OPEN ACCESS

World Journal of Multidisciplinary Studies
ISSN (Online): 3049-0855 | Vol. 3, No. 2, Feb. 2026
Website: https://wasrpublication.com/index.php/wjms

Artificial Intelligence and Machine Learning for Critical Infrastructure Risk Mitigation: Applications, Trade-offs, and Governance Challenges

Anthony Mazza[1*], Lawrence Barnett[2], Angelaquie Cole[2] & Nikki Edwards[2]

[1]Professor, University of the District of Columbia, Masters in Homeland Security Program

[2]Master of Science candidate, University of the District of Columbia, Masters in Homeland Security Program.

Received: 24/11/2025 | Accepted: 02/01/2026 | Published: 05/02/2026

Abstract: Critical infrastructure systems face unprecedented cybersecurity threats that exceed the capacity of traditional risk mitigation approaches. Artificial intelligence (AI) and machine learning (ML) technologies offer transformative capabilities for real-time threat detection, predictive analysis, and automated risk assessment in operational technology (OT) and industrial control system (ICS) environments. However, the deployment of AI/ML in safety-critical infrastructure introduces complex trade-offs between speed and accuracy, availability and security, explainability and performance, and automation and human control. This article provides a comprehensive analysis of AI/ML applications in critical infrastructure protection, evaluates key trade-offs in system design, and addresses limitations, ethical considerations, and governance frameworks necessary for responsible AI deployment. Drawing on recent literature and empirical evidence, this article demonstrates that while AI/ML significantly enhances threat detection and predictive capabilities, successful implementation requires careful attention to organizational readiness, domain expertise integration, explainable AI techniques, and robust governance structures that balance innovation with safety and accountability.

Keywords: Anomaly detection, artificial intelligence, critical infrastructure, cybersecurity, governance, machine learning, operational technology, SCADA systems.

Introduction

Critical infrastructure (CI) encompasses the systems and assets essential to national security, economic prosperity, public health, and safety. The Cybersecurity and Infrastructure Security Agency (CISA) identify sixteen critical infrastructure sectors, including energy, water, transportation, communications, financial services, and healthcare, whose incapacitation would have debilitating societal impacts (Xiang et al, 2025). These infrastructures increasingly face sophisticated cyber threats from nation-state actors, criminal organizations, and insider threats, with high-impact attacks on critical infrastructure increasing by 140% in recent years (Ferrag et al, 2024).

Traditional risk mitigation approaches—relying on periodic audits, manual log reviews, and signature-based intrusion detection—prove inadequate for the scale, speed, and complexity of modern cyber threats (Raman et al, 2024). The convergence of information technology (IT) and operational technology (OT) networks, combined with the proliferation of Internet of Things (IoT) devices in industrial environments, creates expanded attack surfaces that exceed human analytical capacity (Ahmad et al, 2022). Moreover, the sophistication of attacks such as Stuxnet, BlackEnergy, and the Colonial Pipeline ransomware demonstrate that adversaries possess the capabilities and motivation to target industrial control systems with potentially catastrophic consequences (Sarker et al, 2024).

Artificial intelligence and machine learning offer promising solutions to these challenges. AI/ML technologies can process millions of events per second, identify complex patterns invisible to human analysts, predict failures before they occur, and adapt to evolving threat landscapes (Kaur et al, 2023; Zhang et al, 2022). In cybersecurity applications, AI's capabilities, including predictive analytics, machine learning, and autonomous decision-making, allow it to go beyond merely detecting threats but also have a proactive approach against threat actors in cyberspace. Machine learning algorithms enable systems to learn from experiences without explicit programming, making them more effective than static rule-based systems when confronted with novel attack patterns (Macas et al, 2022).

However, the integration of AI/ML into critical infrastructure protection introduces significant challenges. Concerns include lack of transparency in AI decision-making processes, vulnerabilities to adversarial manipulation, ethical considerations regarding autonomous cyber defense, and the potential for AI systems themselves to introduce new failure modes (Stevens, 2020; Guembe et al, 2022). Furthermore, in 2024, organizations experienced an average of 1,308 attacks per week in the first quarter, marking a 28% increase from the last quarter of 2023, with cybercrime losses projected to reach $13.82 trillion by 2028 (Ferrag et al, 2024). This escalating threat environment demands careful consideration of how AI/ML can be responsibly deployed to enhance rather than undermine infrastructure security and resilience.

*__Corresponding Author__

Anthony Mazza*

Professor, University of the District of Columbia, Masters in Homeland Security Program.

This article provides a comprehensive examination of AI/ML applications in critical infrastructure risk mitigation, by establishing the foundational concepts of critical infrastructure risk mitigation and explaining why AI/ML represents a paradigm shift from reactive to proactive security. It then presents detailed analysis of five key AI/ML applications: real-time threat detection, predictive analysis, automated risk assessment, disaster-resilient infrastructure design, and common implementation challenges. The paper analyzes six critical trade-offs that designers must navigate when deploying AI/ML systems in critical infrastructure contexts and addresses fundamental limitations of AI/ML technologies, ethical considerations, and governance frameworks necessary for responsible deployment. The article concludes with recommendations for balancing innovation with safety and accountability in AI-enabled critical infrastructure protection.

CI Risk Mitigation: From Traditional Approaches to the AI-Enabled Paradigm Shift

Evolution of Critical Infrastructure Systems

The development of modern critical infrastructure began in earnest during the Industrial Revolution of the 18th and 19th centuries, when large-scale systems for power generation, water distribution, and transportation emerged to support growing urban populations and industrial economies. These early infrastructures operated as isolated, physically secured systems with manual oversight and control. Security concerns focused almost exclusively on physical protection—locked facilities, fences, guards, and geographic isolation from population centers (Presidential Decision Directive 63, 1998).

Throughout most of the 20th century, critical infrastructure systems remained physically and logically independent and separate, with little interaction or connection with each other or other sectors of the infrastructure (Burgess, 2024). Water treatment plants operated independently from power grids, which functioned separately from transportation networks. Each sector developed its own operational protocols, safety standards, and security measures tailored to physical threats such as sabotage, natural disasters, and equipment failure. This isolation provided inherent security benefits: an attacker would need physical access to a specific facility to cause disruption, and damage would typically remain contained within a single system or geographic area.

The Advent of SCADA and Automation

The landscape began to transform dramatically with the introduction of Supervisory Control and Data Acquisition (SCADA) systems in the 1960s and 1970s. SCADA technology emerged from the need to monitor and control geographically distributed assets more efficiently than manual oversight allowed (History of SCADA, 2023). Early SCADA systems were developed using telephone lines and telemetry to enable remote monitoring and control of industrial processes, particularly in the energy and utilities sectors (Mishra, 2025). These first-generation systems were monolithic in design, centralized on mainframe computers, and operated within single facilities with limited functionality (Unmudl, 2025).

The development of microprocessors and digital communications technology in the 1970s revolutionized SCADA capabilities, making it possible to create more powerful and flexible systems that could monitor and control industrial processes more effectively across greater distances (SCADA Info, 2023). The term "SCADA" was formally coined in the early 1970s, and the rise of programmable logic controllers (PLCs) and remote terminal units (RTUs) during that decade dramatically increased enterprises' ability to automate processes (Inductive Automation, 2018). It is estimated that approximately 3 million SCADA systems are currently in use worldwide (Hildick-Smith, 2022).

SCADA systems became increasingly sophisticated during the 1980s and 1990s, evolving from isolated control systems to distributed architectures that incorporated advanced data logging, alarm management, and remote access capabilities (SCADA Info, 2023). The development of the Internet in the 1990s paved the way for web-based SCADA systems, allowing operators to access and control systems from anywhere in the world (History of SCADA, 2023). This connectivity provided unprecedented operational efficiency but also introduced new vulnerabilities that were not present in physically isolated systems.

The Efficiency-Versus-Security Trade-off in Infrastructure Automation

As infrastructure systems evolved throughout the late 20th and early 21st centuries, a critical design philosophy emerged that would profoundly shape cybersecurity challenges: critical infrastructure automation was developed and deployed with an overwhelming emphasis on operational efficiency and reliability rather than security (Presidential Decision Directive 63, 1998). This efficiency-first approach was driven by several factors. Foremost, were economic imperatives. Infrastructure operators faced intense pressure to reduce operational costs, minimize downtime, and maximize return on investment. Automation enabled remote monitoring and control, reducing the need for on-site personnel and enabling faster response to operational issues (Rockwell Automation, 2022).

The evolution of technology also shaped the operational efficiency emphasis. When many SCADA and ICS systems were designed, they operated on proprietary protocols within isolated networks, creating an assumption that physical isolation provided adequate security. The concept of cyber threats to industrial systems was not well understood during the initial design phases (Hildick-Smith, 2022). Regulatory frameworks historically focused on operational safety and service reliability rather than cybersecurity. Standards such as NERC CIP (North American Electric Reliability Corporation Critical Infrastructure Protection) emerged only after significant cyber incidents demonstrated vulnerabilities (Gordon, 2020).

As advances in technology enabled systems within each sector to become automated and interlinked through computers and communications facilities, the flow of electricity, oil, gas, and telecommunications throughout the country became linked—albeit sometimes indirectly—but the resulting linkages blurred traditional security borders. While this increased reliance on interlinked capabilities and Internet of Things (IoT) technology convergence helped make the economy and nation more efficient and perhaps stronger, it also made the country more vulnerable to disruption and attack (Burgess, 2024). Critical Infrastructure organizations underwent digital transformation, digitizing processes, and adopting IoT to improve efficiency and reliability. The resulting connectivity of operational technology (OT)- the hardware and software used to monitor and control physicalprocesses- to the internet and the convergence between OT and IT created extreme efficiencies, but also new vulnerabilities and exposure to cybersecurity threats (Rockwell Automation, 2022).

The prioritization of efficiency over security has resulted in several persistent consequences and vulnerabilities in critical infrastructure. Many legacy devices were not originally designed with today's cybersecurity standards and remain vital to operations. Integrating them into modern security frameworks is essential to ensure they do not become weak links in the broader system (Device Authority, 2024). Remote Terminal Units (RTUs) and automatic controller devices were developed before industry-wide standards for interoperability existed, resulting in a multitude of control protocols. Larger vendors had incentives to create proprietary protocols to "lock in" their customer base, which meant security was often overlooked and few people beyond developers understood how secure installations were (Gordon, 2020). OT components often lack built-in protections; for example, industrial control systems may not support encryption or modern authentication protocols (Forvis Mazars, 2025). OT equipment operates for 20-30 years, far longer than typical IT systems, meaning infrastructure designed decades ago when cyber threats were minimal continues to operate in today's threat environment (Xiang et al, 2025).

Financial Barriers to Modernization

The deployment of AI/ML threat detection capabilities in existing critical infrastructure faces significant financial constraints that often make comprehensive security retrofitting impractical or infeasible. The age of energy grids infrastructure coupled with outdated technology poses huge obstacles that significantly increase the chance of attacks. To create resilient critical infrastructure, the technology must be upgraded, and equipment should be modernized (American Military University, 2024). However, several budgetary realities complicate these necessary upgrades.

Utilities operate under regulated rate-of-return models that create capital constraints, as they can only recoup approved costs from ratepayers. Security investments often compete with operational improvements that have more direct customer benefits (Maglaras et al, 2023). Public infrastructure sectors including water, wastewater, and transportation systems operate with limited budgets and face political pressures to minimize costs. Cybersecurity investments, while critical, may be perceived as providing less tangible value than visible infrastructure improvements (Maglaras et al, 2023).

Competing investment priorities continue to interfere with security investment. There exists a fundamental trade-off between operational efficiency upgrades and security capabilities. In a country with aging critical infrastructure in desperate need of both operational improvements and hardening capabilities, patchwork modifications become rampant due to lack of political will to prioritize the huge investments necessary for comprehensive modernization.

