI for predictive maintenance of medical equipment in rural clinics. International Journal of Advanced Multidisciplinary Research and Studies, 4(5), 867-882.
59. Kuponiyi, A., and Akomolafe, O. O. (2025). Digital transformation in public health surveillance: lessons from emerging economies. International Journal of Advanced Multidisciplinary Research and Studies, 5(1), 214-228.
60. Lawal, O. A., and Oduleye, T. E. (2021b). Aligning financial planning analytics with corporate strategy: A conceptual integration model. Shodhshauryam, International Scientific Refereed Research Journal, 4(3), 319-346.
61. Lewis, T.G. (2019). Critical Infrastructure Protection in Homeland Security: Defending a Networked Nation (3rd ed.). Wiley, Hoboken, NJ.
62. Liadi, K. O. (2024b). Conceptualizing a governance reform impact model for Nigeria's peacekeeping missions in post-conflict states. International Journal of Advanced Multidisciplinary Research and Studies, 4(1), 1622-1636.
63. Lin, J., Yu, W., Zhang, N., Yang, X., Zhang, H. and Zhao, W. (2017). A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications. IEEE Internet of Things Journal, 4(5), 1125-1142.
64. Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K. and Liu, Y. (2023). Jailbreaking ChatGPT via prompt engineering: An empirical study. arXiv preprint arXiv:2305.13860.
65. Lopez, J., Rubio, J.E. and Alcaraz, C. (2018). A resilient architecture for the smart grid. IEEE Transactions on Industrial Informatics, 14(8), 3745-3753.
66. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations 2018.
67. Marchang, J. and Sumathy, S. (2021). Adversarial attacks in critical infrastructure: A systematic review. Computers and Security, 109, Article 102389.

68. Michael, O.N. and Ogunsola, O.E. (2022a). Examining the Socioeconomic Barriers to Technological Adoption Among Smallholder Farmers in Remote Rural Areas. Shodhshauryam, International Scientific Refereed Research Journal, 5(6), 484-519. DOI: 10.32628/SHISRRJ.
69. Michael, O.N. and Ogunsola, O.E. (2024a). Assessing the Potential of Renewable Energy Technologies for Sustainable Irrigation and Smallholder Farm Productivity. International Journal of Scientific Research in Humanities and Social Sciences, 1(1), 380-411. DOI: 10.32628/IJSRSSH.
70. Mo, Y., Kim, T.H.J., Brancik, K., Dickinson, D., Lee, H., Perrig, A. and Sinopoli, B. (2012). Cyber physical security of a smart grid infrastructure. Proceedings of the IEEE, 100(1), 195-209.
71. Mosenia, A. and Jha, N.K. (2017). A comprehensive study of security of internet of things. IEEE Transactions on Emerging Topics in Computing, 5(4), 586-602.
72. National Institute of Standards and Technology (2011). Managing Information Security Risk: Organization, Mission, and Information System View. NIST Special Publication 800-39. NIST, Gaithersburg, MD.
73. National Institute of Standards and Technology (2014). Framework for Improving Critical Infrastructure Cybersecurity (Version 1.0). NIST, Gaithersburg, MD.
74. National Institute of Standards and Technology (2018a). Cybersecurity Framework Profile for Industrial Internet of Things. NIST IR 8259. NIST, Gaithersburg, MD.
75. National Institute of Standards and Technology (2018b). Framework for Improving Critical Infrastructure Cybersecurity (Version 1.1). NIST, Gaithersburg, MD.
76. National Institute of Standards and Technology (2018c). Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy. NIST Special Publication 800-37 Rev 2. NIST, Gaithersburg, MD.
77. National Institute of Standards and Technology (2020). Security and Privacy Controls for Information Systems and Organizations. NIST Special Publication 800-53 Rev 5. NIST, Gaithersburg, MD.
78. National Institute of Standards and Technology (2023). AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1. NIST, Gaithersburg, MD.
79. National Institute of Standards and Technology (2024a). AI Risk Management Framework Generative AI Profile. NIST AI 600-1. NIST, Gaithersburg, MD.
