algorithms

MDPI

Article

Artificial Intelligence for Cybersecurity: A Scoping Survey of Paradigms, Applications, and Emerging Trends

Amitabh Mishra *, Vasudha Vedula, Asmi Mishra and Shrishti Sharma

Department of Cybersecurity and Information Technology, University of West Florida (UWF), Pensacola, FL 32514, USA; vvedula@uwf.edu (V.V.); am501@students.uwf.edu (A.M.); fs56@students.uwf.edu (S.S.)

* Correspondence: amishra@uwf.edu

Abstract

The rapid evolution of cyberattacks, coupled with the increasing capacity of computing environments and the emergence of artificial intelligence (AI), has significantly complicated the security landscape. While existing studies largely emphasize improving AI model performance for individual cybersecurity tasks, this survey shifts the focus toward operationalizing the deployment rationale and understanding when, where, and why different AI paradigms should be deployed, the capabilities they offer; and the challenges that must be addressed to enable trustworthy and effective real-world cyber defense. This paper aims to provide researchers and practitioners with a comprehensive reference for understanding the evolving role of AI in cybersecurity and the challenges that must be addressed to develop trustworthy and resilient AI-driven cyber defense systems. In this paper, we propose a structured taxonomy to organize various dimensions of AI-driven cybersecurity; review them critically; and finally, discuss key challenges, open problems, and emerging trends.

Keywords: artificial intelligence (AI); internet of things (IoT); cyber-physical systems (CPS); generative AI (GenAI); large language models (LLMs); machine learning (ML); deep learning (DL); reinforcement learning (RL); explainable AI (XAI); federated learning (FL); quantum machine learning (QML)

1. Introduction

The rapid advancement in computer technologies has led to a dramatic transformation in organizational networks and critical infrastructure systems, resulting in a significantly expanded attack surface. While safeguarding information has been a central concern of computer science for decades, the complexity of modern cyberspace has reached an unprecedented scale. Today, attackers no longer rely solely on identifiable, static signatures; instead, they employ polymorphic, adaptive, and multi-stage attacks that actively evade traditional security mechanisms. The evolving cyberspace and security challenges are depicted in Figure 1. Rule-based and signature-based detection systems, which depend on prior knowledge of attack characteristics, increasingly struggle to detect previously unseen or zero-day threats. This shift has motivated extensive research into artificial intelligence (AI) techniques for cybersecurity for more than a decade. AI-driven approaches have demonstrated high potential in areas such as anomaly detection, intrusion detection, malware classification, and automated incident response (IR) by learning complex patterns

Check for updates

Academic Editor: Paul A. Gagniuc

Received: 10 June 2026
Revised: 30 July 2026
Accepted: 3 August 2026
Published: 7 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

from massive volumes of unstructured data. Additionally, the emergence of transformer-based models such as large language models (LLMs) has caused another significant shift in AI applications with respect to security problems and attack vectors. According to a 2025 European survey by ISACA, digital trust professionals increasingly identify AI-assisted attacks as among the fastest-growing categories of cyber risk [1], reflecting a broader shift in organizational threat perception toward AI-enabled adversaries.

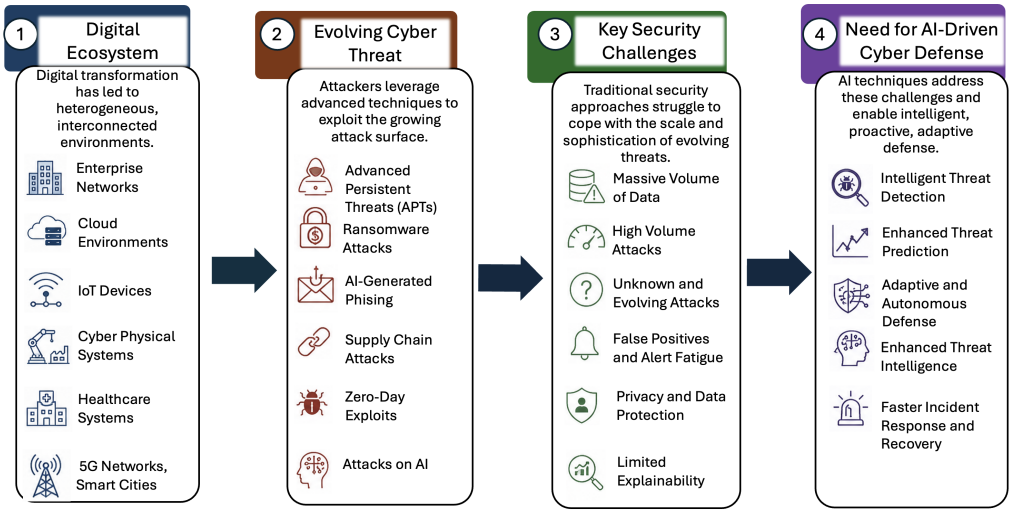

Figure 1. Evolution of the modern cyber threat landscape driving AI adoption in cybersecurity-The rapid growth of interconnected digital infrastructures and the increasing sophistication of threats have outpaced traditional rule-based security mechanisms thus creating need for intelligent, adaptive, and proactive AI-driven cyber defense. This progression motivates the five-dimensional taxonomy introduced in Figure 2, which structures the remainder of the survey.

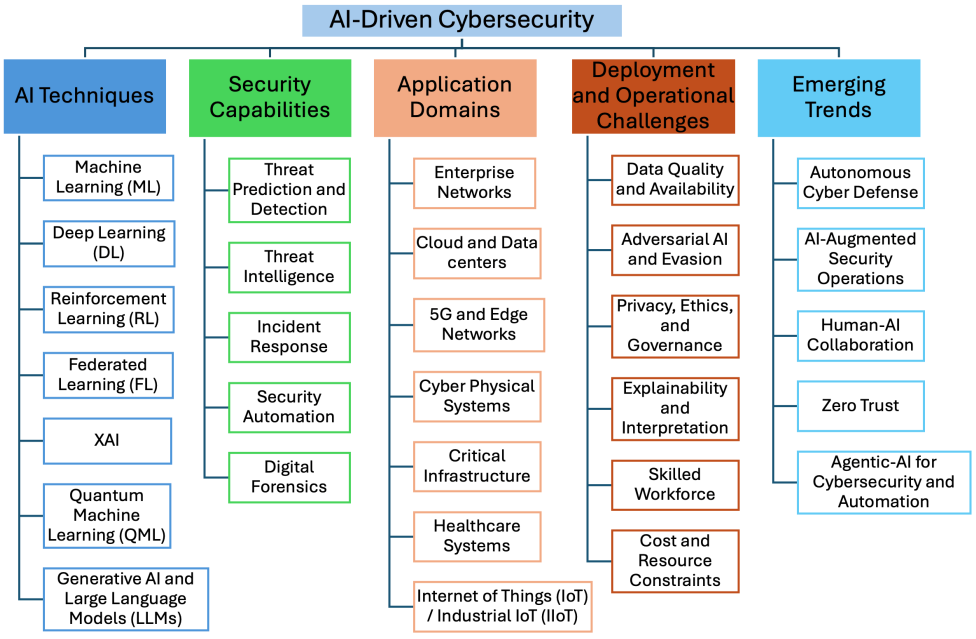

Figure 2. Proposed Taxonomy of AI-Driven Cybersecurity-This taxonomy organizes the AI-driven cybersecurity landscape across five complementary dimensions and provides the conceptual structure for this survey.