Technical Challenges of Retrofitting AI Systems

Securing legacy systems within critical infrastructure presents challenges, as many were not designed with modern cybersecurity requirements in mind. Automating lifecycle management ensures that legacy devices receive necessary updates, patches, and monitoring throughout their operational life, preventing them from becoming entry points for attackers while avoiding disruptions to critical operations (Device Authority, 2024). Critical infrastructure cannot simply be taken offline for upgrades. Any security enhancement must be implemented without disrupting essential services, requiring phased deployment strategies and extensive testing that increase costs and timeline (Rockwell Automation, 2025).

Beyond financial constraints, significant technical barriers impede the integration of AI/ML capabilities into legacy infrastructure. Unpatched legacy infrastructure and other common exposures make industrial organizations attractive targets (Rockwell Automation, 2025). OT devices including PLCs and RTUs have minimal processing power insufficient for sophisticated ML models locally, and network bandwidth constraints limit cloud-based inference capabilities. Legacy SCADA systems may operate on 56K modems, making real-time AI processing impractical (Sarker et al, 2024).

Another challenging reality is the Operating Environments in which equipment is deployed. Critical infrastructure equipment must operate in extreme temperatures (-40°C to +85°C), high vibration, and electromagnetic interference, which limit hardware options for edge ML deployment and require reliability standards that exceed typical IT equipment (Maglaras et al, 2023).

Legacy systems also lack standardized logging formats and data collection protocols, making it extremely difficult to aggregate data necessary for training ML models. Proprietary protocols and incompatible data formats create significant preprocessing burdens (Xiang et al, 2025). Given these constraints, organizations must carefully assess whether AI/ML integration is feasible for their specific context.

Newer facilities with modern SCADA systems, adequate computing resources, standardized protocols, and available capital for investment present good candidates for AI/ML deployment. Systems with partial modernization may support hybrid approaches, such as deploying AI at network gateways while maintaining legacy controls or using centralized cloud-based analytics for non-real-time threat detection. Severely resource-constrained systems with decades-old equipment, proprietary protocols, and minimal computing capacity may find AI/ML retrofitting cost-prohibitive. These systems may require alternative security strategies such as network segmentation, increased physical security, and prioritized replacement schedules rather than AI enhancement. Professional risk and vulnerability assessments by experienced industrial and cybersecurity professionals can find both common and hidden gaps and help organizations determine whether their current defenses are sufficient or whether AI investments represent the most effective use of limited security budgets (Rockwell Automation, 2025).

The Critical Infrastructure Threat Landscape

Critical infrastructure faces a complex threat environment encompassing both traditional and emerging risks. Traditional threats include natural disasters (hurricanes, earthquakes, floods), equipment failures resulting from aging infrastructure, human error, and physical attacks (Maglaras et al, 2023). Emerging cyber threats pose equally significant risks: nation-state actors targeting supervisory control and data acquisition (SCADA) and industrial control systems (ICS), ransomware attacks paralyzing utilities and healthcare facilities, supply chain compromises, insider threats, and vulnerabilities in IoT/OT devices (Ahmed et al, 2022; Whatney, 2022).

Understanding Network Traffic Patterns: Lateral vs. Perimeter

To comprehend modern cybersecurity challenges in critical infrastructure, it is essential to distinguish between two fundamental types of network traffic: *Perimeter Traffic (North-South Traffic)* and *Lateral Traffic (East-West Traffic)*

Perimeter firewalls primarily inspect inbound and outbound traffic, known as north-south traffic (Palo Alto Networks, 2025). This traffic crosses the boundary between an organization's internal network and external networks such as the Internet. Traditional security models focused heavily on perimeter defense, using firewalls and intrusion detection systems to monitor and control what enters or leaves the network. The network perimeter is the boundary between an organization's secured internal network and the Internet or any other uncontrolled external network—in other words, the edge of what an organization has control over (Cloudflare, 2025).

Traffic moving laterally within the network, or east-west traffic, may not be monitored by perimeter firewalls, potentially allowing internal threats to go undetected (Palo Alto Networks, 2025). Lateral movement refers to the techniques attackers use to navigate through a network after gaining initial access, moving from system to system to identify and compromise additional assets (Illumio, 2025). Lateral movement is observed in 25% of cyberattacks, with nine out of ten organizations currently exposed to at least one attack path and 80% having paths exposing critical assets (Zero Networks, 2025).

The distinction between perimeter and lateral traffic is critical for critical infrastructure security because most cybersecurity investments historically concentrated on perimeter defenses—firewalls, intrusion detection systems, and access controls at the network edge. Legacy security infrastructures are generally flat network architectures that rely on a perimeter firewall as their only point of traffic inspection and control. Since network boundaries don't exist as they used to, and most data center traffic is east-west, traditional port-based firewalls provide limited value (Palo Alto Networks, 2025).

But attackers exploit internal movement. When threats penetrate the perimeter, they are free to move laterally in the network to access virtually any data, application, asset or services. With virtually unhindered access, attackers can easily exfiltrate a full range of valuable assets, often before the breach has even been detected (Palo Alto Networks, 2025). Effective AI/ML threat detection requires monitoring both perimeter (north-south) and lateral (east-west) traffic. Monitoring internal (east-west) traffic is critical to preventing lateral movement. Most organizations are good at perimeter defenses, but once inside, attackers have free rein (Illumio, 2025). Machine learning systems must analyze network flow patterns to detect unauthorized lateral movement, suspicious command-and-control communications, and abnormal connections to Human-Machine Interfaces (HMIs).

The convergence of IT and OT networks creates particularly concerning risks related to lateral movement. At present, OT networks have an inconsistent deployment of security policies and standards wherein IT networks have strong security policies; this is in part due to the timing for the build and upgrades to both IT and OT networks. Industrial control systems supervise physical processes through sensor data and perform remote monitoring, control, and diagnostic functions, but ICS cyber threats are growing at an alarming rate on industrial automation applications. Notable incidents illustrate the severity of these threats: the Stuxnet worm targeting Iranian nuclear facilities, the BlackEnergy attack causing Ukrainian power outages, and the Triton malware compromising Saudi Aramco safety systems (Langner, 2011; Lee et al, 2016; Sarker et al, 2024).

Limitations of Traditional Risk Mitigation Frameworks

Traditional risk mitigation follows a structured framework encompassing risk identification, analysis, treatment, and monitoring (Maglaras et al, 2023). However, this approach faces critical limitations when applied to modern critical infrastructure environments. *Scale challenges* emerge from the millions of sensors and data points requiring continuous monitoring across geographically dispersed assets with complex interdependencies (Xiang et al, 2025). *Speed challenges* manifest as human analysis proves too slow for real-time threat response, with security teams overwhelmed by alert fatigue and time-to-detection measured in months rather than minutes (Linkov & Kott, 2019). *Complexity challenges* arise from sophisticated multi-stage attacks employing zero-day vulnerabilities, encrypted payloads, and polymorphic malware that evade signature-based detection (Mishra et al, 2022). *Resource challenges* include severe shortages of skilled cybersecurity professionals, budget constraints, legacy system maintenance burdens, and competing operational priorities (Elmaghraby & Losavio, 2014).

The AI/ML Value Proposition for Critical Infrastructure

Artificial intelligence and machine learning offer a transformative response to the limitations of traditional risk mitigation strategies in critical infrastructure. ML models can learn from historical attack data to predict future threats, enabling organizations to take preemptive action. Unlike manual processes that struggle with scale and speed, AI-driven systems can continuously monitor vast networks of sensors and endpoints without fatigue or attention degradation, processing millions of events per second to identify anomalies in real time (Chowdhury, 2024). This capability dramatically reduces detection times from hours or days to mere seconds, enabling organizations to respond to threats before they escalate (Gujar, 2024).

Beyond speed, AI/ML excels in pattern recognition and predictive analytics. Advanced algorithms uncover subtle correlations and complex attack signatures that human analysts would likely miss, particularly in high-dimensional data environments (Sarker, 2023). By learning from historical data, machine learning models forecast potential failures and cyberattacks, shifting security postures from reactive to proactive (Basiru et al, 2023; Crawford et al, 2023). This predictive capability supports maintenance planning, resource optimization, and operational resilience (Kaur et al, 2023).

Automation is another critical advantage. AI reduces the burden of routine tasks such as log reviews and vulnerability scans, freeing skilled personnel to focus on strategic decision-making (Gugueoth et al, 2023). Furthermore, these systems adapt over time, continuously improving as they encounter new threats—an essential feature in dynamic environments where static rule-based systems quickly become obsolete. Continuous learning from new threats enables model improvement over time, unlike static rule-based systems that require manual updating (Mohamed, 2025).

Paradigm Shifts Enabled by AI/ML

The integration of artificial intelligence and machine learning into critical infrastructure protection represents more than a technological upgrade; it signals a fundamental transformation in how organizations approach security. Traditional methods have long relied on reactive measures, responding only after an incident has occurred. AI/ML systems, by contrast, enable a proactive posture, predicting and preventing disruptions before they materialize through anomaly detection and predictive analytics (Raman et al, 2024). This shift is equally evident in the movement from rules-based to behavior-based security. Signature matching, once the cornerstone of intrusion detection, is limited to identifying known threats. AI-powered anomaly detection expands this capability by recognizing previously unseen attack patterns through behavioral analysis, offering protection against zero-day exploits and novel adversarial techniques (Sarker et al, 2024).

Another paradigm change lies in the transition from periodic to continuous monitoring. Scheduled audits and inspections provide only snapshots of system health, leaving gaps that adversaries can exploit. AI systems, however, enable real-time visibility and continuous assessment, ensuring that infrastructure operators maintain an up-to-date understanding of their security posture (Chithaluru et al, 2023).

Finally, AI/ML fosters integration across traditionally siloed environments. IT and OT systems have historically been managed separately, creating blind spots in security coverage. AI-enabled platforms unify visibility across converged infrastructures, allowing operators to detect and respond to threats that traverse both digital and physical domains (Bruce, 2025). Together, these paradigm shifts demonstrate how AI/ML technologies are transforming the core principles of critical infrastructure security. While they offer unprecedented opportunities for resilience and efficiency, they also introduce new risks that demand careful governance and human oversight (Xiang et al, 2025).

However, these capabilities come with attendant risks and challenges. As critical infrastructure operators and providers seek to harness the benefits of new artificial intelligence capabilities, they must also manage associated risks from both AI-enabled cyber threats and potential vulnerabilities in deployed AI systems (Xiang et al, 2025). The following sections examine specific applications, implementation challenges, and governance considerations necessary for responsible AI deployment in critical infrastructure contexts.

AI/ML Applications in Critical Infrastructure Protection

There are five key applications of AI/ML technologies in critical infrastructure risk mitigation: real-time threat detection, predictive analysis for component failures, automated risk assessment tools, disaster-resilient infrastructure design, and common implementation challenges.

Real-Time Threat Detection Using Machine Learning

Real-time threat detection represents one of the most mature applications of ML in critical infrastructure security. Detection techniques with machine learning algorithms on public datasets, suitable for intrusion detection of cyber-attacks in SCADA systems, as the first line of defense, have been detailed. The goal is continuous monitoring of network traffic, system logs, and sensor data with immediate identification of suspicious patterns and automated alerting within seconds to minutes of detection (Inoue et al, 2017).

Three primary ML approaches enable real-time threat detection in critical infrastructure environments. *Supervised learning* employs classification algorithms including Random Forests, Support Vector Machines (SVM), and Neural Networks trained on labeled datasets distinguishing "normal" from "attack" traffic (Alsamiri & Alsubhi, 2019). Supervised learning is particularly effective for detecting known threats, such as phishing emails, malware, or Distributed Denial of Service (DDoS) attacks. Models can be trained on historical data to recognize patterns associated with these threats, allowing them to flag suspicious activities in real-time (Thawait, 2024). This approach excels at detecting known malware signatures and DDoS patterns but requires substantial labeled training data, which is often scarce in OT environments (Huda et al, 2018).