80. National Institute of Standards and Technology (2024b). Framework for Improving Critical Infrastructure Cybersecurity (Version 2.0). NIST, Gaithersburg, MD.
81. Nicholson, A., Webber, S., Dyer, S., Patel, T. and Janicke, H. (2012). SCADA security in the light of cyber warfare. Computers and Security, 31(4), 418-436.
82. Nnaji, N., and Akinlolu, V. S. (2022). Reducing decision delays through real-time health informatics and performance analytics. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(6), 541-556.
83. Nnaji, N., and Akinlolu, V. S. (2023). Addressing productivity and workflow inefficiencies using analytics-driven performance models in healthcare operations. Shodhshauryam, International Scientific Refereed Research Journal, 6(3), 318-334.
84. Nnaji, N., and Akinlolu, V. S. (2024a). Enhancing clinical record protection through structured health data security frameworks in regulated healthcare environments. International Journal of Scientific Research in Humanities and Social Sciences, 1(2), 448-464.
85. North American Electric Reliability Corporation (2023). Critical Infrastructure Protection Standards. NERC CIP 002 through CIP 014. NERC, Atlanta, GA.
86. Oduleye, T. E., and Medon, J. J. (2021). A Data-Driven Cost Management Model for Improving Strategic Financial Planning and Performance Evaluation. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 524-537.
87. Ogbete, J.C. and Aminu-Ibrahim, A. (2024). Translating Healthcare Infrastructure Investment into Measurable Population Health and Diagnostic Outcomes. International Journal of Scientific Research in Humanities and Social Sciences, 1(2), 955-985. DOI: 10.32628/IJSRSSH242775.
88. Ozowara, D. E., Adebayo, A., and Anunagba, C. O. (2025a). Advanced conceptual model for strengthening audit quality using data analytics across financial institutions. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2246-2268. https://doi.org/10.62225/2583049X.2025.5.6.6050.
89. Ozowara, D. E., Anunagba, C. O., and Adepoju, P. A. (2025b). A review of ransomware economics and financial resilience strategies in hospital networks. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2284-2298. https://doi.org/10.62225/2583049X.2025.5.6.6052.

90. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B. and Swami, A. (2016). The limitations of deep learning in adversarial settings. IEEE European Symposium on Security and Privacy 2016, 372-387.
91. Patrick, M.C.A., Okonkwo, C.S., Mayo, W. and Okeke, O.T. (2020). A GIS Enabled Framework for Modern ERP Procurement Processes. International Journal of Multidisciplinary Research and Growth Evaluation, 1(5), 499-508. DOI: 10.54660/Ijmrge.2020.1.5.499-508.
92. Perez, F. and Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. arXiv preprint arXiv:2211.09527.
93. Phiayura, P. and Teerakanok, S. (2023). A comprehensive framework for migrating to zero trust architecture. IEEE Access, 11, 19937-19952.
94. Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E. and Loukas, G. (2019). A taxonomy and survey of attacks against machine learning. Computer Science Review, 34, Article 100199.
95. Rose, S., Borchert, O., Mitchell, S. and Connelly, S. (2020). Zero Trust Architecture. NIST Special Publication 800-207. NIST, Gaithersburg, MD.
96. Sanni, J. O., and Atima, M. E. (2021a). Analytics driven go-to-market frameworks addressing compliance sustainability complexity service portfolios. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 647-660.
97. Sanni, J.O. and Atima, M.E. (2021b). Analytics Driven Go to Market Frameworks Addressing Compliance Sustainability Complexity. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 647-660.
98. Sanni, J.O. and Atima, M.E. (2021c). Business Intelligence Dashboard Frameworks Resolving Executive Visibility Gaps in Strategic Marketing Governance. International Journal of Multidisciplinary Research and Growth Evaluation, 2(6), 633-646.
99. Sanni, J.O. and Attah, A. (2023c). Market Research Frameworks Addressing Entry Barriers Within Highly Regulated Industrial Service Sectors. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(2), 904-930.