While AI-driven cybersecurity is a highly researched field that is served by several comprehensive surveys, there are still critical gaps in the existing literature. The existing reviews provide massive bibliometric data or detailed classifier analysis; however, they

often lack the operational guidance required for real-world deployment. As shown in Table 1, our work is distinguished from prior surveys by combining a formal taxonomy with a prescriptive deployment rationale that integrates emerging paradigms like FL and LLMs. To our knowledge, this emphasis is not the primary focus of existing comprehensive reviews in this space.

Table 1. Structured comparison of this paper with contemporary AI cybersecurity surveys.

Survey Ref	Year	Primary Focus	Key Contribution	Deployment Framework
Ofusori et al. [2]	2024	General AI Review	Bibliometric analysis of 939 papers.	No
Ali et al. [3]	2019	General AI/ML Review	Systematic review of ML algorithms and model performance benchmarking.	No
Mohamed [4]	2023	Trends in AI/ML	State-of-the-art survey of IEEE/ACM/Springer articles.	No
Randieri et al. [5]	2025	Innovations Scoping	Future perspectives and scoping review.	No
Ferrag et al. [6]	2024	FL, DL, and Blockchain	Alignment and hardware design security.	No
Our Paper	2026	Cyber Defense	Formal taxonomy and 'When, Where, Why' matrix.	Yes

This survey shifts the focus from "which model performs best" toward understanding when, where, and why different AI paradigms (e.g., centralized DL vs. FL vs. LLMs) should be deployed. We aim to provide an operational perspective that balances detection accuracy with practical constraints such as data privacy, computational overhead, and explainability. Throughout this paper, AI is used as the umbrella term for the broader ecosystem of intelligent techniques discussed here, while ML, DL, FL, RL, XAI, LLMs, GenAI, and QML are each treated as distinct subfields or paradigms within that ecosystem, with their individual architectures, capabilities, and deployment characteristics detailed in Section 2. The primary contributions of this survey are summarized as follows:

- We propose a structured taxonomy of AI-driven cybersecurity organized across five core dimensions: technical paradigms, defensive capabilities, application domains, operational challenges, and emerging trends as shown in Figure 2.
- We operationalize the "when, where, and why" promise by providing a concrete decision matrix and deployment criteria for selecting AI models based on the threat model, data locality, and resource constraints as shown in Table 2.
- We provide a comprehensive synthesis of contemporary AI paradigms—including ML, DL, reinforcement learning (RL), FL, LLMs, and emerging quantum machine learning (QML). Further, critically analyze their capabilities, strengths, limitations, and suitability for cybersecurity tasks.
- Finally, we examine the adoption of AI across diverse cybersecurity domains, including enterprise networks, cloud computing, IoT, CPS, healthcare, and defense, highlighting common security objectives, domain-specific challenges, and representative real-world use cases.

Table 2. Deployment Decision Matrix for AI Paradigms in Cyber Defense.

Paradigm	When to Deploy	Where to Place	Why (Rationale)
FL	High privacy, decentralized data.	Edge Gateways, Local LANs	Preserves data locality; prevents raw log leakage
DL	High-volume, complex traffic patterns.	Network Perimeter, Cloud	Excels at feature extraction from raw flows
LLMs	Semantic analysis, CTI synthesis.	SOC, Management Layer	Processes unstructured data; automates reports
Classical ML	Resource-constrained, simple anomalies.	IoT End-devices, Sensors	Low compute overhead; high interpretability

This survey adopts a narrative, thematically organized approach to literature synthesis, with a focus on conceptual synthesis and operational guidance. Sources were selected for relevance, recency, and coverage across the five taxonomy dimensions, rather than through a PRISMA-compliant formal systematic protocol; accordingly, the review does not claim exhaustive or bias-controlled coverage of the field, consistent with its framing as a comprehensive survey rather than a systematic review.

The paper is organized as follows. Section 2 presents the background of AI techniques and paradigms for cybersecurity. Section 3 presents the security capabilities and key application domains. Section 4 outlines challenges, emerging trends, and open problems, and Section 5 concludes the paper.

2. AI Techniques and Paradigms for Cybersecurity

AI has evolved into a diverse ecosystem of learning paradigms, each offering distinct strengths, limitations, and deployment characteristics for cybersecurity. Often modern security solutions employ different paradigms depending on the nature of the security problem, availability of data and computational resources, operational objectives, and privacy requirements. This section provides an overview of the major AI techniques and paradigms applied specifically in cybersecurity, ranging from classical ML models, also known as shallow learning models, to advanced DL models, federated learning frameworks, and advanced transformer-based models such as large language models (LLMs). It further discusses their strengths, comparative characteristics, practical trade-offs, and suitability for different cybersecurity scenarios based on the summary provided in Figure 3.

AI Techniques/ Capabilities	Generalization (Transfer Knowledge to Unseen Attacks)	Scalability (Ability to handle large amounts of data)	Adaptability (Dynamic to change)	Explainability (Interpretable model predictions)	Robustness (Resiliency to noise and poisoning attacks)	Computational Efficiency	Data Efficiency (Ability to learn from limited data)	Privacy Preservation (Protect sensitive information)
Machine Learning (ML)	◐	◐	◐	◐	◐	◐	◐	◐
Deep Learning (DL)	●	●	●	◐	◐	◐	◐	○
Reinforcement Learning (RL)	○	◐	●	○	◐	◐	○	○
Federated Learning (FL)	◐	●	◐	○	◐	◐	◐	●
Transformer-Based Foundation Models	●	◐	◐	◐	◐	◐	◐	◐
Quantum Machine Learning (QML)	△	△	△	△	△	△	△	△

● Strong Capability ◐ Moderate Capability △ Emerging / Potential ○ No Capability

Figure 3. Multi-Dimensional Capability Matrix of AI Techniques-The matrix provides a comprehensive view of how different AI techniques and paradigms contribute to a wide range of security capabilities across diverse application domains. Different techniques offer complementary strengths and gaps. The matrix was constructed by synthesizing reported capability claims and limitations from the reviewed literature and should be interpreted as a qualitative guide rather than a quantitative benchmark.

2.1. Machine Learning (ML) Models

Classical machine learning algorithms have formed the foundation of AI-driven cybersecurity for over a decade, offering interpretable and computationally efficient approaches to threat detection. The most widely deployed ML algorithms in cybersecurity include

Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), K-Nearest Neighbors (KNN), and Naive Bayes classifiers [7,8].

2.1.1. Learning Approaches

Supervised ML methods dominate the cybersecurity literature, particularly for classification tasks such as intrusion detection and malware identification. SVMs have been extensively applied due to their effectiveness in high-dimensional feature spaces and ability to handle non-linear decision boundaries through kernel methods [7,8]. RF classifiers are favored for their ensemble nature, which provides robustness against overfitting and inherent feature importance ranking capabilities. Decision Trees offer interpretability advantages, making them suitable for security operations centers (SOCs) where explainability is critical.

Unsupervised learning techniques, including K-Means clustering and Gaussian Mixture Models, are employed for anomaly detection scenarios where labeled attack data is scarce or unavailable [8]. These methods identify deviations from normal behavior patterns without requiring explicit attack signatures, making them particularly valuable for detecting novel or zero-day threats. However, unsupervised approaches often suffer from high false positive rates, limiting their deployment in production environments without human oversight [9].

2.1.2. Performance Characteristics

ML algorithms typically achieve accuracy rates between 85–95% on standard benchmark datasets [10–13], with Random Forest and SVM consistently ranking among the top performers [7,9]. The computational efficiency of ML methods enables real-time processing of network traffic, with inference times often measured in milliseconds (ms). However, ML approaches face limitations in handling complex, high-dimensional data such as raw network packets or binary executables, where feature engineering becomes a critical bottleneck [8].