Unsupervised learning utilizes clustering algorithms such as K-means and DBSCAN for anomaly detection without predefined attack signatures (Kim et al, 2019). This approach proves particularly valuable for identifying novel attack patterns, such as unusual SCADA command sequences that deviate from established operational baselines (Inoue et al, 2017). Towards this end, a class of anomaly detectors, created using data-centric approaches, are gaining attention. Using machine learning algorithms such approaches can automatically learn the process dynamics and control strategies deployed in an ICS.

Deep learning techniques, including Convolutional Neural Networks (CNNs) for pattern recognition and Recurrent Neural Networks (RNNs) for time-series analysis, enable detection of subtle timing anomalies in industrial control signals (Huda et al, 2018). Deep Learning techniques appear as a suitable solution for detecting such complicated attacks. Long Short-Term Memory (LSTM) networks prove especially effective for modeling temporal dependencies in SCADA communications (Kim et al, 2019).

Machine learning has revolutionized real-time threat detection in critical infrastructure by enabling systems to identify and respond to cyber threats with speed and precision. Several practical applications illustrate its impact.

A major application is in network intrusion detection, where ML models continuously monitor both lateral (east–west) and perimeter (north–south) traffic. These systems detect unauthorized lateral movement, suspicious command-and-control communications, and abnormal connections to Human-Machine Interfaces (HMIs), which often signal early stages of an attack (Ni, 2023).

Another critical use case is endpoint behavior analysis. Here, machine learning examines execution patterns on engineering workstations to identify anomalies such as fileless malware, privilege escalation attempts, and unusual scripting activity, including PowerShell commands commonly exploited in advanced attacks (Linkov & Kott, 2019).

Scada/ICS protocol analysis represents a third application. ML algorithms parse industrial protocols, such as Modbus, DNP3, and OPC, to detect unauthorized read/write commands to Programmable Logic Controllers (PLCs), replay attacks, and compliance violations that could compromise physical processes (Anwar et al, 2021).

Finally, physical security integration combines video analytics with ML to strengthen perimeter defenses. These systems analyze camera feeds and sensor data to detect unauthorized personnel near substations, identify abandoned objects, and even track drones attempting to breach restricted zones (Akoglu et al, 2010).

Predictive Analysis for Critical Infrastructure Components

Predictive analysis leverages historical operational data to forecast future failures or degradation before they occur, enabling the transition from reactive and preventive to predictive maintenance paradigms (Yigit et al, 2025). This capability reduces unplanned downtime, optimizes resource allocation, and extends asset lifespan through data-driven interventions (Mohammed, 2025). This type of analysis can also be used to detect threats to vulnerabilities prior to successful attacks.

Three ML technique categories enable predictive analysis in critical infrastructure. *Time series forecasting* employs ARIMA models, Facebook's Prophet framework, and LSTM neural networks to predict equipment performance trends, such as forecasting transformer oil degradation rates based on temperature, load, and environmental factors (Yigit et al, 2025). *Survival analysis* utilizes Cox proportional hazards models and Weibull analysis to estimate time-to-failure probabilities and predict remaining useful life of critical components like turbines and generators (Casalicchio et al, 2010; Song & Kawai, 2023). *Regression models* including linear, polynomial, and ensemble methods correlate sensor data with specific failure modes, such as predicting bearing wear from vibration signature analysis (Lang et al, 2024).

Key applications demonstrate predictive analysis value in critical infrastructure contexts. *Predictive maintenance* monitors equipment health metrics—vibration, temperature, oil quality, electrical characteristics—to schedule maintenance interventions before failures occur, reducing emergency repairs and optimizing spare parts inventory (Yigit et al, 2025). This approach can predict generator failures 2-4 weeks in advance, enabling scheduled downtime rather than emergency outages.

Grid load forecasting predicts electricity demand patterns to optimize generation dispatch, balance loads across distribution networks, and effectively integrate renewable energy sources through day-ahead wind and solar generation forecasts (Yigit et al, 2025).

Infrastructure degradation modeling assesses structural health of bridges, pipelines, and dams to prioritize repair and replacement investments while ensuring proactive compliance with safety regulations, such as predicting pipeline corrosion rates based on age, material, soil conditions, and operating parameters (Song & Kawai, 2023). *Cascading failure prevention* model interdependencies between infrastructure systems to simulate failure propagation scenarios and implement preemptive load shedding or rerouting, such as preventing blackout cascades during severe weather events (Mohammed, 2025).

Automated Risk Assessment Tools and Uses

Automated risk assessment employs AI/ML for continuous, data-driven evaluation of security posture with dynamic risk scoring based on real-time threat intelligence (Sarker et al, 2024). This approach enables automated vulnerability prioritization, remediation recommendations, and integration with governance, risk, and compliance (GRC) frameworks. Three ML approaches facilitate automated risk assessment. *Bayesian networks* model probabilistic relationships between risk factors, updating risk estimates as new evidence emerges to calculate the likelihood of successful cyberattacks given current vulnerabilities, threat intelligence, and defensive posture (Sarker et al, 2024). *Natural language processing (NLP)* parses threat intelligence feeds, Common Vulnerabilities and Exposures (CVE) databases, and security advisories to extract actionable insights from unstructured data, automatically correlating newly disclosed vulnerabilities with asset inventories (Raman et al, 2024). *Reinforcement learning* optimizes security control configurations by learning optimal defensive strategies through simulation, such as adaptive firewall rule optimization that balances security and operational requirements (Sarker et al, 2024).

Automated risk assessment in critical infrastructure environments demonstrates how AI/ML can transform static, manual evaluations into dynamic, continuous processes that adapt to evolving threats. By leveraging probabilistic modeling, natural language processing, and reinforcement learning, organizations can move beyond traditional checklists and audits toward systems that provide real-time situational awareness and actionable insights. Several key use cases illustrate the breadth of these applications.

Vulnerability management is one of the most immediate benefits of automation. AI-driven systems continuously discover assets, scan for vulnerabilities, and correlate findings with contextual information such as asset criticality and operational relevance. Unlike traditional patch management, which often prioritizes vulnerabilities based solely on severity scores, automated risk assessment tools can weigh the importance of OT-specific vulnerabilities differently from those focused on IT, ensuring that remediation efforts align with operational priorities (Musa et al, 2024).

Threat modeling automation extends this capability by simulating potential attack paths across converged IT/OT networks. Using frameworks such as MITRE ATT&CK for ICS, AI systems can identify chokepoints, single points of failure, and likely propagation routes for ransomware or advanced persistent threats. This allows operators to anticipate adversary behavior and implement preemptive defenses, such as segmentation or adaptive firewall rules, before attacks occur (Assante & Lee, 2015).

Compliance automation addresses the growing burden of regulatory requirements in critical infrastructure sectors. AI-enabled platforms can continuously monitor systems against standards such as NERC CIP or TSA directives, automatically collect evidence for audits, and generate dashboards that highlight compliance gaps. This reduces the manual workload on compliance teams, ensuring that organizations maintain regulatory alignment even as systems evolve (Sarker et al, 2024).

Third-party risk assessment is increasingly vital as supply chains become more interconnected. AI/ML tools can evaluate vendor security postures, monitor for emerging threats in supplier ecosystems, and analyze responses to security questionnaires. By automating these processes, organizations gain visibility into risks introduced by external partners, reducing the likelihood of compromised components entering critical systems (Xiang et al, 2025). AI-enabled resilient design optimizes infrastructure for robustness and recovery by incorporating redundancy, diversity, and adaptability principles informed by simulation of disaster scenarios to identify vulnerabilities (Maglaras et al, 2023).

Machine Learning Techniques for Resilient Design

Three ML approaches support disaster-resilient infrastructure design. *Optimization algorithms* including genetic algorithms and particle swarm optimization design redundant systems with minimal cost, determining optimal placement of backup generators, battery storage, and redundant communication links while balancing capital expenditure against availability requirements (Mohammed, 2025). *Agent-based modeling* simulates complex system behaviors under stress conditions, tests recovery strategies in virtual environments, and models evacuations during chemical plant incidents or power grid failures (Casalicchio et al, 2010). *Digital twin technology* creates virtual replicas of physical infrastructure to test design changes and incident responses safely, simulating grid behavior during cyberattacks or equipment failures without risking actual operations (Maglaras et al, 2023).

Four applications demonstrate AI-enabled resilient design. *Redundancy optimization* identifies optimal backup system configurations, balances cost against availability requirements, and designs for graceful degradation rather than catastrophic failure, comparing N+1 versus 2N uninterruptible power supply (UPS) configurations for data centers (Mohammed, 2025).

Network topology optimization designs mesh networks with multiple failover paths, minimizes single points of failure, and optimizes for both efficiency and resilience through smart grid microgrid segmentation strategies (Yigit et al, 2025).

Automated disaster response implements preprogrammed response playbooks triggered by AI detection, enables autonomous load shedding, rerouting, and isolation, and coordinates responses across multiple infrastructure systems, such as automatic water system isolation during contamination detection (Maglaras et al, 2023).

Climate adaptation planning model infrastructure performance under future climate scenarios, identifies assets at risk from sea level rise and extreme weather, and prioritizes hardening investments, such as flood risk assessment for coastal substations (Yigit et al, 2025).

Common Challenges in AI/ML Implementation

Despite significant promise, AI/ML deployment in critical infrastructure faces six major challenges that organizations must address for successful implementation. Sensor malfunctions produce erroneous readings, network latency causes data gaps, legacy systems lack standardized logging, and environmental interference from electromagnetic interference (EMI) and physical obstructions degrades data quality. These issues cause training on corrupted data to produce unreliable models, anomaly detectors triggered by sensor noise rather than genuine threats, missing data that prevents pattern recognition, and model performance degradation over time (Inoue et al, 2017).

Mitigation strategies include data quality monitoring and cleansing pipelines, sensor calibration and validation protocols, imputation techniques for missing data such as forward fill and model-based approaches, ensemble methods robust to noise, and transfer learning from similar but cleaner datasets (Kim et al, 2019). Vulnerabilities and information of successful attacks are not currently logged in a central database that can be used to improve ML and modeling techniques.

A critical challenge emerges from the gap between data science and operational technology expertise. Data scientists often lack understanding of OT/ICS operations, while OT engineers lack familiarity with ML algorithms, creating challenges in feature engineering that requires deep domain knowledge and model interpretation needing operational context (Xiang et al, 2025). This expertise gap results in poorly chosen features missing critical signals, models optimized for wrong objectives, false positives from misunderstanding "normal" operations, and inability to explain model decisions to stakeholders (Sarker et al, 2024). This gap can also result in incomplete vulnerability modeling based on missing pattern detection derived from ML.

Mitigation approaches include cross-functional teams combining data scientists with domain experts, extensive feature engineering workshops involving operational staff, explainable AI (XAI) techniques for model transparency, simulation environments enabling data scientists to learn OT processes, and subject matter expert validation at every development stage (Xiang et al, 2025).

Another challenge manifests from the human dynamic. IT and OT teams typically operate in organizational silos with divergent priorities—security versus availability—creating cultural conflicts, resistance to change from operational staff, and procurement and deployment approval delays (Sarker et al, 2024). These barriers result in models developed but never deployed, insufficient integration with existing workflows, lack of buy-in from end users, under resourced mitigation investment, and inadequate maintenance and updates (Xiang et al, 2025).

Mitigation strategies encompass executive sponsorship mandating collaboration, joint IT/OT security operations centers, gradual deployment through pilot programs, change management and training initiatives, and clearly defined roles and responsibilities using RACI (Responsible, Accountable, Consulted, Informed) matrices (Maglaras et al, 2023).