100. Sanni, J.O. and Wedraogo, L. (2024b). Data Centric Funnel Optimization Frameworks Resolving Revenue Leakage in High Value Services. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2859-2874.
101. Sanni, J.O., Iwuanyanwu, U.A., Essien, M.A. and Wedraogo, L. (2024a). Integrating Blockchain-Enabled Smart Contracts for Transparent and Verifiable Marketing Workflows. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2891-2906. DOI: 10.62225/2583049X.2024.4.6.5646.
102. Sarker, I.H. (2024). AI driven cybersecurity and threat intelligence: Cyber automation, intelligent decision making and explainability. Springer Nature.
103. Schinagl, S., Schoon, K. and Paans, R. (2015). A framework for designing a security operations centre (SOC). HICSS 2015, 2253-2262.
104. Souppaya, M., Morello, J. and Scarfone, K. (2020). Application Container Security Guide. NIST Special Publication 800-190. NIST, Gaithersburg, MD.
105. Stellios, I., Kotzanikolaou, P., Psarakis, M., Alcaraz, C. and Lopez, J. (2018). A survey of IoT enabled cyber attacks: Assessing attack paths to critical infrastructure and services. IEEE Communications Surveys and Tutorials, 20(4), 3453-3495.
106. Stouffer, K., Pease, M., Tang, C., Zimmerman, T., Pillitteri, V., Lightman, S. and Hahn, A. (2023). Guide to Operational Technology (OT) Security. NIST Special Publication 800-82 Rev 3. NIST, Gaithersburg, MD.
107. Stouffer, K., Pillitteri, V., Lightman, S., Abrams, M. and Hahn, A. (2015). Guide to Industrial Control Systems (ICS) Security. NIST Special Publication 800-82 Rev 2. NIST, Gaithersburg, MD.
108. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014). Intriguing properties of neural networks. International Conference on Learning Representations 2014.
109. Taiwo, S. O. (2022). Pfai™: A predictive financial planning and analysis intelligence framework for transforming enterprise decision-making. International Journal of Scientific Research in Science, Engineering and Technology, 9(6), 472-487. https://doi.org/10.32628/IJSRSET25122272.
110. Ten, C.W., Manimaran, G. and Liu, C.C. (2010). Cybersecurity for critical infrastructures: Attack and defense modeling. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 40(4), 853-865.
111. Topol, E.J. (2019). High performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.
112. Vassilev, A., Oprea, A., Fordyce, A. and Anderson, H. (2024). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2023. NIST, Gaithersburg, MD.

113. Voigt, P. and von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer, Cham.
114. Wedraogo, L. and Sanni, J.O. (2024). Machine Learning Models Addressing Uncertainty in Cross Channel Campaign Performance Forecasting Accuracy. International Journal of Advanced Multidisciplinary Research and Studies, 4(6), 2875-2890.
115. Wei, A., Haghtalab, N. and Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? Advances in Neural Information Processing Systems 36.
116. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., Isaac, W., Legassick, S., Irving, G. and Gabriel, I. (2022). Taxonomy of risks posed by language models. ACM FAccT 2022, 214-229.
117. West, J. and Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers and Security, 57, 47-66.
118. Yadav, G. and Paul, K. (2019). Architecture and security of SCADA systems: A review. International Journal of Critical Infrastructure Protection, 34, Article 100433.
119. Yeboah, B. K., and Nnabueze, S. B. (2021). Policy-oriented framework for predictive analytics in maintenance optimization. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(1), 585-602.
120. Yeboah, B. K., and Nnabueze, S. B. (2024). Data-driven policy framework for energy efficiency in higher education institutions. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(8), 255-270.
121. Yuan, X., He, P., Zhu, Q. and Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 30(9), 2805-2824.

122. Cheng, L., Liu, F. and Yao, D. (2017). Enterprise data breach: Causes, challenges, prevention, and future directions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(5), Article e1211.