2.1.3. Comparative Advantages

The primary advantages of classical ML include interpretability, lower computational requirements compared to deep learning, and effectiveness with smaller training datasets. These characteristics make ML particularly suitable for resource-constrained environments such as IoT devices and embedded systems, as well as scenarios where regulatory compliance demands explainable decision-making [14].

2.2. Deep Learning (DL) Models

Deep learning has emerged as a transformative paradigm in cybersecurity, offering the ability to automatically learn hierarchical feature representations from raw data without extensive manual feature engineering. The most prominent DL architectures deployed in cybersecurity include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Deep Belief Networks (DBNs), and Autoencoders [8,15–20].

2.2.1. Prominent Architectures

Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in malware detection and network intrusion detection by treating binary executables or network traffic as image-like data structures [8,17–21]. The convolutional layers automatically extract spatial hierarchies of features, enabling detection of complex attack patterns that evade traditional signature-based methods. Recent work has shown CNNs achieving accuracy rates exceeding 99% on malware classification tasks when trained on sufficient

data [8,19]. For network intrusion detection, CNN-based systems have been deployed for LAN security, demonstrating real-time detection capabilities [18].

Recurrent neural networks (RNNs) and their Long Short-Term Memory (LSTM) variants excel at modeling sequential dependencies in time-series data, making them particularly effective for analyzing network traffic flows and detecting multi-stage attacks [8,16,17,22–24]. LSTM networks address the vanishing gradient problem inherent in standard RNNs, enabling learning of long-term dependencies critical for identifying sophisticated attack campaigns. Hybrid CNN-LSTM architectures have emerged as a powerful approach, combining CNN's spatial feature extraction with LSTM's temporal modeling capabilities [17]. These hybrid models have achieved detection accuracies of 99.6% on benchmark datasets while maintaining low false positive rates [17].

Deep Belief Networks (DBNs) and Autoencoders, constructed from stacked Restricted Boltzmann Machines (RBMs), provide unsupervised pre-training capabilities that improve performance when labeled data is limited [8]. Autoencoders, particularly denoising autoencoders, are employed for anomaly detection by learning compressed representations of normal behavior and flagging inputs that cannot be accurately reconstructed [8,25,26]. These approaches are valuable for detecting novel attacks that differ significantly from training data distributions.

2.2.2. Performance and Scalability

Deep learning models consistently achieve state-of-the-art performance on cybersecurity benchmark datasets, with reported accuracies frequently exceeding 95–99% [8,16,17,19]. However, these performance gains come at the cost of increased computational requirements, longer training times, and reduced interpretability [8]. The "black box" nature of deep neural networks poses challenges for security analysts seeking to understand why specific threats were flagged, limiting trust and adoption in critical infrastructure contexts [14].

2.2.3. Emerging Architectures

Recent work has explored Transformer architectures and attention mechanisms for cybersecurity applications, leveraging their success in natural language processing. These models show promise for analyzing log files, security alerts, and threat intelligence reports, though their application in cybersecurity remains less mature than in other domains [27].

2.3. Reinforcement Learning (RL)

Reinforcement learning (RL)-based security frameworks are designed with a primary focus of adaptiveness and automation where agents learn optimal defense strategies through interaction with the environment, receiving rewards based on detection accuracy, response effectiveness, or resource utilization. This approach is particularly suitable for dynamic decision-making tasks such as intrusion response, attack mitigation, and security policy optimization. RL techniques have been explored in software-defined networking (SDN) environments [16] for adaptive traffic control, DDoS mitigation, and moving target defense. By continuously learning from observed attack behavior, RL agents can adjust defense mechanisms in real time. However, RL models face challenges related to training stability, convergence, and safety, especially when deployed in live production environments where exploration may introduce risk.

2.4. Federated Learning (FL)

Federated learning (FL), a paradigm shift in AI model training, typically addresses critical privacy and scalability concerns in cybersecurity. In FL, models are trained collaboratively across multiple organizations or devices without sharing raw data, thereby

preserving privacy and reducing regulatory risk. This paradigm is particularly relevant for Internet of Things (IoT), Industrial Internet of Things (IIoT), and cross-enterprise threat intelligence sharing.

FL enables security models to operate across edge, fog, and cloud layers, allowing local detection while benefiting from global knowledge aggregation. Despite these advantages, FL introduces new challenges, including communication overhead, heterogeneous data distributions, and vulnerability to poisoning and inference attacks.

2.5. Explainable AI (XAI)

XAI techniques aim to make model decisions transparent and understandable to human analysts, thus improving trust, usability, and compliance with regulatory requirements. XAI techniques aim to explain why a decision was made, rather than just reporting what decision was made. This distinction is especially important in high-stakes domains such as cybersecurity, healthcare, and defense, where automated actions can have severe operational consequences, including system downtime or service disruption. A lack of transparency in AI-driven systems creates problems for security operations. When decisions cannot be clearly explained, analysts may struggle to validate alerts, diagnose false positives, or justify automated response actions. This opacity increases operational risk and undermines trust in automated systems. Common approaches include feature attribution methods, rule extraction, and surrogate models [28].

2.6. Generative AI (GenAI) and Large Language Models (LLMs)

The emergence of GenAI and LLMs represents a paradigm shift in cybersecurity, introducing both powerful defensive capabilities and novel attack vectors. This dual nature, where the same technologies can be weaponized by adversaries or deployed for defense, characterizes the current state of generative AI in cybersecurity [29,30].

2.6.1. Defensive Applications

LLMs have been applied to multiple cybersecurity tasks, including automated threat intelligence analysis, security code review, vulnerability detection, incident response automation, and security documentation generation [30]. The natural language understanding capabilities of models like GPT-4 and specialized security-focused LLMs enable parsing of unstructured threat reports, correlation of indicators of compromise (IoCs), and generation of actionable security recommendations [29]. Generative AI systems can assist security analysts by automating routine tasks, synthesizing threat intelligence from multiple sources, and providing contextual explanations of security events [29,31].

2.6.2. Generative Adversarial Networks (GANs)

GANs have been employed for data augmentation in cybersecurity, generating synthetic attack samples to address class imbalance problems in training datasets [8,32]. GAN-based intrusion detection systems for CAN bus networks have demonstrated the ability to detect 1954 malicious messages in 0.18 s, offering advantages in scalability and effectiveness [32]. However, GANs also pose security risks, as they can be used to generate adversarial examples that evade detection systems or create sophisticated phishing content.

2.6.3. Security Challenges of LLMs

The deployment of LLMs in cybersecurity introduces significant security and privacy concerns. LLMs are vulnerable to prompt injection attacks, jailbreaking attempts, data poisoning, and model extraction attacks [33–35]. Adversaries can exploit LLMs to generate polymorphic malware, craft convincing phishing emails, automate vulnerability discovery, and scale social engineering attacks [29]. The survey by Xu et al. provides a

systematic analysis of LLM applications in cybersecurity, highlighting both opportunities and risks [30].

2.6.4. Trustworthiness and Governance

Critical challenges remain in ensuring the trustworthiness of LLM-based security systems, including concerns about hallucinations (generation of false information), bias in security recommendations, lack of explainability, and potential for misuse [33]. Governance frameworks for responsible deployment of generative AI in security contexts are still emerging, with ongoing debates about transparency, accountability, and ethical considerations.