Another major challenge comes from “false positives.” ML models frequently err on the side of caution and/or lack of domain expertise in either the business and/or cyber security, generating excessive alerts that overwhelm security teams, bury real threats in noise, and lead operators to ignore alerts through "cry wolf" effects (Kim et al, 2019). This phenomenon reduces responsiveness to genuine threats, wastes investigation time and resources, degrades trust in AI systems, and creates potential for complete system disablement (Anwar et al, 2021).

Mitigation approaches include threshold tuning and Receiver Operating Characteristic (ROC) curve optimization, multi-stage alert validation combining automated and human review, alert prioritization and risk scoring, continuous model retraining with feedback loops, integration with Security Orchestration, Automation, and Response (SOAR) platforms, and anomaly ranking rather than binary classification (Chicco & Jurman, 2020).

Conceptual drift is another major issue. Infrastructure environments evolve through new equipment deployments, configuration changes, and operational pattern modifications, causing ML models trained on historical data to become outdated as adversaries adapt evasion techniques (Sarker et al, 2024). Consequences include declining detection accuracy, increased false negatives missing threats, increased false positives flagging normal changes as anomalies, and continuous retraining requirements (Huda et al, 2018). This impact is exacerbated by cyber security not always being designed into new equipment deployments required to increase operating efficiency. Mitigation strategies encompass continuous monitoring of model performance metrics, automated retraining pipelines, online learning algorithms adapting in real-time, A/B testing of model versions, version control with rollback capabilities, and regular model audits and validation (Kim et al, 2019).

Lastly, sophisticated attackers can poison training data through data integrity attacks, craft evasion attacks with inputs designed to bypass detection, perform model inversion revealing sensitive training data, and exploit transferability of attacks across similar models (Guembe et al, 2022). These attacks reduce effectiveness against sophisticated adversaries, create potential for catastrophic failures, and raise privacy concerns around model training data. Mitigation approaches include adversarial training incorporating attack samples in training datasets, ensemble methods more difficult to evade simultaneously, input validation and sanitization, differential privacy techniques, model monitoring for unusual prediction patterns, and red team exercises testing robustness (Sarker et al, 2024).

Trade-offs in AI/ML System Design

The deployment of AI/ML in critical infrastructure requires navigating complex trade-offs between competing objectives. This section analyzes six critical trade-off dimensions that shape system design decisions.

Speed versus Accuracy

Speed demands detection and response to threats in real-time (milliseconds to seconds), while accuracy requires minimizing false positives and false negatives through complex models and extensive data analysis (Chicco & Jurman, 2020). These objectives conflict because complex models such as deep neural networks provide higher accuracy but slower inference times, real-time processing requires lightweight models that may sacrifice accuracy, and thorough analysis consumes time that attackers can exploit (Huda et al, 2018).

The critical infrastructure context intensifies this trade-off. Industrial processes operate on tight timing constraints where milliseconds matter, false positives can trigger unnecessary shutdowns costing millions in losses, false negatives allow attacks to succeed undetected, and speed matters most in life-safety scenarios such as chemical plants and nuclear facilities (Maglaras et al, 2023).

Organizations should *prioritize speed* when life-safety implications demand immediate response, high-frequency events require real-time decisions, attackers can act quickly through automated malware propagation, such as autonomous emergency shutdown systems (Sarker *et al*, 2024). Organizations should *prioritize accuracy* when high costs of false positives create operational disruption, time is available for human validation, and consequences of false negatives are catastrophic, such as differentiating scheduled from emergency maintenance decisions (Kim et al, 2019). *Balanced approaches* employ multi-tiered detection with fast initial screening followed by deeper analysis, risk-based thresholds applying speed for high-risk assets and accuracy for others, and ensemble methods combining fast and accurate models (Anwar et al, 2021).

Availability versus Security

Availability demands continuous system operation (99.99%+ uptime), while security controls may interrupt operations through patching, scanning, and incident response (Maglaras et al, 2023). These objectives conflict because security patches require downtime for testing and deployment, monitoring tools add network latency, incident response may necessitate isolating critical systems, and defensive measures such as firewalls and segmentation can block legitimate traffic (Sarker et al, 2024).

The critical infrastructure context exacerbates this tension. Utilities cannot afford unplanned outages due to regulatory penalties and customer impact, healthcare systems prioritize patient care over security updates, manufacturing downtime is measured in thousands of dollars per minute, and traditional IT security practices including frequent patching and aggressive scanning prove incompatible with OT requirements (Xiang et al, 2025).

Organizations should *prioritize availability* when life-safety systems cannot be taken offline, high operational costs of downtime exist, compensating controls are available through network segmentation and monitoring, such as running hospital life-support systems on unpatched operating systems (Maglaras et al, 2023). Organizations should *prioritize security* when known active threats target the system, regulatory requirements mandate controls, and systems are not in critical paths where isolation or replacement is possible, such as immediately patching internet-facing OT gateways (Sarker et al, 2024).

Balanced approaches implement defense-in-depth with multiple layers compensating for individual weaknesses, risk-based patching addressing only critical vulnerabilities with thorough testing, non-intrusive monitoring using passive network taps and read-only access, planned maintenance windows for updates, and redundant systems enabling patching of one while another operates (Xiang et al, 2025).

AI/ML contributions include predictive maintenance scheduling downtime during low-demand periods, anomaly detection providing security without intrusive scanning, automated risk assessment identifying truly critical patches, and virtual patching through Web Application Firewalls (WAF) and Intrusion Prevention System (IPS) rules providing interim protection (Yigit et al, 2025).

Recent advances in computer vision and object recognition software have enabled automated monitoring and situational awareness across critical infrastructure sectors. Object recognition, driven by machine vision software, facilitates real-time detection of unauthorized intrusions, facial recognition, abandoned objects, and equipment anomalies on camera feeds, drastically reducing the need for human oversight. Modern deployments frequently combine edge computing and “on-device” AI to deliver scalability, rapid response and robust privacy- critical for surveillance and operational technology safety. Video analytics applications allow for unattended object detection, virtual fencing and automated alarm systems in infrastructure environments, ensuring threat detection and emergency responses (Sharma et al, 2021; Zhou et al, 2017). Leading platforms such as *Lumana AI*, leverage advanced image processing and deep learning algorithms to support industrial automation, quality control, and security monitoring.

Explainability versus Performance

Explainability requires humans to understand why ML made decisions through interpretability and transparency, while performance seeks highest possible accuracy often through complex "black box" methods (Sarker, 2023). These objectives conflict because simple models such as decision trees and linear regression are interpretable but less accurate, complex models including deep neural networks and gradient boosting are accurate but opaque, and feature interactions in ensemble methods defy simple explanation (Huda et al, 2018).

The critical infrastructure context demands explainability. Operators must trust and act on ML recommendations, regulatory compliance may require audit trails and decision justification, debugging false positives requires understanding model logic, and liability concerns arise in safety-critical applications (Stevens, 2020).

Organizations should *prioritize explainability* when regulatory requirements mandate transparency, operators lack trust in ML systems, high-stakes decisions require justification, and model debugging and improvement are needed, such as explaining why ML flagged a sensor reading as anomalous (Sarker, 2023). Organizations should *prioritize performance* when detection accuracy is paramount, decisions are automatically executed without human review, and competitive advantage depends on model sophistication, such as high-frequency trading algorithms and real-time threat detection (Huda et al, 2018). Many of the issues relating to explainability should be alleviated if the initial modelling is well documented; this should result from the collaboration between the technologists and the operational engineers described above. Even black box ML decisions made in the absence of human intervention should be easily described if the initial model is well documented.

Balanced approaches employ explainable AI (XAI) techniques including SHAP (SHapley Additive exPlanations) values, LIME (Local Interpretable Model-agnostic Explanations), and attention mechanisms; model-agnostic interpretation methods; hybrid systems combining interpretable and complex models for consensus; and post-hoc explanations generated after decisions (Sarker, 2023).

Cost versus Capability

Cost constraints encompass budget limitations on hardware, software, and personnel, while capability demands advanced AI requiring expensive infrastructure and expertise (Xiang et al, 2025). These objectives conflict because high-performance computing including GPUs for deep learning is expensive, commercial AI platforms charge per sensor/endpoint/API call, data science talent commands premium salaries, and cloud computing costs scale with data volume (Sarker et al, 2024).

The critical infrastructure context intensifies cost pressures. Utilities operate on regulated rate-of-return creating capital constraints, public sector infrastructure including water and transportation has limited budgets, legacy systems may require costly upgrades for AI integration, and cybersecurity is often underfunded relative to operational priorities (Maglaras et al, 2023).

There is a second aspect of this type of tension which includes the fundamental trade-offs between operational efficiency upgrades and security capabilities. In a country with aging critical infrastructure in desperate need of both operational improvements and hardening capabilities, patch work modifications become rampant. This is due to the lack of political will to prioritize the huge investments necessary to build out (for instance) water distribution replacement and improvements to the electrical grid. These ad hoc improvements may lead to a lack of investment in security capabilities, especially in light of the fact that cyber capabilities are not always included in the initial design of the operational improvements and optimization Strategies

Infrastructure optimization balances cloud versus on-premises deployment for control and cost, employs open-source tools (scikit-learn, TensorFlow) versus commercial platforms, implements edge computing to process data locally reducing bandwidth costs, and leverages existing hardware without assuming GPU requirements (Sarker et al, 2024). *Personnel optimization* upskills existing staff versus hiring specialists, outsources model development while keeping operations in-house, partners with universities for research collaborations and internships, and pursues consortium approaches sharing costs across industry (Xiang et al, 2025).

Data optimization employs transfer learning reusing models trained on public datasets, generates synthetic data to augment limited real data, and implements federated learning for collaborative training without sharing data (Sarker et al, 2024). *Capability prioritization* focuses on high-ROI use cases first (predictive maintenance before exotic threat detection), adopts crawl-walk-run approaches starting simple and adding complexity as value is proven, and evaluates commercial off-the-shelf (COTS) solutions versus custom development (Yigit et al, 2025).

The lack of a regulatory environment that necessitates the reporting of successful cyber-attacks in a completely coordinated fashion and best practices to prevent them create challenges for individual companies to improve their security capabilities. The current administration’s reduction in infrastructure investments and the federal workforce that assists in the implementation of those investments will only exacerbate this issue.

Privacy versus Utility

Privacy demands protection of sensitive data about infrastructure, operations, and individuals, while utility requires extensive data for ML model training and improvement (Guembe et al, 2022). These objectives conflict because more data improves model accuracy but increases privacy risk, sharing data across organizations enhances threat intelligence but exposes operational secrets, detailed logs enable forensics but reveal patterns to adversaries, and centralized data facilitates ML training but creates attractive targets for attackers (Stevens, 2020).

The critical infrastructure context raises unique privacy concerns. SCADA data reveals operational secrets and vulnerabilities, employee behavior data raises workforce privacy concerns, cross-sector data sharing is restricted by competitive and national security considerations, and regulatory frameworks including GDPR and CCPA impose privacy obligations (Xiang et al, 2025).

The integration of artificial intelligence and machine learning into critical infrastructure environments introduces unique privacy challenges. Operational data collected from SCADA systems, IoT devices, and industrial sensors often contains sensitive information about system performance, vulnerabilities, and even personal data related to employees or customers. If exposed, this data could be exploited by adversaries to identify weaknesses or launch targeted attacks. Privacy-preserving techniques, therefore, play a critical role in ensuring that AI/ML systems can deliver analytical power without compromising confidentiality, regulatory compliance, or trust.