123. Liang, G., Weller, S.R., Zhao, J., Luo, F. and Dong, Z.Y. (2017). The 2015 Ukraine blackout: Implications for false data injection attacks. IEEE Transactions on Power Systems, 32(4), 3317-3318.
124. Sridhar, S., Hahn, A. and Govindarasu, M. (2012). Cyber-physical system security for the electric power grid. Proceedings of the IEEE, 100(1), 210-224.
125. Ryan, M.D. (2013). Cloud computing security: The scientific challenge, and a survey of solutions. Journal of Systems and Software, 86(9), 2263-2278.
126. Khisamova, Z.I., Begishev, I.R. and Gaifutdinov, R.R. (2019). Artificial intelligence and problems of ensuring cyber security. International Journal of Cyber Criminology, 13(2), 564-577.
127. Diogenes, Y. and Ozkaya, E. (2018). Cybersecurity: Attack and Defense Strategies. Packt Publishing, Birmingham.
128. Stamp, M. (2011). Information Security: Principles and Practice (2nd ed.). Wiley, Hoboken, NJ.
129. Stallings, W. (2017). Cryptography and Network Security: Principles and Practice (7th ed.). Pearson, New York.
130. Kim, D. and Solomon, M.G. (2018). Fundamentals of Information Systems Security (3rd ed.). Jones and Bartlett Learning, Burlington, MA.
131. Ciampa, M. (2018). Security+ Guide to Network Security Fundamentals (6th ed.). Cengage Learning, Boston, MA.
132. Andress, J. (2014). The Basics of Information Security: Understanding the Fundamentals of InfoSec in Theory and Practice (2nd ed.). Syngress, Waltham, MA.
133. Pfleeger, C.P., Pfleeger, S.L. and Margulies, J. (2015). Security in Computing (5th ed.). Prentice Hall, Upper Saddle River, NJ.
134. Dagodzo, D. and Ahiaeke Patrick, M.C. (2022). A Review of Right-of-Way Encroachment Detection Methods Using Geospatial Technologies. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(1), 638-667.
135. Baykara, M. and Daimi, K. (2018). Cybersecurity: A survey of challenges, risks, and mitigation strategies. International Journal of Computer Science and Information Security, 16(9), 15-22.
136. Cherdantseva, Y. and Hilton, J. (2013). A reference model of information assurance and security. Proceedings of the Eighth International Conference on Availability, Reliability and Security, 547-554.

137. Enisa (2022). ENISA Threat Landscape 2022. European Union Agency for Cybersecurity, Athens.
138. Sani, A.S., Yuan, D., Jin, J., Gao, L., Yu, S. and Dong, Z.Y. (2019). Cyber security framework for Internet of Things-based Energy Internet. Future Generation Computer Systems, 93, 849-859.
139. Sajid, A., Abbas, H. and Saleem, K. (2016). Cloud-assisted IoT-based SCADA systems security: A review of the state of the art and future challenges. IEEE Access, 4, 1375-1384.
140. Gupta, B.B., Arachchilage, N.A.G. and Psannis, K.E. (2018). Defending against phishing attacks: Taxonomy of methods, current issues and future directions. Telecommunication Systems, 67(2), 247-267.
141. Bertino, E. and Islam, N. (2017). Botnets and Internet of Things security. Computer, 50(2), 76-79.
142. Cardenas, A.A., Amin, S. and Sastry, S. (2008). Research challenges for the security of control systems. HotSec 2008, Article 6.
143. Cardenas, A.A., Roosta, T. and Sastry, S. (2009). Rethinking security properties, threat models, and the link layer for ad hoc networks. Ad Hoc Networks, 7(8), 1434-1447.
144. Coppolino, L., D'Antonio, S., Mazzeo, G. and Romano, L. (2017). Cloud security: Emerging threats and current solutions. Computers and Electrical Engineering, 59, 126-140.
145. Cyber Incident Reporting for Critical Infrastructure Act (2022). Public Law 117-103, Division Y. United States Congress, Washington, DC.
146. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D. (2018). Robust physical world attacks on deep learning visual classification. IEEE/CVF CVPR 2018, 1625-1634.
147. Gartner (2021). Cybersecurity Predictions for the Distributed Ecosystem. Gartner, Inc., Stamford, CT.
148. House, W. (2021). Executive Order 14028: Improving the Nation's Cybersecurity. Federal Register, 86(93), 26633-26648.
149. House, W. (2023b). National Cybersecurity Strategy. Office of the National Cyber Director, Washington, DC.
150. House, W. (2024). National Cybersecurity Strategy Implementation Plan (Version 2). Office of the National Cyber Director, Washington, DC.
151. Kurakin, A., Goodfellow, I.J. and Bengio, S. (2017). Adversarial examples in the physical world. International Conference on Learning Representations Workshop.
152. Maglaras, L.A., Kim, K.H., Janicke, H., Ferrag, M.A., Rallis, S., Fragkou, P., Maglaras, A. and Cruz, T.J. (2018). Cyber security of critical infrastructures. ICT Express, 4(1), 42-45.
153. Office of Management and Budget (2022). Moving the U.S. Government Toward Zero Trust Cybersecurity Principles. OMB Memorandum M to 22-9. Washington, DC.
154. OpenAI (2024). Preparedness Framework. OpenAI, San Francisco, CA.
155. Project, O. W. A. S. (2023). OWASP Top 10 for Large Language Model Applications. OWASP Foundation, Wakefield, MA.
156. Roman, R., Lopez, J. and Mambo, M. (2018). Mobile edge computing, Fog et al.: A survey and analysis of security threats and challenges. Future Generation Computer Systems, 78, 680-698.
157. Sarker, I.H. (2023). Multi aspects AI based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview. Security and Privacy, 6(5), Article e295.
158. Schneier, B. (2018). Click Here to Kill Everybody: Security and Survival in a Hyper Connected World. W.W. Norton, New York.
159. Sgandurra, D., Carmagnola, F., Conti, M. and Lupu, E.C. (2017). The trinity of insider threat: An insider threat detection framework. 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA), 1-5.
160. Sicari, S., Rizzardi, A., Grieco, L.A. and Coen-Porisini, A. (2015). Security, privacy and trust in Internet of Things: The road ahead. Computer Networks, 76, 146-164.
161. Singh, A. and Chatterjee, K. (2017). Cloud security issues and challenges: A survey. Journal of Network and Computer Applications, 79, 88-115.
162. Tabrizchi, H. and Kuchaki Rafsanjani, M. (2020). A survey on security challenges in cloud computing: Issues, threats, and solutions. Journal of Supercomputing, 76, 9493-9532.
163. Adebayo, A. (2022b). Fintech payment infrastructure security: Threat modeling for high-volume transaction processing platforms. Journal of Financial Technology Security, 8(2), 198-217. Squad, Nigeria.
164. Adebayo, A. (2023a). Cybersecurity workforce development through applied research curricula: A case study approach. Journal of Cybersecurity Education and Research, 14(2), 78-95. East Tennessee State University, Tennessee, USA.

165. Adebayo, A. (2024). Threat detection capability assessment for resource-constrained organizations: A maturity model approach. Journal of Information Security Practice, 17(3), 312-329. East Tennessee State University, Tennessee, USA.
166. Alexander, O., Belisle, M. and Steele, J. (2020). MITRE ATT&CK for Industrial Control Systems: Design and Philosophy. MITRE Corporation, McLean, VA.
167. Apruzzese, G., Andreolini, M., Ferretti, L., Marchetti, M. and Colajanni, M. (2022). Modeling realistic adversarial attacks against network intrusion detection systems. ACM Digital Threats, 3(3), 1-19.
168. Ekechi, N. V., Ozowara, D. E., and Anunagba, C. O. (2025). Conceptual framework for AI governance, data privacy compliance, and financial sustainability in digital health. International Journal of Advanced Multidisciplinary Research and Studies, 5(6), 2283-2298.