2.7. Quantum Machine Learning

Quantum machine learning (QML) represents an emerging frontier in cybersecurity, leveraging quantum computing principles to potentially overcome computational limitations of classical approaches. While still largely in the research phase, QML shows theoretical promise for enhancing threat detection capabilities [36–40].

2.7.1. Quantum Neural Networks

Quantum neural networks (QNNs) exploit quantum superposition and entanglement to process information in fundamentally different ways than classical neural networks. Recent work has demonstrated QNN-based network anomaly detection systems deployed on noisy intermediate-scale quantum (NISQ) computers, though performance remains limited by current hardware constraints [39]. Quantum deep learning approaches for anomaly detection have been proposed to enhance network security, with theoretical analyses suggesting potential advantages in pattern recognition tasks [41].

2.7.2. Quantum-Enhanced Intrusion Detection

Several studies have explored quantum machine learning techniques for intrusion detection systems (IDS). Kalinin and Krundyshev investigated quantum ML techniques for security intrusion detection, demonstrating feasibility on quantum simulators [37]. The QML-IDS system proposed by Abreu et al. represents one of the first practical implementations of quantum machine learning for intrusion detection, evaluated on standard cybersecurity datasets [38].

2.7.3. Quantum Malware Detection

QML has been applied to mobile malware detection and classification, with a focus on explainability which is a critical requirement for security applications [40]. The quantum approach aims to provide both improved detection accuracy and interpretable decision-making processes, addressing a key limitation of classical deep learning methods.

2.7.4. Current Limitations and Future Potential

Despite theoretical promise, QML in cybersecurity faces significant practical challenges. Current quantum hardware is noisy, error-prone, and limited in qubit count, restricting the size and complexity of problems that can be addressed [41]. Most QML cybersecurity research relies on quantum simulators rather than actual quantum computers, and the quantum advantage over classical methods remains unproven for most security tasks [37–39]. However, as quantum computing technology matures, QML may offer breakthrough capabilities for cryptanalysis, optimization of security configurations, and detection of sophisticated attacks that overwhelm classical computational resources.

3. Security Capabilities and Applications

The integration of AI into cybersecurity has enabled a wide range of applications including enterprise networks, cloud infrastructures, IoT ecosystems, healthcare systems and many more. This section focuses on security capabilities and their application in various domains. The applicability is categorized based on different security tasks that can be achieved by the capabilities of AI-driven security framework depicted in Figure 2. The choice of a specific AI technique depends on the task, characteristics of the data the technique will be trained on, and the operational and deployment requirements. Figure 3 compares key compatibilities dimensions and Table 2 details the deployment decision matrix for major AI paradigms in cybersecurity.

3.1. Security Capability: Threat Prediction and Anomaly Detection

Threat prediction and anomaly detection is one of the most extensively researched security capabilities in the AI-driven cybersecurity literature, and it occupies a foundational position in the Figure 2 taxonomy. AI models such as ML and DL learn complex, nonlinear relationships within high-dimensional data, enabling identification of subtle deviations from learned behavioral baselines that conventional signature-based systems cannot detect. By analyzing network traffic, system logs, user behavior, and application activity, AI-driven detection mechanisms can uncover malicious behavior in real time and generalize to previously unseen attack variants, including zero-day threats and polymorphic malware [42,43].

Classical ML algorithms such as Support Vector Machines (SVMs), Random Forests (RF), Decision Trees (DT), and k-Nearest Neighbors (kNNs) have long dominated this space due to their interpretability and computational efficiency [7,8]. DL architectures such as CNNs, RNNs, LSTM networks, and autoencoders have subsequently demonstrated superior performance on benchmark datasets, achieving accuracy rates frequently exceeding 99% on tasks such as malware classification and network intrusion detection [8,17,19]. However, as Figure 3 illustrates, this performance gain comes at the cost of reduced explainability and increased sensitivity to adversarial manipulation, a trade-off that is particularly consequential in high-stakes deployments.

A persistent structural limitation across this body of work is dataset quality. Widely used benchmarks like NSL-KDD, CICIDS2017, UNSW-NB15 are collected in controlled laboratory environments and tend to over-represent known attack types while underrepresenting benign traffic diversity [10–13,44]. Sommer and Paxson [45] critically observe that this contradicts the fundamental logic of anomaly detection, which requires models trained primarily on normal behavior to flag deviations. The prevalence of class-imbalanced, replay-based datasets introduces optimistic performance estimates that may not transfer to real production environments. Model calibration and contextual awareness therefore remain unresolved challenges across the threat-detection literature. The subsections below apply this critical lens to each major application domain within this security capability.

3.1.1. Application Domain: Network and Critical Infrastructure Security

Critical networked infrastructure spanning energy grids, water systems, government continuity systems, healthcare networks, national defense, emergency services, and hazardous material facilities is highly vulnerable to cyberattacks. Disruptions to these interconnected systems can trigger cascading impacts on public safety and economic stability, as evidenced by the 2007 Estonia cyberattacks and subsequent cyber operations associated with the Ukraine conflict [46].

Lansky et al. [8] provide a systematic review of DL-based IDS for this domain, cataloging methodologies, reported effectiveness, and key limitations, including strong dependence on the quality of training datasets. Ahmad et al. [43] complement this with an

extensive review of ML and DL-based NIDS, identifying DL as the dominant recent trend for performance improvement. Nguyen et al. [16] demonstrate the effectiveness of reinforcement learning (RL) for controlling and mitigating cyberattacks in SDN environments, presenting an adversarial Deep RL method that enhances security in autonomous systems and emphasizing the need for automated IDS capable of handling large volumes of audit data. Sanchez-Zas et al. [47] propose an unsupervised anomaly detection system using K-means clustering deployed at scale over Apache Spark [48], demonstrating practical suitability for heterogeneous real-time security device environments where labeled attack data is unavailable.

Deployment Readiness: DL-based NIDS have achieved laboratory-level maturity but face significant barriers to production deployment such as computational overhead, high false positive rates under real-traffic distributions, and the absence of standardized deployment pipelines. RL-based systems such as those proposed in [16] remain at an experimental stage, dependent on simulated environments that do not fully capture the complexity of live production networks. Unsupervised systems like [47] offer higher deployment feasibility in resource-constrained or heterogeneous environments but sacrifice specificity.

Strengths and Limitations: DL-based approaches offer strong performance on known attack signatures and generalize better to novel variants than classical ML, but their black-box nature limits operational trust in critical infrastructure contexts [14]. RL offers adaptability and real-time response optimization, but training instability and safety risks during live exploration constrain its deployment. Unsupervised approaches are more privacy-preserving and require no labeled data, but suffer from elevated false-positive rates that create alert fatigue [9].

Assumptions: The majority of studies assume that historical traffic distributions remain stationary. This assumption is routinely violated by evolving attack tactics, techniques, and procedures (TTPs). Many DL evaluations further assume that training data is clean and unmanipulated, which is challenged by poisoning attacks.

Evaluation Metrics and Datasets: Standard metrics include accuracy, precision, recall, F1-score, and false positive rate (FPR). Benchmark datasets include NSL-KDD [49], UNSW-NB15 [13], CICIDS2017 [11], and CSE-CIC-IDS2018 [12]. These datasets are widely used but are acknowledged to contain redundant records, lack modern encrypted traffic, and reflect controlled environments that do not replicate the statistical properties of production networks.

Model Robustness: Systems evaluated exclusively on clean benchmark data are demonstrably fragile against adversarial evasion attacks. Adversarially crafted network packets can bypass both ML and DL classifiers, as documented in the adversarial attack taxonomy shown in Table 3. DRL systems deployed in adversarial SDN environments [16] show promising robustness under adversarial conditions, but generalizability across attack types remains limited.