One of the most widely adopted approaches is *federated learning*, which allows models to be trained locally on distributed devices or systems. Instead of transmitting raw data to a central server, only model updates are shared, significantly reducing the risk of data leakage. This approach is particularly valuable in critical infrastructure contexts where operational data cannot leave secure environments due to regulatory or safety constraints (Kairouz et al, 2021). Trade-offs include communication overhead and coordination complexity, exemplified by utilities collaboratively training threat detection without sharing operational data (Sarker et al, 2024).

Another cornerstone method is *differential privacy*, which introduces statistical noise into datasets or model outputs to obscure individual data points while preserving overall utility (Stevens, 2020). This ensures that sensitive information, such as employee credentials or facility-specific operational metrics, cannot be reverse-engineered from AI outputs. Differential privacy has been widely applied in healthcare and financial services, and its adoption in critical infrastructure contexts is growing as organizations seek to balance transparency with confidentiality (Dwork & Roth, 2014). *Homomorphic encryption* represents a more advanced technique, enabling computations to be performed directly on encrypted data. This ensures that sensitive inputs remain confidential throughout the analytic process, even when processed by external systems or cloud environments. While computationally intensive, homomorphic encryption is increasingly being optimized for industrial applications where confidentiality is paramount (Acar et al, 2018). Trade-offs include significant computational overhead (10-1000x slower), such as third-party AI vendors analyzing data without seeing plaintext (Guembe et al, 2022).

Synthetic data generation creates artificial data preserving statistical properties for sharing instead of real data. Trade-offs include potentially missing real-world patterns, such as generating synthetic SCADA traffic for model training and testing (Sarker et al, 2024). *Data minimization* collects and retains only necessary data, aggregating before analysis (e.g., hourly averages versus second-by-second readings). Trade-offs include less granular data potentially missing subtle patterns, such as anonymizing IP addresses in logs while retaining only port and protocol information (Xiang et al, 2025).

Secure multi-party computation (SMPC) further enhances collaborative risk analysis by allowing multiple stakeholders to jointly compute functions over their inputs without revealing the inputs themselves. For example, utilities across different regions could collectively assess systemic vulnerabilities without exposing proprietary operational data to one another (Evans et al, 2018). Finally, *trusted execution environments* (TEEs) provide hardware-based secure enclaves that isolate sensitive computations from the rest of the system. TEEs protect against insider threats and malware by ensuring that critical processes, such as anomaly detection or predictive maintenance, are executed in a secure and verifiable environment (Sabt et al, 2015).

Together, these privacy-preserving techniques form a layered defense strategy that enables AI/ML systems to operate effectively in sensitive environments. By embedding federated learning, differential privacy, encryption, and secure enclaves into system architectures, organizations can achieve both analytical power and regulatory compliance. This balance is essential for fostering trust in AI-enabled critical infrastructure, ensuring that innovation does not come at the expense of confidentiality or resilience.

Automation versus Human Control

Automation enables AI to make decisions and act autonomously providing speed, consistency, and scalability, while human control requires humans to review and approve actions ensuring accountability, judgment, and safety (Maglaras et al, 2023). These objectives conflict because automated response provides fastest mitigation critical for fast-moving threats, human validation prevents mistakes from becoming disasters, trust and accountability require human oversight, yet speed of cyber-attacks may exceed human reaction time (Sarker et al, 2024).

The critical infrastructure context intensifies this trade-off. Automated shutdowns can prevent catastrophic failures but also cause unnecessary outages, false positives are more costly when automation acts without validation, regulatory frameworks often require human accountability, and adversaries can exploit predictable automated responses (Xiang et al, 2025).

The Sheridan-Verplank scale provides five automation levels (Parasuraman et al, 2000):

i. *Level 1: Manual* - AI provides raw data only with humans analyzing and deciding, such as security logs presented to analysts.

ii. *Level 2: Advisory* - AI recommends actions with humans deciding whether to accept, such as "Suggest isolating this device?"

iii. *Level 3: Consent Required* - AI proposes and prepares to execute with human approval required, such as "Ready to block this IP - approve?"

iv. *Level 4: Veto Available* - AI executes after brief delay with humans able to cancel if present, such as "Blocking in 60 seconds unless canceled."

v. *Level 5: Fully Autonomous* - AI decides and acts immediately with humans informed after the fact, such as automatic emergency shutdown.

Organizations should employ *fully autonomous automation (Level 5)* when life-safety implications demand immediate action, high confidence in detection exists through validated well-understood scenarios, actions are reversible without harm, and speed is critical where human review is too slow, such as emergency reactor shutdown on anomaly detection (Maglaras et al, 2023).

Organizations should employ *human-in-the-loop (Levels 2-4)* when consequences of false positives are severe, novel or ambiguous situations require judgment, regulatory requirements mandate human oversight, and building trust in new ML systems is necessary, such as blocking critical suppliers based on anomaly alerts (Sarker et al, 2024).

A graduated automation approach implemented in phases builds trust: Phase 1 (6 months) employed advisory mode only with AI recommending while humans always decided; Phase 2 (6 months) required consent for low-risk actions such as blocking known-bad IPs; Phase 3 (ongoing) made veto available for medium-risk actions with consent for high-risk actions. Results included building trust gradually, reducing alert fatigue, and preventing automation disasters (Kim *et al*, 2019).

The irony of automation poses a critical challenge: as systems become more automated, humans become less practiced at manual intervention, yet human expertise is most needed during automation failures (Parasuraman & Riley, 1997). Critical infrastructure implications include operators losing situational awareness when AI handles routine decisions, skills atrophying when rarely exercised, humans struggling to intervene quickly during automation failures, and over-reliance on AI creating single points of failure (Maglaras et al, 2023). In “auto-drive cars” for example, demanding human interaction with the AI based driving systems might not actually be viable. As drivers become more reliant on the hands-off driving function, expecting the driver to intervene in situations where AI fails to anticipate, may be unreasonable.

Mitigation strategies include regular manual drills and simulations, maintaining "human-on-the-loop" rather than "out-of-the-loop" roles, implementing transparent AI enabling operators to understand automation actions, ensuring graceful degradation with manual mode always available, and continuous training on manual procedures (Sarker et al, 2024).

Limitations, Ethics, Governance and Best Practices

Limitations

There are fundamental limitations of AI/ML technologies in critical infrastructure contexts, ethical considerations arising from deployment, and governance frameworks necessary for responsible implementation. Machine learning is fundamentally constrained by training data quality and availability. Cyberattacks on critical infrastructure are rare, providing limited attack examples for supervised model training. Historical data may not represent future threats, particularly zero-day attacks having no training examples by definition (Inoue et al, 2017).

Data bias emerges when training data from one environment fails to generalize to others, seasonal and operational variations are not captured in limited datasets, and historical bias perpetuates as models learn past mistakes, such as models trained on summer data failing during winter operations (Kim et al, 2019). Label inaccuracy compounds these issues as mislabeled data teaches incorrect patterns, ambiguous cases create confusion about whether incidents represent attacks or mistakes, and hindsight bias influences labeling when outcomes are known, such as equipment failures mislabeled as normal operations corrupting models (Huda et al, 2018).

Another challenge develops when institutions in similar industries employ machine learning in isolated environments. For example, law firms use the same case law in attempts to add value to their specific clients, as opposed to sharing ML with other law firms using the same data sets. This will lead to the use of redundant IT resources and an inherent bias in specific interpretations of the case law. It will also invariably lead to content ownership issues, as individual firms attempt to monetize their “value added” using their specific domain expertise.

Machine learning recognizes patterns from training data but struggles with genuinely new situations. Advanced Persistent Threats (APTs) use novel techniques to evade detection, zero-day exploits by definition have no historical patterns, black swan events fall outside models' experience, and adversarial innovation outpaces model retraining (Guembe et al, 2022). Stuxnet (2010) exemplifies this limitation, employing a novel attack chain never seen before that traditional anomaly detection would struggle to recognize (Langner, 2011). Mitigation strategies include ensemble methods combining multiple detection approaches, out-of-distribution detection recognizing when inputs are unlike anything seen before, human oversight for high-confidence novel patterns, and red team exercises discovering blind spots. Fundamentally, organizations must accept that AI augments rather than replaces human expertise for truly novel threats (Sarker et al, 2024).

Complex ML models function as "black boxes" whose decisions are difficult to interpret. This matters for critical infrastructure because operators must trust and act on ML recommendations, debugging false positives requires understanding model reasoning, regulatory audits may require decision justification, safety-critical decisions need transparent logic, and adversaries can exploit unexplainable vulnerabilities (Sarker, 2023). Technical challenges include deep neural networks with millions of parameters, non-linear complex feature interactions, ensemble methods combining multiple models compounding opacity, and real-time decisions precluding detailed analysis. Consequences include "AI said so" providing insufficient justification for critical actions, operators potentially overriding or ignoring unexplainable alerts, difficulty improving models without understanding failures, and liability concerns in safety-critical applications (Stevens, 2020).

In addition to challenges in explaining or documenting the AI decision making process, there are generally obstacles in determining the efficacy in the target datasets used to facilitate ML. For example, when organizations attempt to determine the potential damages from a successful attack, the available data describing the impact of like attacks might not be available; this can make the prioritization of decisions that determine when and how to prevent such attacks, problematic. ML models face deliberate deception or manipulation through four attack types. *Evasion attacks* involve attackers crafting inputs to bypass detection, such as slightly modifying malware to evade signature-based detection or creating adversarial examples with imperceptible changes fooling classifiers (Guembe et al, 2022). *Poisoning attacks* corrupt training data causing models to learn incorrect patterns, such as injecting false normal behavior making real attacks appear normal. Model inversion attacks query models to extract training data, creating privacy breaches revealing sensitive operational information such as inferring SCADA network topology from model responses. *Model stealing* recreates proprietary models through queries, enabling offline analysis to find vulnerabilities such as extracting security vendors' threat detection models (Sarker et al, 2024).

Critical infrastructure implications are severe: nation-state adversaries possess resources for sophisticated attacks, long dwell times allow patient adversarial exploration, and high-value targets justify significant attacker investment. Mitigation includes adversarial training incorporating attack samples in training datasets, ensemble methods difficult to evade simultaneously, input validation and sanitization, differential privacy techniques, model monitoring for unusual prediction patterns, and red team exercises (Guembe et al, 2022).

Institutions adverse to negative publicity or regulatory scrutiny will contribute to the lack of complete transparency of the impact of successful attacks, and stifle attempts to share best practices among participants in similar industries, causing less optimal solution development in mitigating known vulnerabilities.

Critical infrastructure environments also impose practical constraints on ML deployment. Real-time requirements mean industrial processes operate on millisecond time scales where ML inference must complete within tight deadlines, yet complex models may be too slow, such as protective relays needing to act within 16ms (one power cycle) (Maglaras et al, 2023). Limited computing resources exist as OT devices including PLCs and RTUs have minimal processing power insufficient for sophisticated ML models locally, network bandwidth constraints limit cloud-based inference, such as legacy SCADA systems with 56K modems (Sarker et al, 2024). Harsh environments involve extreme temperatures, vibration, and electromagnetic interference limiting hardware options for edge ML deployment with reliability requirements exceeding typical IT standards, such as substation equipment operating at -40°C to +85°C (Maglaras et al, 2023). Long lifecycles mean OT equipment operates for 20-30 years requiring ML models and infrastructure maintainable over decades with technology obsolescence risk, exemplified by Windows XP remaining common in OT environments where ML tools may not provide support (Xiang et al, 2025).

Ethical Considerations

ML models can perpetuate or amplify societal biases with three critical infrastructure manifestations. *Resource allocation* involves predictive maintenance prioritizing some assets over others, raising questions about whether underserved communities receive less proactive maintenance, such as rural substations receiving fewer upgrades than urban ones due to ROI calculations (Sarker, 2023).