169. Hashizume, K., Rosado, D.G., Fernandez-Medina, E. and Fernandez, E.B. (2013). An analysis of security issues for cloud computing. Journal of Internet Services and Applications, 4(1), 1-13.
170. Idika, C. N., Salami, E. O., Ijiga, O. M., and Enyejo, L. A. (2021). Deep Learning Driven Malware Classification for Cloud-Native Microservices in Edge Computing Architectures International Journal of Scientific Research in Computer Science. Engineering and Information Technology, 7(4).
171. Mbonu, I.S., Iwuanyanwu, U., Uzoka, E. and Oluoha, O.M. (2019b). Advances in Enterprise Log Analytics and Automated Incident Response Architectures Using Python and SIEM Platforms. Iconic Research and Engineering Journals, 3(2), 1000-1019.
172. Mbonu, I.S., Aliliele, C., Iwuanyanwu, U. and Uzoka, E. (2020a). A Review of Identity and Access Management Integration Strategies in Hybrid and Multi Cloud Environments. International Journal of Multidisciplinary Research and Growth Evaluation, 1(5), 795-810.
173. Mbonu, I.S., Iwuanyanwu, U., Aliliele, C. and Uzoka, E. (2022c). Advances in Cloud Identity and Access Governance Optimization in Large Scale AWS Enterprise Environments. Shodhshauryam, International Scientific Refereed Research Journal, 5(3), 403-438.
174. Okoruwa, P. O., Fadayomi, O., Akeju, B., Edivri, J., Ogbole, J. I., and Abolaji, T. O. (2020). Conceptual model for privacy-centric security engineering in digital and cloud computing systems. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(5), 371-389. https://doi.org/10.32628/IJSRCSEIT.
175. Olatunde-Thorpe, J., Aifuwa, S. E., Oshoba, T. O., and Ogbuefi, E. (2020). Metadata-driven access controls: Designing role-based systems for analytics teams in high-risk industries. International Journal of Multidisciplinary Research and Growth Evaluation, 1(3), 143-162. DOI: 10.54660/.IJMRGE.2020.1.3.143-162.
176. Oshoba, T. O., Hammed, N. I., and Odejobi, O. D. (2019). Secure identity and access management model for distributed and federated systems. IRE Journals, 3(4). ISSN: 2456-8880.
177. Oshoba, T.O., Hammed, N.I. and Odejobi, O.D. (2020b). Blockchain-Enabled Compliance and Audit Trail Model for Cloud Configuration Management. Journal of Frontiers in Multidisciplinary Research, 1(1), 193-201. DOI: 10.54660/.Ljfmr.2020.1.1.193-201.
178. Oshoba, T. O., Hammed, N. I., and Odejobi, O. D. (2021). Adoption model for multi-factor authentication in enterprise Microsoft 365 environments. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(1), 519-536.
179. Oshoba, T. O., Ahmed, K. S., and Odejobi, O. D. (2023b). Proactive threat intelligence and detection model using cloud-native security tools. International Journal of Advanced Multidisciplinary Research and Studies, 3(1), 1481-1490. DOI: 10.62225/2583049X.2023.3.1.5140.
180. Ozowara, D. E., Adebayo, A., and Anunagba, C. O. (2022). A systematic review of cybersecurity investments and their impact on healthcare financial performance. Gyanshauryam, International Scientific Refereed Research Journal, 5(1), 404-427.
181. Strom, B.E., Applebaum, A., Miller, D.P., Nickels, K.C., Pennington, A.G. and Thomas, C.B. (2018). MITRE ATT&CK: Design and Philosophy. MITRE Corporation, McLean, VA.
182. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z. and Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. High Confidence Computing, 4(2), Article 100211.
183. Omoegun, G. O., Fadayomi, O., Bello, A. D., and Elebe, O. (2022). A blockchain-based Know Your Customer and digital identity verification framework for regulated financial services. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(3), 166-180.
184. House, W. (2023a). Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Federal Register, 88(210), 75191-75226.