Governance Implications: Automated threat detection in critical infrastructure raises significant accountability concerns. Misclassifications—particularly false negatives—can have catastrophic consequences. Regulatory frameworks such as NERC CIP in energy sector [50] and IEC 62443 for industrial systems [51] impose compliance requirements, such as strict logging, validation rules, and fail-safe operations, on automated security decisions. The deployment of black-box DL models in these contexts requires XAI mechanisms to satisfy audit and accountability obligations.

Table 3. Overview of Adversarial Attacks.

Attack Type	Target	Description	Impact
Evasion Attacks	Trained models or classifiers	Adversarial inputs manipulate model predictions during inference	Bypass detection and misclassification
Poisoning Attack	Training datasets	Handcrafted malicious samples injected during training	Corrupted models and compromised threat detection and prediction
Prompt Injection	LLM and Security copilots	Malicious prompts to manipulate model behavior	Incorrect and unsafe outputs and bypass safeguards
Model Inversion	Trained models or classifiers	Inferring sensitive training information from model outputs	Privacy leakage and exposing sensitive data
Jailbreaking	LLMs and transformer-based models	Handcrafted prompted to bypass restrictions	Model generates malicious and prohibited content

3.1.2. Application Domain: Cyber-Physical Systems (CPS)

Conventional Industrial Control Systems (ICSs) were originally designed for reliability and longevity in isolated environments. Digital transformation has progressively converged these systems into Cyber-Physical Systems (CPSs), where computational components interact directly with physical processes through sensors, actuators, and control logic. This convergence introduces cyber interdependencies that legacy security mechanisms—firewalls, basic anomaly detection, application whitelisting—are ill-equipped to address [52,53].

In CPS environments, security failures extend beyond data compromise to include physical damage, safety hazards, and service disruption, as attackers can exploit HMIs, Safety Instrumented Systems, and network devices to manipulate setpoints, disable alarms, or alter process behavior [54]. The Stuxnet attack remains the defining case study, demonstrating how cyber intrusions can deliberately induce physical damage by targeting control logic rather than information assets [55]. To address these risks, the literature increasingly employs time-series anomaly detection using ML and DL models, complemented by physics-based models and digital twins that continuously compare real-time system behavior against virtual replicas for early detection of cyber-physical anomalies.

Deployment Readiness: CPS security AI is at a moderate deployment readiness level. Physics-informed models and digital twin approaches are seeing increasing adoption in smart grid and industrial automation contexts, but integration complexity, latency requirements, and the absence of standardized CPS security datasets limit broader deployment. Safety-critical environments impose strict latency and reliability constraints that many DL models cannot satisfy in real time.

Strengths and Limitations: The fusion of data-driven ML/DL anomaly detection with physics-based models provides a complementary detection capability that is more robust to distributional shift and more interpretable to domain engineers. However, CPS environments generate highly domain-specific time-series data that does not transfer well across sectors (e.g., models trained on smart grid data do not generalize to water treatment plants). The scarcity of labeled attack data in operational CPS environments further restricts supervised learning approaches.

Assumptions: Models assume that the physics of the underlying process are stable and well-characterized. In practice, legitimate operational changes (maintenance modes, seasonal variations) can trigger false positives if not modeled explicitly. Most evaluations

assume that the attacker cannot observe or manipulate the detection system, an assumption that is increasingly untenable for persistent, targeted adversaries.

Evaluation Metrics and Datasets: CPS security studies report detection rate, false alarm rate, and in some cases, time-to-detection. Publicly available CPS datasets are limited; the SWaT (Secure Water Treatment) [56] and BATADAL (Battle of the Attack Detection Algorithms) [57] datasets are frequently used but are narrow in scope and do not represent ICS diversity.

Model Robustness: The absence of adversarial robustness evaluation in most CPS security studies is a significant gap. Models evaluated only on benign operational data and standard attack scenarios may fail against adaptive adversaries who have knowledge of the deployed detection logic—a realistic threat model for APT actors targeting critical infrastructure.

Governance Implications: Safety and security are tightly coupled in CPS environments. Autonomous AI-driven responses such as automatically shutting down a process upon anomaly detection can carry the risk of triggering safety incidents worse than the attack itself. Human-in-the-loop oversight is therefore not merely a governance preference but an operational necessity in CPS security, requiring clear escalation paths and override mechanisms.

3.1.3. Application Domain: IoT and Other Autonomous Systems

The rapid growth of IoT devices introduces security challenges rooted in scale, heterogeneity, and resource constraints. IoT devices operate in unattended environments, often lack the computational resources for on-device security, and may be produced by manufacturers with limited security expertise, creating a broad and difficult-to-monitor attack surface that requires a defense-in-depth approach.

The literature has converged on several complementary strategies for AI-driven IoT security. Device behavior profiling using ML classifiers enables identification of rogue or compromised devices through deviation from normal traffic signatures [58,59]. Malware analysis for firmware-vulnerable devices has leveraged both static analysis of binary code and dynamic behavioral analysis [60,61]. Edge AI paradigms—deploying lightweight ML models on or near IoT devices—are gaining traction as a means of reducing latency and preserving data locality [61,62]. At the network level, the combination of AI techniques with Software-Defined Networking (SDN) provides enhanced visibility and adaptive defense against DDoS attacks, including low-rate stealthy variants [63,64]. Vehicle security represents an emerging sub-domain, with AI-driven safety features being increasingly adopted to safeguard vehicular networks and connected autonomous vehicles [65]. As indicated in Table 2, classical ML models with low computational overhead are well-suited for deployment on IoT end- devices and sensors, whereas FL enables collaborative learning across IoT environments without raw data sharing, directly addressing the privacy and data locality constraints characteristic of this domain.

Deployment Readiness: Edge AI for IoT is the most deployable paradigm in this domain, with commercial deployments in industrial IoT (IIoT) and smart home contexts. FL for IoT security is advancing toward production readiness, with implementations demonstrated on heterogeneous device federations, but communication overhead and model synchronization across resource-constrained devices remain practical barriers.

Strengths and Limitations: Classical ML offers low inference latency and interpretability suitable for embedded contexts. FL preserves data privacy across distributed device networks and enables collective threat detection without centralizing sensitive telemetry. However, FL introduces new vulnerabilities to poisoning attacks from compromised participating nodes, and heterogeneous data distributions across devices can degrade global

model quality. DL-based approaches offer superior pattern recognition but are generally infeasible for direct deployment on resource-constrained IoT endpoints.

Assumptions: IoT security models frequently assume that device traffic profiles remain stable over time. In practice, firmware updates, changing usage patterns, and environmental factors alter normal behavior, requiring adaptive or continually updated baselines. FL-based systems assume that the federation includes a sufficient proportion of honest nodes to resist poisoning—an assumption that may not hold in adversarial IoT deployments.

Evaluation Metrics and Datasets: Precision, recall, F1-score, and detection latency are the predominant metrics. Representative IoT datasets include Bot-IoT [66], ToN-IoT [67], N-BaIoT [68], and CICIoT2023 [69]. These datasets vary in realism, attack diversity, and feature representation, and most suffer from synthetic or laboratory-generated traffic that does not capture the full heterogeneity of production IoT environments.

Model Robustness: IoT ML models are particularly susceptible to adversarial manipulation because device-level security controls are minimal. Evasion attacks crafted to mimic normal device traffic can bypass behavior-profiling classifiers. FL models face the additional threat of Byzantine attacks, where compromised devices contribute poisoned gradients to corrupt the global model.