Threat detection bias emerges when models trained on historical incidents reflect biased investigation patterns, questioning whether certain employee groups are disproportionately flagged as insider threats, such as foreign-born employees flagged more often due to biased training data (Stevens, 2020). *Access and service* issues arise when AI optimizes grids for profitability potentially disadvantaging low-income areas, questioning whether AI perpetuates inequitable infrastructure investment such as smart grid benefits concentrated in affluent neighborhoods (Xiang et al, 2025).

Mitigation strategies include fairness metrics measuring demographic parity and equalized odds, bias audits of training data and model outputs, diverse development teams identifying blind spots, community engagement in AI deployment decisions, and regulatory frameworks requiring fairness assessments (Sarker, 2023).

Determining responsibility when AI makes consequential mistakes raises complex accountability questions. When ML models fail to detect attacks resulting in infrastructure damage, unclear responsibility exists among data scientists, operators who trusted the system, vendors, and management, with legal liability ambiguous for AI-driven decisions (Stevens, 2020).

False positive harm occurs when ML triggers unnecessary shutdowns costing millions in lost production, raising questions about who compensates for losses and faces consequences, with insurance and indemnification unclear for AI errors. Automated actions taken by fully autonomous systems without human review create accountability gaps when actions prove wrong, with moral hazard tempting blame on "the algorithm" (Xiang et al, 2025).

Ethical frameworks address these challenges through meaningful human control principles maintaining ultimate human authority, traceability providing audit trails documenting AI decisions and human oversight, clear responsibility assignment using RACI matrices for AI systems, insurance and liability frameworks adapted for AI, and regulatory requirements for AI accountability in safety-critical applications (Sarker, 2023; Stevens, 2020).

Stakeholders affected by AI may not understand or consent to its use, raising concerns in three contexts. *Employee monitoring* involves AI analyzing employee behavior for insider threat detection where workers may not know they're being monitored, potentially violating privacy and creating surveillance cultures eroding trust, questioning whether employees have rights to know and consent (Stevens, 2020).

Public impact emerges when AI optimizes grid operations affecting millions of citizens not consulted on AI deployment in public infrastructure, with mistakes impacting communities having no say, questioning what level of public engagement is required. *Third-party vendors* deploy AI systems where asset owners may not fully understand capabilities or limitations, with "trust us" black-box solutions questioning how much transparency vendors can be required to provide (Xiang et al, 2025).

Best practices include transparency reports explaining AI use in critical infrastructure, public comment periods for significant AI deployments, worker notification and consultation for monitoring systems, explainable AI making systems more transparent, and independent audits of high-impact AI systems (Sarker, 2023).

The potential to misuse AI is enormous and the moral jeopardy is immense. Artificial intelligence and machine learning technologies are inherently *dual-use*, meaning they can be applied for both beneficial and harmful purposes. In critical infrastructure contexts, this duality creates profound governance challenges. The same algorithms that enable anomaly detection, predictive maintenance, and disaster resilience can also be weaponized by adversaries to exploit vulnerabilities, design sophisticated attacks, or destabilize essential services.

One primary concern is *adversarial exploitation*. Attackers can poison training datasets or craft adversarial inputs that bypass detection systems, effectively turning defensive AI into a liability. For example, models trained to identify abnormal SCADA traffic can be manipulated to misclassify malicious commands as benign, undermining trust in automated defenses (Guembe et al, 2022).

Another risk lies in the development of *autonomous cyber weapons*. Agentic AI systems, designed to act independently in defensive contexts, could be repurposed to launch coordinated offensive operations without human oversight. Such systems might autonomously probe networks, exploit vulnerabilities, and propagate malware at speeds far beyond human capacity, raising ethical and strategic concerns about escalation in cyberspace sabotage (Xiang et al, 2025).

Surveillance infrastructure misuse emerges when critical infrastructure monitoring systems could be repurposed for population surveillance, with network traffic analysis tracking individual activities, such as smart grid data revealing household behaviors (Stevens, 2020). *Synthetic data misuse* is also a growing threat. Generative AI models, which can produce realistic sensor data or simulate infrastructure performance, may be exploited to create falsified operational logs. These synthetic datasets could deceive monitoring systems, mask intrusions, or mislead operators during incident response (Bommasani et al, 2021).

Finally, *infrastructure sabotage through optimization inversion* highlights the dual-use dilemma. AI tools designed to optimize grid performance or pipeline efficiency can be inverted to identify single points of failure. An adversary could use these insights to deliberately overload systems, trigger cascading failures, or disrupt essential services such as energy or water distribution (Rolnick et al, 2019).

The weaponization of AI/ML underscores the need for robust governance frameworks. International agreements, export controls, and ethical guidelines must ensure that advanced AI capabilities are not diverted toward malicious ends. Transparency, explainability, and human-in-the-loop oversight remain essential safeguards. Moreover, organizations must adopt red-teaming practices, adversarial testing, and continuous monitoring to anticipate how their own AI systems might be exploited.

Governance approaches include export controls on advanced critical infrastructure security AI, ethical guidelines for security researchers emphasizing responsible disclosure, international norms against critical infrastructure targeting, security by design preventing unauthorized repurposing, and distinction between defensive and offensive AI capabilities (Stevens, 2020; Sarker et al, 2024).

Governance Frameworks

The existing regulatory environment encompasses sector-specific and cross-sector frameworks. *Energy sector* regulations include NERC CIP (Critical Infrastructure Protection) and Federal Energy Regulatory Commission (FERC) cybersecurity mandates. *Water sector* operates under America's Water Infrastructure Act and state-level regulations. *Transportation* complies with Transportation Security Administration (TSA) pipeline security directives and Federal Aviation Administration (FAA) drone regulations. *Healthcare* adheres to Health Insurance Portability and Accountability Act (HIPAA) security rules and Food and Drug Administration (FDA) medical device cybersecurity requirements (Sarker *et al*, 2024; Xiang et al, 2025).

Cross-sector frameworks include NIST Cybersecurity Framework (voluntary in most sectors), CISA Critical Infrastructure Security Guidance, Executive Orders on improving critical infrastructure cybersecurity, and Department of Homeland Security (DHS) Chemical Facility Anti-Terrorism Standards (CFATS) (Xiang et al, 2025).

AI-specific emerging regulations encompass EU AI Act with risk-based classification and high-risk systems requirements, proposed Algorithmic Accountability Act in the United States, NIST AI Risk Management Framework, and IEEE standards for AI ethics (P7000 series) (Sarker, 2023; Xiang et al, 2025). The resource shift by this administration from the federal to state governments, along with an emphasis on deregulation, will serve to further diminish the centralized capabilities to protect our critical infrastructure from malicious foreign actors.

Technical standards provide implementation guidance through three categories. *ICS/OT security standards* include IEC 62443 (Industrial Automation and Control Systems Security), ISA/IEC 62443-4-2 (Technical Security Requirements for IACS Components), and NIST SP 800-82 (Guide to Industrial Control Systems Security) (Sarker et al, 2024). *AI/ML standards* encompass ISO/IEC 23894 (AI Risk Management), ISO/IEC 5338 (AI Lifecycle Management), and IEEE P2863 (Organizational Governance of AI) (Sarker, 2023). *Data standards* include ISO 27001 (Information Security Management), NIST Privacy Framework, and FAIR (Factor Analysis of Information Risk) (Xiang et al, 2025).

In addition to fewer federal resources, implementation challenges include standards lagging behind technology evolution, voluntary standards lacking enforcement mechanisms, conflicting standards across jurisdictions, and resource burdens on smaller organizations (Sarker et al, 2024).

Organizational policy mandates cane be effective. Essential policy components for AI governance include five categories. *AI use policy* defines approved use cases for AI in critical infrastructure, prohibited uses such as fully autonomous safety-critical decisions, risk assessment requirements before deployment, and approval workflows with authority levels (Xiang et al, 2025).

Data governance policy establishes data classification and handling requirements, retention and deletion schedules, access controls and audit logging, and third-party data sharing restrictions (Stevens, 2020). *Model governance policy* specifies model development lifecycle requirements, testing and validation standards, deployment approval processes, performance monitoring and retraining triggers, and model versioning with rollback procedures (Sarker, 2023).

Human oversight policy defines required human review for different risk levels, escalation procedures for anomalous AI behavior, override authority and documentation requirements, and training requirements for AI operators (Xiang et al, 2025). *Incident response policy* addresses AI-specific incident types including model failure and adversarial attacks, response procedures with responsible parties, communication protocols internally and externally, and post-incident review and improvement processes (Sarker et al, 2024).

Risk Management and Stakeholder Engagement are key to any successful governance policy. AI-specific risk assessment encompasses technical risks including model failure modes and consequences, adversarial attack surfaces, data poisoning vulnerabilities, and integration risks with legacy systems (Guembe et al, 2022). Operational risks involve over-reliance on AI causing skill degradation, alert fatigue and desensitization, misinterpretation of AI outputs, and automation bias (Maglaras et al, 2023). Strategic risks include vendor lock-in, technology obsolescence, competitive disadvantage if AI fails, and regulatory non-compliance (Xiang *et al*, 2025).

Risk treatment strategies employ redundancy through multiple detection methods beyond AI, diversity using different algorithms, vendors, and data sources, defense-in-depth with layered security treating AI as one component, resilience enabling graceful degradation when AI is unavailable, and insurance covering AI-related incidents (Sarker et al, 2024).

Stakeholder engagement encompasses internal stakeholders including executive leadership making strategic decisions on AI investment, operations teams using AI systems daily providing feedback, IT/OT security teams deploying and maintaining systems, and legal/compliance reviewing regulatory compliance (Xiang et al, 2025). External stakeholders include regulators verifying compliance, vendors providing AI systems, industry peers sharing best practices, public and customers with service expectations, and academia providing research and evaluation (Sarker, 2023; Sarker et al, 2024).

Governance Best Practices and Emerging Challenges

Effective governance practices include executive sponsorship ensuring C-suite champions AI initiatives, cross-functional teams collaborating from day one, starting small with gradual scaling through pilot programs, continuous monitoring of AI governance, transparency through explainable AI and open communication, human-centered design augmenting rather than replacing expertise, and regulatory engagement with proactive communication (Xiang et al, 2025; Sarker et al, 2024).

Common pitfalls to avoid include technology-first approaches deploying AI without clear use cases, siloed development with data scientists lacking operational input, "set and forget" deployments without ongoing monitoring, overpromising that AI will solve all problems, ignoring culture by imposing AI on resistant workforces, vendor over-reliance on black-box solutions, and compliance-only mindsets checking regulatory boxes without genuine security improvement (Sarker et al, 2024).

Emerging challenges include AI supply chain security concerns about vulnerabilities in open-source ML libraries, poisoned pre-trained models, and cloud AI service dependencies requiring mitigation through Software Bill of Materials (SBOM), model provenance tracking, and vendor security assessments (Guembe et al, 2022). AI system composition challenges arise from multiple AI systems interacting unpredictably, creating emergent behaviors and cascading failures requiring system-level testing and circuit breakers (Maglaras et al, 2023).

Geopolitical considerations involve foreign-developed AI in U.S. critical infrastructure raising supply chain risks, data sovereignty and localization requirements, export controls on advanced AI technology, and balancing international cooperation against national security concerns (Xiang et al, 2025). The evolving threat landscape demands continuous adaptation as adversarial AI capabilities advance rapidly, nation-states dedicate resources to AI-enabled attacks, and quantum computing threatens current encryption impacting AI security, requiring continuous threat intelligence, adaptive defenses, and quantum-resistant cryptography (Sarker et al, 2024; Guembe et al, 2022). The prioritization of foreign adversaries in AI related investments stands in contrast with the U.S. government's de-emphasis of these types of investments.

Recommendations and Directions of Future Research

Based on the findings presented in this article, several actionable recommendations emerge for critical infrastructure operators and security professionals implementing AI/ML systems, as well as several opportunities to contribute to the scholarship of this discipline.

Recommendations

Successful AI-integration projects *adopt a phased implementation approach*. Pilot programs should be limited in scope (10-15% of infrastructure) and operating in shadow mode to establish baselines and tune models before full deployment. Organizations should resist the pressure for "big bang" deployments that risk catastrophic failures that undermine stakeholder confidence. The project should be staffed by dedicated *cross-functional* teams, combining data scientists, OT engineers, security analysts, and operational staff with clearly defined roles and responsibilities. The expertise gap between AI specialists and domain experts represents one of the most significant implementation barriers. Regular feature engineering workshops involving operational staff, simulation environments enabling data scientists to learn OT processes, and subject matter expert validation at every development

The project focus should be on *implementing explainable AI techniques*. Deploy XAI methods including SHAP values, LIME, and attention mechanisms to provide human-readable explanations for ML decisions, particularly for high-stakes actions. Use complex ensemble models for detection accuracy combined with simplified decision trees for human-understandable explanations—performance and explainability need not be mutually exclusive. Explainability builds operator trust, enables debugging of false positives, and satisfies regulatory audit requirements.

Establish robust governance frameworks. Develop comprehensive policies covering AI use cases, data governance, model lifecycle management, human oversight requirements, and incident response procedures specific to AI failures. Organizations should move beyond compliance-only mindsets to create cultures of responsible AI deployment. Regular model audits, adversarial testing through red team exercises, bias assessments, and fairness metrics should become standard practice.

Secure executive sponsorship from both IT and OT leadership to *address organizational barriers* and mandate collaboration across traditional silos. Implement change management programs that communicate AI benefits in operational terms (predictive maintenance, reduced false alarms, improved asset visibility) rather than purely security-focused justifications.

Balance automation with human control. Implement graduated automation levels starting with advisory mode (Level 2) for 6-12 months, progressing to consent-required (Level 3) for low-risk actions, and reserving fully autonomous operation (Level 5) only for validated life-safety scenarios. Regular manual drills and simulations maintain operator skills and situational awareness, mitigating the automation paradox where human intervention capabilities atrophy precisely when most needed during system failures.

Model drift and concept drift are inevitable as infrastructure evolves and adversaries adapt; systems requiring manual retraining every few years will become obsolete. *Invest in continuous improvement.* Establish automated retraining pipelines, performance monitoring dashboards, and feedback loops incorporating operator input. Combine improvements in operational efficiency with best-in-class security updates: investments in operational efficiencies in critical infrastructure without addressing the corresponding security concerns will not result in achieving the projected benefits to the underlying business.

Organizations with limited budgets should consider *open-source ML libraries* (scikit-learn, TensorFlow) and algorithms (Isolation Forest) deployed on existing infrastructure before investing in commercial platforms. Prudent investment demonstrates that cost constraints need not preclude AI adoption. Organizations must, however, ensure they possess or can acquire necessary expertise to implement, maintain, and secure open-source solutions.

Endeavor to participate in information sharing initiatives. Join industry-specific Information Sharing and Analysis Centers (ISACs) and consortia enabling collaborative threat intelligence development. Federated learning approaches allow organizations to improve ML models collectively while preserving competitive and operational secrets. Threat detection achieved through multi-utility collaboration illustrates the power of collective defense against common adversaries.

Directions for Future Research

This analysis identifies seven critical areas warranting further investigation:

i. *Adversarial robustness in critical infrastructure contexts*: While general adversarial machine learning research has advanced significantly, specific studies examining nation-state capabilities to evade ICS-focused ML models remain limited. Research should investigate attack techniques specifically targeting OT protocols, SCADA communications, and industrial sensors, along with defenses optimized for resource-constrained OT environments. Collaborative efforts between academic researchers, infrastructure operators, and security agencies could leverage testbed environments to study adversarial attacks without risking operational systems.

ii. *Explainable AI techniques for operational technology*: Current XAI methods primarily target IT applications. Research is needed on explanation techniques tailored to OT contexts where operators require understanding of physical processes and operational states rather than abstract features. Investigation of causal inference methods that explain not just "what" the model detected but "why" the anomaly matters in terms of physical consequences would enhance operator decision-making and trust.

iii. *Privacy-preserving collaborative learning frameworks*: While federated learning shows promise, practical implementations face challenges including communication efficiency, handling non-IID (non-independent and identically distributed) data across organizations, and preventing malicious participants from poisoning global models. Research should develop robust aggregation mechanisms, secure multi-party computation protocols optimized for critical infrastructure constraints, and governance frameworks for multi-stakeholder AI consortia.

iv. *Human factors in AI-augmented operations*: Limited research examines how OT operators interact with AI recommendations in high-pressure, safety-critical situations. Studies should investigate cognitive load implications of AI alerts, trust calibration mechanisms preventing both over-reliance and under-reliance on AI, and training approaches maintaining human skills despite increasing automation. Longitudinal studies tracking operator performance and situational awareness over years of AI system deployment would provide valuable insights.

v. *Resilience and recovery from AI system failures*: Research should examine graceful degradation strategies when AI systems fail, mechanisms for detecting AI system compromise or malfunction, and procedures for transitioning between automated and manual control modes. Case studies documenting AI failures in critical infrastructure (when such information becomes declassified or publicly available) would provide crucial lessons for improving system designs.

vi. *Economic models for AI investment in critical infrastructure*: Generalized frameworks for evaluating AI investments across diverse infrastructure types, organizational scales, and risk profiles would assist decision-makers in resource allocation. Research should examine total cost of ownership including maintenance and retraining expenses, opportunity costs of delayed adoption, and methods for quantifying intangible benefits such as improved regulatory relationships and enhanced organizational capabilities. The analysis also should study the projected impacts of successful cyber-attacks on specific infrastructure components affecting individual verticals (i.e. electrical grids or SCADA controls of dams). This needs to be coupled with the generation of actuarial tables that describe the costs associated with successful cyber-attacks that can be used to recalibrate cost estimates.

vii. *Standardization and interoperability*: As AI becomes integrated into critical infrastructure at scale, standards for model formats, explanation interfaces, performance metrics, and security testing will become essential. Research should inform development of these standards, considering both technical feasibility and practical adoption challenges across fragmented infrastructure sectors and international jurisdictions.

Conclusion

The integration of artificial intelligence and machine learning into critical infrastructure protection represents a fundamental transformation in how society defends essential systems upon which modern life depends. This transformation brings immense promise: the ability to detect threats at machine speed, predict failures before they occur, and adapt to evolving attack landscapes. The quantified benefits realized provide compelling evidence that AI/ML can significantly enhance infrastructure security and resilience.

This article has also documented significant challenges, limitations, and risks that demand humility and careful governance. AI/ML systems remain fundamentally dependent on training data quality and representativeness, struggle with truly novel threats, can be manipulated by sophisticated adversaries, may perpetuate biases and create accountability gaps, and introduce new failure modes even as they address traditional ones. The path forward requires balancing innovation with safety, automation with human control, and efficiency with explainability. Governmental regulations that require the reporting and instantiation of data describing the incidents and associated costs of successful attacks is essential in this analysis.

Most critically, successful AI/ML deployment in critical infrastructure is not primarily a technical challenge but an organizational and governance challenge. Technical sophistication without organizational readiness, domain expertise, stakeholder trust, and ethical frameworks will not produce systems that enhance security and resilience. The absence of the estimated costs of integrating security and resilience into operational improvements will result in artificially low estimates of these types of upgrades The standardization of these estimates through organizations such as NIST will improve and streamline the cost/benefit analysis of improvements to critical infrastructure components.

As critical infrastructure operators, policymakers, researchers, and vendors collectively navigate this transformation, several principles should guide responsible AI deployment. First, AI should augment rather than replace human expertise, maintaining meaningful human control over consequential decisions while leveraging machine capabilities for tasks exceeding human capacity. Second, transparency and explainability should be non-negotiable requirements, enabling operators to understand, trust, validate, and improve AI systems over time. Third, governance frameworks must balance innovation with safety, establishing guardrails without stifling beneficial applications. Fourth, collaboration across organizations, sectors, and nations should enable collective defense against common adversaries while respecting competitive and security boundaries through privacy-preserving techniques.

Finally, the critical infrastructure community must maintain realistic expectations about what AI/ML can and cannot achieve. These technologies offer powerful tools for addressing unprecedented cybersecurity challenges, but they are tools requiring skilled human operators, robust governance structures, and continuous vigilance. The adversaries targeting critical infrastructure—nation-state actors with significant resources, criminal organizations with profit motives, and insider threats with system knowledge—will adapt to AI-enabled defenses just as defenders adapt to new attack techniques. The contest between attackers and defenders is fundamentally dynamic, requiring continuous evolution of defensive capabilities.

The future of critical infrastructure protection will undoubtedly involve increasingly sophisticated AI/ML systems. This article has sought to provide a comprehensive analysis of applications, trade-offs, and governance challenges to inform responsible deployment that enhances security and resilience while managing attendant risks. The path forward requires technical excellence, organizational maturity, ethical commitment, and collaborative engagement among all stakeholders. If pursued thoughtfully, AI-enabled critical infrastructure protection can significantly enhance societal resilience against evolving cyber threats while preserving the safety, accountability, and human control essential for systems upon which lives depend.

References

1. Ahmad, A., Desouza, K. C., Maynard, S. B., Naseer, H., & Baskerville, R. L. (2022). How integration of cyber security management and incident response enables organizational learning. *Journal of the Association for Information Science and Technology*, *73*(8), 1176-1192. https://doi.org/10.1002/asi.24620
2. Akoglu, L., Tong, H., & Koutra, D. (2010). Graph-based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery*, *29*(3), 626-688. https://doi.org/10.1007/s10618-014-0365-y
3. Alsamiri, J., & Alsubhi, K. (2019). Internet of Things cyber attacks detection using machine learning. *International Journal of Advanced Computer Science and Applications*, *10*(12), 627-634. https://doi.org/10.14569/IJACSA.2019.0101281
4. Anwar, R. W., Abdullah, T., & Pastore, F. (2021). Firewall best practices for securing smart healthcare environment: A review. *Applied Sciences*, *11*(19), 9183. https://doi.org/10.3390/app11199183
5. Assante, M. J., & Lee, R. M. (2015). *The industrial control systems cyber kill chain*. SANS Institute. https://icscsi.org/library/Documents/White_Papers/SANS%20-%20ICS%20Cyber%20Kill%20Chain.pdf
6. Basiru, M. O., Zubair, A., & Williams, C. (2023). Artificial intelligence for cybersecurity: Offensive tactics, mitigation techniques and future directions. *Applied Artificial Intelligence*, *37*(1), 2201883.
7. Basiru, M. O., Zubair, A., & Williams, C. (2023). Artificial intelligence for cybersecurity: Offensive tactics, mitigation techniques and future directions. *Applied Artificial Intelligence*, *37*(1), 2201883. https://academic-journals.eu/pl/download?path=%2Fuploads%2FZm9sZGVycHVibWVkaWE1Ng%3D%3D%2Fdocuments%2Facig_erwin_adi_final2.pdf
8. Bruce, V. (2025). AI As a Double-Edged Sword for OT/ICS Cybersecurity. *Solutions Review*. Rockwell Automation. https://solutionsreview.com/endpoint-security/ai-as-a-double-edged-sword-for-ot-ics-cybersecurity/#:~:text=As%20OT/IT%20convergence%20continues,of%20AI%20strategy%20that%20wins
9. Burgess, C. (2024). Legacy systems are the Achilles' heel of critical infrastructure cybersecurity. *CSO Online*. https://www.csoonline.com/article/2514214/legacy-systems-are-the-achilles-heel-of-critical-infrastructure-cybersecurity.html#:~:text=The%20importance%20of%20critical%20infrastructure%20cannot%20be,%E2%80%94%20something%20we%20literally%20cannot%20live%20without
10. Casalicchio, E., Galli, E., & Tucci, S. (2010). Agent-based modelling of interdependent critical infrastructures. *Int. J. Syst. Syst. Eng.*, *2*, 60-75. https://art.torvergata.it/bitstream/2108/41437/1/IJSSE%202%281%29%20Casalicchio%20et%20al.pdf
11. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7
12. Chithaluru, P., Al-Turjman, F., Kumar, M., & Stephan, T. (2023). I-AREOR: An energy-balanced clustering protocol for implementing green IoT in smart cities. *Sustainable Cities and Society*, *90*, 104366. https://doi.org/10.1016/j.scs.2022.104366
13. Chowdhury, R. (2024). AI-driven business analytics for operational efficiency. *World Journal of Advanced Engineering Technology and Sciences.* DOI:10.30574/wjaets.2024.12.2.0329
14. Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Najada, H. A. (2023). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, *10*(1), 1-54. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=qCM4FZwAAAAJ&citation_for_view=qCM4FZwAAAAJ:UeHWp8X0CEIC
15. Device Authority. (2024). Critical Infrastructure Under Siege: How Automation Can Safeguard Against Cyber Threats.