Governance Implications: Consumer IoT devices collect personally identifiable information (PII), triggering data protection obligations under frameworks such as GDPR and CCPA. AI models that monitor or profile device behavior must be designed with privacy-by-design principles. Industrial IoT deployments in critical sectors are additionally subject to sector-specific compliance requirements, and autonomous security responses on industrial equipment may require human approval to prevent operational disruption.

3.1.4. Application Domain: Healthcare Systems

The digitization of healthcare has created an extensive attack surface centered on Electronic Health Records (EHRs), medical IoT devices, diagnostic imaging systems, and interconnected hospital networks. Healthcare organizations have become high-priority targets for ransomware and data theft due to the high value and sensitivity of patient data and the life-critical nature of the services they provide [70,71].

Healthcare infrastructures hold comprehensive patient health and medical histories in EHR systems that must be exchanged securely among personnel, specialists, and authorized caregivers while remaining protected as highly sensitive PII [72]. Breaches such as the Anthem cyberattack [73] underscore the severity of unauthorized disclosure and regulatory violations [74]. AI-driven cybersecurity systems including ML-based IDS, anomaly detection, and behavioral analytics safeguard healthcare environments by continuously monitoring network activity, medical device behavior, and EHR access patterns to identify and mitigate threats in real time. The healthcare domain is distinctive in that it spans multiple application domains from the Figure 2 taxonomy simultaneously: it involves IoT-class medical devices, CPS-class clinical equipment (e.g., infusion pumps, ventilators), enterprise networks, and cloud-based EHR platforms. This convergence demands multi-layered AI security architectures.

Deployment Readiness: AI-driven healthcare security is at an early-to-moderate deployment readiness level. ML-based anomaly detection has seen commercial adoption in healthcare SIEM platforms, but the regulatory approval pathway for AI- driven autonomous security decisions in clinical environments is underdeveloped. Integration with legacy clinical systems is a significant practical barrier.

Strengths and Limitations: Behavioral analytics offer the ability to detect insider threats and unauthorized EHR access patterns that traditional rule-based systems cannot identify. However, the sensitivity of healthcare data constrains the use of centralized ML training,

making FL an increasingly important paradigm for this domain. The limited availability of publicly accessible, labeled healthcare security datasets substantially restricts model development and benchmarking.

Assumptions: Healthcare security models assume relatively stable access patterns for clinical staff, which may not hold during emergencies, staff rotations, or system migrations. Models trained on one hospital's data may not generalize to another due to differences in workflow, EHR systems, and device inventories.

Evaluation Metrics and Datasets: Detection accuracy, false positive rate, and time-to-detection are the primary metrics. Publicly available healthcare-specific cybersecurity datasets are scarce; researchers typically adapt general-purpose IDS datasets such as CICIDS [12] or use proprietary clinical datasets that cannot be shared.

Model Robustness: Healthcare AI security models are subject to both adversarial evasion (crafting access patterns that mimic legitimate clinical activity) and data poisoning (manipulating training data to suppress detection of specific threat actors). Robustness evaluation is largely absent from the published healthcare security literature.

Governance Implications: Healthcare AI security is subject to stringent regulatory requirements: HIPAA (US), GDPR Article 9 (EU), and national equivalents impose obligations on data processing, breach notification, and the use of automated decision-making for sensitive health data. The life-critical nature of healthcare systems requires that autonomous AI-driven responses—such as isolating a compromised medical device—be executed with extreme caution, as false positives can interrupt patient care.

3.2. Security Capability: Incident Response Automation

Incident response (IR) is a structured, coordinated process aimed at detecting, containing, and recovering from cyber incidents including advanced persistent threats (APTs) and data breaches while minimizing damage and preventing recurrence [75]. Traditional IR processes follow defined phases: preparation, detection, containment, eradication, recovery, and post-incident analysis. These phases are increasingly executed within AI-enabled Security Operations Centers (SOCs) that serve enterprise networks, cloud environments, and 5G/edge infrastructure simultaneously.

The primary challenge driving AI adoption in IR is the scale mismatch between modern threat environments and human analytical capacity: high alert volumes, significant false-positive rates, and the speed of AI-enabled attacks combine to create alert fatigue that degrades SOC performance [76]. AI-driven techniques such as ML classifiers, NLP-based alert triage, and behavioral analytics have been integrated as intelligent layers over Security Information and Event Management (SIEM) platforms to prioritize alerts by severity and enable faster coordination of response actions [77]. This evolution has culminated in Security Orchestration, Automation, and Response (SOAR) platforms, which automate response actions across diverse security tools using predefined playbooks while retaining human oversight [78,79].

Empirical validation from a financial-sector case study by Waelchli and Walter [78] demonstrates that SOAR mechanisms reduce human susceptibility to social engineering attacks by coordinating mitigation once indicators are identified—even when attacks lack detectable malicious payloads. Modern SOAR platforms including Fortinet FortiSOAR, IBM QRadar SOAR, and Palo Alto Cortex XSOAR integrate ML-based anomaly detection to reduce false positives and automate response actions at machine speed. As designated in Table 2, LLMs are particularly suited to the SOC and management layer for processing unstructured threat reports, synthesizing cyber threat intelligence (CTI), and automating security documentation—tasks that are central to the IR workflow.

Deployment Readiness: SOAR-integrated AI for IR is commercially deployed across enterprise and financial-sector SOCs. LLM-based automation of threat report analysis and playbook generation is at early production readiness, with commercial products such as Microsoft Security Copilot and SentinelOne Purple AI demonstrating integration into real-world workflows. However, full automation of containment and remediation decisions remains at a pilot stage in most organizations, with human-in-the-loop (HITL) requirements constraining the degree of autonomous action.

Strengths and Limitations: AI-augmented IR substantially reduces mean time to respond (MTTR) and analyst workload, enabling SOCs to scale against AI-accelerated attacks. However, the black-box nature of many ML models in IR systems creates transparency deficits that limit analyst trust and regulatory compliance [80,81]. Automation bias—the tendency of analysts to over-rely on automated recommendations—is an emerging operational risk. Explainable AI (XAI) techniques, including LIME and SHAP, partially address interpretability gaps [82] but introduce additional computational overhead and do not yet generalize well to complex DL models.

Assumptions: SOAR playbooks assume that attack patterns conform to predefined response categories; novel attack types that do not match existing playbooks may be mishandled or escalated with insufficient context. LLM-based components assume that structured threat intelligence exists and is accessible, and that model outputs are sufficiently accurate to be acted upon without additional validation—an assumption that is undermined by hallucination risks.

Evaluation Metrics and Datasets: Standard IR metrics include MTTR, false positive rate, alert reduction ratio, and containment success rate. Empirical evaluation of AI-driven IR systems is predominantly conducted through case studies or controlled scenarios; standardized IR benchmark datasets do not yet exist, limiting rigorous comparative evaluation.

Model Robustness: Automated IR systems are high-value targets for adversarial manipulation: attackers who understand the deployed detection logic can craft attacks that evade automated triage, exploit automation bias in analysts, or deliberately trigger false alarms to obscure genuine intrusions. LLM-based IR components are vulnerable to prompt injection attacks that could manipulate automated response recommendations.

Governance Implications: Automated IR decisions carry significant accountability implications, particularly when containment actions such as network isolation and account suspension affect production systems or customer data. The responsibility gap identified in the literature [83]—where accountability for automated actions is distributed across system designers, security teams, and organizational leadership rather than a single decision-maker—is especially acute in IR contexts. Regulatory frameworks such as NIS2 (EU) [84] and sector-specific standards require documented, auditable incident response processes; automated systems must therefore maintain comprehensive logs and support human override.

3.3. Security Capability: Threat Intelligence and Security Analytics

Cyber threat intelligence (CTI) involves the systematic collection and analysis of data on potential or active threats to transform raw information into actionable insights about adversaries' motives, techniques, targets, and vulnerabilities. Effective CTI enables organizations to move from reactive defense to proactive threat anticipation [85]. It is classified into strategic, operational, tactical, and technical tiers—from high-level trend analysis to machine-consumable indicators of compromise (IoCs)—forming a hierarchy whose upper layers require human analytical capability and whose lower layers are amenable to AI automation.

AI-driven security analytics enable predictive threat intelligence by analyzing historical attack patterns, adversary behavior, and contextual indicators to forecast potential

attack scenarios [86]. When integrated with SIEM and SOAR platforms, these analytics transform CTI from a passive information source into an active decision-support layer by enriching alerts with contextual intelligence and prioritizing risks. GenAI is increasingly applied to CTI workflows: LLMs enable automated parsing of unstructured threat reports, dark web content, social media, and security forums; VAEs and GANs enhance pattern discovery, cross-source correlation, and synthetic threat scenario generation [87]. Commercial platforms including Microsoft Security Copilot, Recorded Future AI, and SentinelOne Purple AI represent early production deployments of LLM-powered CTI. The WannaCry ransomware campaign illustrates both the critical need for real-time AI-enhanced CTI and its limitations: rapid global propagation exploited the gap between threat discovery and intelligence dissemination, a gap that AI-augmented CTI systems are specifically designed to close.

Deployment Readiness: LLM-based CTI automation is at early-to-moderate production readiness. Commercial platforms demonstrate viability for structured intelligence extraction and report generation, but accuracy and reliability on adversarial or low-quality intelligence sources remain inconsistent. ML-based IoC correlation and threat scoring are more mature and widely deployed in enterprise SIEM stacks.

Strengths and Limitations: AI-driven CTI dramatically scales the analytical capacity of threat intelligence teams, enabling processing of data volumes and source diversity that are infeasible for manual analysis. LLMs provide unprecedented capability for natural language understanding of unstructured threat data. However, LLMs hallucinate—generating plausible but incorrect intelligence—which is particularly dangerous in CTI contexts where acting on false information can misallocate defensive resources or trigger inappropriate responses [33]. Multimodal GenAI additionally introduces risks: image and video-based intelligence generation can be weaponized for adversarial content creation [87].

Assumptions: CTI systems assume that historical attack patterns and known adversary TTPs remain predictive of future behavior—an assumption that becomes less reliable as adversaries adapt and change tactics. LLM-based CTI systems further assume that source data quality is sufficient for reliable extraction; in practice, threat intelligence feeds contain significant noise, duplicate reporting, and deliberate disinformation.

Evaluation Metrics and Datasets: CTI quality is assessed through precision of IoC extraction, coverage of threat actor TTPs such as MITRE ATT&CK coverage and analyst-rated utility scores. Standardized CTI evaluation datasets are limited; MITRE ATT&CK provides a structured framework for adversary behavior taxonomy but does not constitute a labeled evaluation dataset.

Model Robustness: CTI systems are vulnerable to adversarial disinformation campaigns: threat actors can deliberately inject misleading information into public threat feeds, social media, and dark web forums to manipulate AI-driven CTI systems. LLMs are additionally susceptible to prompt injection when processing adversarially crafted threat reports [34]. Robustness of CTI models under deliberate intelligence poisoning has received limited research attention.

Governance Implications: CTI sharing is subject to legal and ethical constraints: intelligence derived from dark web monitoring may involve legally ambiguous collection methods, and cross-organizational threat sharing raises data protection obligations. AI-generated CTI must be clearly distinguished from analyst-validated intelligence in operational workflows to prevent automation bias from translating AI errors into operational decisions. LLM-based CTI components should be subject to output validation requirements and hallucination monitoring protocols.

3.4. Security Capability: Security Automation in Cyber Warfare and Military Networks

Modern conflict has transitioned into the digital domain, with AI enabling both offensive cyber operations and AI-augmented defense of military networks [88,89]. Military capabilities such as communication, command and control (C2), logistics, targeting are increasingly dependent on hybrid cyber power, making cybersecurity a strategic national priority [90]. Effectively integrated with traditional military domains, cyber power creates operational asymmetries and reduces the need for physical troop deployment [91].

The emergence of Autonomous Warfare Systems integrates AI into unmanned vehicles, robotics, and weapons platforms capable of selecting and engaging targets with minimal human intervention [92–95]. While these systems promise increased precision, endurance, and reduced risk to soldiers, they introduce profound ethical, legal, and cybersecurity challenges. Fully autonomous systems that rely on facial recognition and algorithmic target identification raise serious accountability, international humanitarian law (IHL) compliance, and human rights protection concerns [96–100]. The complexity and decentralized nature of these systems make them difficult to control, predict, and secure, rendering them vulnerable to system failures, spoofing, adversarial AI attacks, and cyberattacks that could produce catastrophic outcomes before human intervention is possible [101]. The absence of comprehensive international regulatory frameworks exacerbates risks of misuse, civilian harm, and geopolitical instability [102–104]. In the context of cybersecurity operations specifically, AI-driven security automation enables real-time detection and response to cyber operations targeting military networks. RL-based autonomous agents are a leading research paradigm for adaptive cyber defense in military contexts [105], with multi-agent RL (MARL) architectures that combine LLMs and rule-based systems proposed for automating monitoring, detection, and mitigation in military network environments [106]. While these broader autonomous-warfare debates are not themselves cybersecurity problems, they establish the risk context in which AI-driven cyber defense of military networks must operate. The governance gaps like accountability, meaningful human control, and regulatory lag that concern autonomous weapons also apply directly to autonomous cyber-defense agents discussed here.

Deployment Readiness: AI-driven autonomous cyber defense for military networks is at an early laboratory stage. The CAGE competition framework [105] provides simulated training environments, but the gap between simulated and operational military network complexity is substantial. Ethical, legal, and policy constraints further limit deployment timelines for fully autonomous systems.

Strengths and Limitations: Autonomous cyber defense agents offer machine-speed response to cyber threats, reduce the workload on human cyber operators, and can continuously adapt to evolving adversarial tactics. However, RL-based agents trained in simulation exhibit significant performance degradation when transferred to real environments. Autonomous warfare systems and autonomous cyber agents lack the capacity to interpret ethical dilemmas, and the accountability vacuum created by removing humans from the decision loop violates foundational principles of IHL.

Assumptions: RL-based cyber defense research assumes that simulated training environments adequately represent operational conditions—an assumption that is increasingly questioned in the autonomous cyber defense literature. The targeting subsystems of autonomous warfare systems assume that sensor data accurately represents the operational environment and that algorithmic classification of threat targets meets the legal standards for proportionality and distinction required under IHL.

Evaluation Metrics and Datasets: Military network security evaluations use simulation-based metrics including detection rate, response time, and mission success rate in

adversarial scenarios. Operational datasets from military network environments are classified and unavailable for open research, limiting external validation.

Model Robustness: Military AI systems are high-priority adversarial targets. Spoofing attacks against sensor inputs, adversarial manipulation of perception systems, and cyberattacks on communication channels can mislead autonomous agents into catastrophic decisions [101]. Robust defense mechanisms for military AI systems require adversarial training under classified threat models that are inaccessible to the open research community.