https://deviceauthority.com/critical-infrastructure-under-siege-how-automation-can-safeguard-against-cyber-threats/

16. Elmaghraby, A.E. and Losavio, M. (2014). Cyber Security Challenges in Smart Cities: Safety, security and privacy. Journal of Advanced Research, 5(4). DOI:10.1016/j.jare.2014.02.006
17. Ferrag, M. A., Debbah, M., & Choo, K. K. R. (2024). Artificial intelligence for cyber-physical systems security: A survey. *IEEE Communications Surveys & Tutorials*, *26*(1), 5-39. https://doi.org/10.1109/COMST.2023.3321551
18. Forvis Mazars. (2025). Addressing Rising Cyberthreats on US Critical Infrastructure. https://www.forvismazars.us/forsights/2025/09/addressing-rising-cyberthreats-on-us-critical-infrastructure
19. Gordon, J. (2020). Critical Infrastructure Protection- the Essential Guide. *Industrial*. https://industrialcyber.co/features/critical-infrastructure-protection-a-beginners-guide/
20. Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The emerging threat of ai-driven cyber attacks: A review. *Applied Artificial Intelligence*, *36*(1), 2037254. https://doi.org/10.1080/08839514.2022.2037254
21. Gugueoth, V., Safavat, S., Shetty, D. K., & Kotecha, K. (2023). Machine learning-based security framework for IoT-enabled industrial control systems. *IEEE Access*, *11*, 28235-28253. https://doi.org/10.1109/ACCESS.2023.3259847
22. Gujar, S.S. (2024). "Real-Time Threat Detection and Response Using AI for Securing Critical Infrastructure," *2024 Global Conference on Communications and Information Technologies (GCCIT)*, BANGALORE, India, 2024, pp. 1-7, doi: 10.1109/GCCIT63234.2024.10862978. https://ieeexplore.ieee.org/document/10862978
23. Hildick-Smith, A. (2022). Security for Critical Infrastructure SCADA Systems. *SANS Institute*. https://www.sans.org/white-papers/1644
24. Huda, S., Abawajy, J., Alazab, M., Abdollalihian, M., Islam, R., & Yearwood, J. (2018). Hybrids of support vector machine wrapper and filter based framework for malware detection. *Future Generation Computer Systems*, *78*, 491-502. https://doi.org/10.1016/j.future.2017.07.035
25. Illumio. (2025). Cybersecurity 101: What is Lateral Movement? https://www.illumio.com/cybersecurity-101/lateral-movement
26. Inductive Automation. (2018). What is SCADA? Supervisory Control and Data Acquisition. https://inductiveautomation.com/resources/article/what-is-scada
27. Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., & Sun, J. (2017). Anomaly detection for a water treatment system using unsupervised machine learning. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 1058-1065). IEEE. https://doi.org/10.1109/ICDMW.2017.149
28. Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, 101804. https://doi.org/10.1016/j.inffus.2023.101804
29. Kim, J., Shin, N., Jo, S. Y., & Kim, S. H. (2019). Method of intrusion detection using deep neural network. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 313-316). IEEE. https://doi.org/10.1109/BIGCOMP.2019.8679508
30. Lang, X., Nilsson H., and Mao, W. *(*2024). *IOP Conf. Ser.: Earth Environ. Sci.* **1411** 012046 https://iopscience.iop.org/article/10.1088/1755-1315/1411/1/012046
31. Langner, R. (2011). Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy*, *9*(3), 49-51. https://doi.org/10.1109/MSP.2011.67
32. Lee, R., Assante, M., and Conway, T. (2016). Analysis of the cyber attack on the Ukrainian power grid. *Electricity Information Sharing and Analysis Center (E-ISAC)*, 388. https://nsarchive.gwu.edu/sites/default/files/documents/3891751/SANS-and-Electricity-Information-Sharing-and.pdf
33. Linkov, I. and Kott, A. (2019). Fundamental Concepts of Cyber Resilience: Introduction and Overview. *Cyber Resilience of Systems and Networks* (pp.1-25). DOI:10.1007/978-3-319-77492-3_1
34. Macas, M., Wu, C., & Fuertes, W. (2022). A survey on deep learning for cybersecurity: Progress, challenges, and opportunities. *Computer Networks*, *212*, 109032. https://doi.org/10.1016/j.comnet.2022.109032
35. Maglaras, L., Drivas, G., Nogueira, K., Janicke, H., & Ferrag, M. A. (2023). Cybersecurity in the era of artificial intelligence: A systematic literature review. *ACM Computing Surveys*, *56*(5), 1-36. https://doi.org/10.1145/3571156
36. Marley, M. (2025). How to Prevent Lateral Movement: Cybersecurity Risks and Strategies. *Zero Networks*. https://zeronetworks.com/blog/how-to-prevent-lateral-movement-cybersecurity-risks-strategies
37. Mishra, P. (2025). History of ICS & SCADA Systems. *Study.com*. https://study.com/academy/lesson/history-of-ics-scada-systems.html
38. Mishra, S., Sharma, S. K., & Alowaidi, M. A. (2022). Analysis of security issues in cloud environment. *Computers & Security*, *112*, 102508. https://doi.org/10.1016/j.cose.2022.102508
39. Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. *Knowl Inf Syst* **67**, 6969–7055. https://doi.org/10.1007/s10115-025-02429-y
40. Mohammed, M. (2025). "Emerging artificial intelligence methods in civil engineering: A Comprehensive Review", *Rafidain J. Eng. Sci.*, vol. 3, no. 1, pp. 280–293, Feb. 2025, doi: 10.61268/939e6941. https://rjes.iq/index.php/rjes/article/view/155
41. Musa, U. S., Chizari, H., & Adetoye, A. O. (2024). Vulnerability management in operational technology environments: A systematic review. *Computers & Security*, *138*, 103655. https://doi.org/10.1016/j.cose.2023.103655
42. Ni, M. (2023). A review on machine learning methods for intrusion detection system. Proceedings of the 2023 International Conference on Software Engineering and Machine Learning. *Applied Computational Engineering, 27(1):57-64*. DOI:10.54254/2755-2721/27/20230148
43. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253. https://doi.org/10.1518/001872097778543886
44. Palo Alto Networks. (2025). What Is Network Segmentation? https://www.paloaltonetworks.com/cyberpedia/what-is-network-segmentation
45. Palo Alto Networks. (2025). What Is a Perimeter Firewall? https://www.paloaltonetworks.com/cyberpedia/what-is-a-perimeter-firewall

46. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, *30*(3), 286-297. https://doi.org/10.1109/3468.844354
47. Presidential Decision Directive 63. (1998). Critical Infrastructure Protection. *The White House*. https://irp.fas.org/offdocs/pdd/pdd-63.htm
48. Raman, R., Achuthan, K., Vinod Kumar, K., Venkataraghavan, S. S., & Nedungadi, P. (2024). Artificial intelligence in cyber security: Research advances, challenges, and opportunities. *Artificial Intelligence Review*, *57*(1), 1-59. https://doi.org/10.1007/s10462-023-10588-5
49. Rockwell Automation. (2022). Critical Infrastructure Cybersecurity Fundamentals. https://www.rockwellautomation.com/en-us/company/news/blogs/cyber-fundamentals.html
50. Rockwell Automation. (2025). Critical Infrastructure Cybersecurity Solutions. Retrieved from https://www.rockwellautomation.com/en-us/capabilities/industrial-cybersecurity/industry-services/critical-infrastructure.html
51. Sarker, I. H. (2023). Machine learning for intelligent data analysis and automation in cybersecurity: Current and future prospects. *Annals of Data Science*, *10*(6), 1473-1498. https://doi.org/10.1007/s40745-022-00444-2
52. Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2024). Ai-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science*, *5*(4), 1-22. https://doi.org/10.1007/s42979-024-02700-w
53. SCADA Info. (2023). History of SCADA. https://www.scadainfo.com/history-of-scada/
54. Sharma, R., Pandey, R., & Nigam, A. (2021). Real-Time Object Detection for Visually Challenged. In *Machine Learning and Information Processing* (pp. 579–589). Springer Singapore. https://doi.org/10.1007/978-981-33-4859-2_28
55. Song, L., & Kawai, K. (2023). Survival analysis for predictive maintenance of infrastructure systems. *Reliability Engineering & System Safety*, *231*, 109023. https://doi.org/10.1016/j.ress.2022.109023
56. Stevens, T. (2020). *Knowledge in the grey zone: AI and cybersecurity*. Digital Society Collaboratory, Berlin Social Science Center. https://doi.org/10.1057/s42984-020-00007-w.
57. Thawait, N. K. (2024). "Machine Learning in Cybersecurity: Applications, Challenges and Future Directions, " *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, vol. 10, no. 3, pp. 16–27, doi: 10.32628/CSEIT24102125. https://www.researchgate.net/publication/380327525_Machine_Learning_in_Cybersecurity_Applications_Challenges_and_Future_Directions
58. Unmudl. (2025). Understanding SCADA Systems: An In-Depth Guide. https://www.unmudl.com/blog/scada-systems
59. Whatney, M. (2022). Cybersecurity Threats to and Cyberattacks on Critical Infrastructure: a Legal Perspective. European Conference on Cyberwarfare and Security, 21(1):319-327 DOI:10.34190/eccws.21.1.196
60. Xiang, H., Li, X., Liao, X., Cui, W., Liu, F., and Li, D. (2025). Artificial Intelligence in Renewable Energy Systems: Applications and Security Challenges. *MDPI, Energy*. 18(8):1931 DOI:10.3390/en18081931
61. Yigit, Y., Ferrag, M. A., Ghanem, M. C., Sarker, I. H., Maglaras, L. A., Chrysoulas, C., Moradpoor, N., Tihanyi, N., & Janicke, H. (2025). Generative AI and LLMs for Critical Infrastructure Protection: Evaluation Benchmarks, Agentic AI, Challenges, and Opportunities. *Sensors*, *25*(6), 1666. https://doi.org/10.3390/s25061666. https://www.mdpi.com/1424-8220/25/6/1666
62. Zhang, C., Chen, Y., Meng, Q., & Zhang, R. (2022). A deep learning approach for network intrusion detection based on NSL-KDD dataset. In *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)* (pp. 1003-1008). IEEE. https://doi.org/10.1109/ICPECA53709.2022.9718847
63. Zhou, X., Feng, Y.-J., & Zhou, X. (2017). Real-Time Object Detection Using Efficient Convolutional Networks. In *Biometric Recognition* (pp. 569–576). Springer International Publishing. https://doi.org/10.1007/978-3-319-69923-3_68