Governance Implications: The governance implications of AI-driven military operations are among the most consequential in this survey. The responsibility gap [83] and the requirement for meaningful human control [107] are critical principles in this domain, where automated decisions may have irreversible lethal consequences. International governance frameworks like the Convention on Conventional Weapons (CCW) discussions on LAWS, ICRC guidance on autonomous weapons provide a partial regulatory foundation but lack binding enforcement mechanisms. Defense policymakers face the urgent challenge of balancing technological advantage with civilian safety and ethical accountability [102–104].

3.5. Security Capability: Digital Forensics

Digital forensics is the process of the systematic collection, preservation, analysis, and presentation of digital evidence. This field is increasingly augmented by AI techniques that accelerate evidence processing, malware attribution, and incident reconstruction at a scale infeasible for manual methods. While less prominently represented in the current survey corpus than detection and response capabilities, digital forensics constitutes an explicit security capability in the Figure 2 taxonomy and merits critical assessment.

ML-based malware classification using static analysis such as PE header features, byte histograms and dynamic behavioral analysis like API call sequences has reached high maturity levels for known malware families presented in Table 4 [108–115]. DL-based approaches including convolutional analysis of malware binary visualizations (grayscale images) and sequence models over API call logs demonstrate strong classification performance on benchmark datasets such as EMBER, SOREL-20M, and Malimg [108,109,111]. LLMs represent an emerging capability for automated malware report generation, vulnerability attribution, and forensic documentation synthesis. GenAI introduces a dual-use concern in digital forensics: the same generative capabilities that support synthetic attack sample generation for training data augmentation [32] can be weaponized to produce adversarial malware variants that evade forensic classifiers, deepfake artifacts that complicate attribution, and synthetic evidence that undermines forensic integrity.

Deployment Readiness: ML-based malware classification is commercially deployed in endpoint detection and response (EDR) products and threat intelligence platforms. LLM-based forensic automation is at early proof-of-concept stage. Forensic AI tools face specific deployment barriers including chain-of-custody requirements for evidence admissibility and explainability requirements for use in legal proceedings.

Strengths and Limitations: AI-driven malware analysis dramatically scales detection capacity and reduces analyst time for known family classification. However, models trained on historical malware families demonstrate limited generalization to novel, obfuscated, or polymorphic variants. Static analysis features are particularly susceptible to obfuscation; dynamic behavioral analysis is more robust but requires sandbox execution infrastructure.

Assumptions: Malware classification models assume that sample labeling in training datasets is accurate and consistent. That is an assumption frequently violated in datasets assembled from heterogeneous sources with inconsistent labeling methodologies. Models further assume that the distribution of malware families in the training set reflects operational prevalence, which changes over time as new malware campaigns emerge.

Table 4. Malware Detection Datasets.

Datasets	Attacks	Features	Limitation
EMBER [108]	Windows malware families	Portable Executable (PE) header features, byte histograms	Suitable for static analysis
SOREL-20M [109]	Ransomware, Dropper, Worm, Trojan, etc	PE metadata behavioral labels	large computational requirements
Microsoft Malware Challenge [110]	Ramnit, Kelihos, Lollipop, Obfuscator	Assembly code, byte code	Old malware families
Malimg [111]	25 malware families with various malware types	Grayscale malware images	Limited semantic behavior information
Malnet [112]	Various malware families	Large malware image corpus	Resource intensive training
CICMalMem 2020 [114]	Ransomware, file-less malware	Memory dump artifacts	Limited malware family diversity
VirusShare [113]	Diverse malware binaries	Raw malware samples	Inconsistencies in labeling
API Call Sequence Dataset [115]	Execution behavior data	API call sequences	High-dimensional sequential data

Evaluation Metrics and Datasets: Classification accuracy, F1-score per family, false negative rate, and detection time are standard metrics. Key datasets, presented in Table 4 include EMBER [108], SOREL-20M [109], Microsoft Malware Classification Challenge [110], Malimg [111], MalNet [112], VirusShare [113], CICMalMem-2022 [114], and the APT Call Sequence Dataset [115].

Model Robustness: Malware classifiers are a primary adversarial target: evasion attacks that modify malware to mimic benign software behavior routinely defeat both signature-based and ML-based classifiers. Adversarial robustness evaluation is inconsistently reported in the forensics literature, limiting assessment of real-world reliability.

Governance Implications: AI-assisted digital forensics carries specific legal implications: evidence processed or classified by AI systems must meet admissibility standards, which typically require explainable, reproducible, and auditable decision processes. Automated malware attribution—identifying a threat actor responsible for an incident—carries geopolitical and legal implications and must be treated with exceptional caution when used to inform policy or legal responses.

The subsections above demonstrate that the security capabilities and application domains defined in Figure 2 do not map uniformly to a single AI paradigm. Different techniques offer complementary strengths and gaps, as illustrated in Figure 3.

The synthesis reveals several cross-cutting patterns that warrant emphasis. First, there is a systematic mismatch between reported benchmark performance and operational deployment readiness across all security capabilities: models achieving 99%+ accuracy in controlled evaluations consistently encounter higher error rates under real-world distribution shift, adversarial conditions, and data quality limitations. Second, model robustness against adversarial manipulation is the most underexplored dimension across the literature—particularly in CPS, healthcare, IoT, and military domains—suggesting that current AI security deployments may carry substantially higher operational risk than published performance metrics imply. Third, governance and accountability frameworks lag significantly behind technical capability, with the responsibility gap, automation bias, and explainability deficit posing systemic risks that technical performance improvements alone cannot resolve. Finally, as both Figure 3 and Table 2 illustrate, no single AI paradigm

dominates across all security capabilities and application domains: effective deployments typically require paradigm-mixing—for example, classical ML for resource-constrained IoT detection combined with FL for privacy-preserving aggregation and LLMs for SOC-level CTI synthesis—guided by the deployment decision criteria consolidated in Table 2.

4. Challenges, Emerging Trends, and Open Research Directions

The fourth and fifth dimensions of the proposed taxonomy in Figure 2 represent the critical interface between AI-driven cybersecurity research and its operational realization. Dimension 4—Deployment and Operational Challenges—enumerates the structural barriers that constrain reliable field deployment: data quality and availability, adversarial AI and evasion, privacy, ethics and governance, explainability and interpretation, skilled workforce, and cost and resource constraints. Dimension 5—Emerging Trends—captures the frontier capabilities that are beginning to reshape security operations: autonomous cyber defense, AI-augmented Security Operations Centers (SOCs), human–AI collaboration, zero-trust architecture, and agentic AI for cybersecurity automation. This section synthesizes the literature across both dimensions. Each challenge and trend is analyzed critically along axes like technical severity or maturity, affected AI paradigms (cross-referencing capability matrix in Figure 3), current mitigation or enabling approaches and discuss our insights.

4.1. Deployment and Operational Challenges

4.1.1. Datasets

The performance of AI-driven security systems is fundamentally bounded by the availability of high-quality, representative, and well-labeled training data. In cybersecurity, this constraint is particularly acute as ground-truth labels for novel attacks are rare, class distributions are severely skewed towards benign traffic, and publicly available benchmark datasets such as those presented in Tables 4–6 have well-documented coverage gaps like outdated attack profiles, synthetic or lab-generated traffic, and limited encrypted-traffic realism [77]. The consequence is that models evaluated on benchmark datasets frequently fail to generalize to live, heterogeneous production environments where attack distributions shift continuously.

Technical Severity: High. Class imbalance causes skewed decision boundaries that inflate detection rates on majority classes while under-detecting rare, high-impact threats. The gradual divergence between the statistical distribution of training data and live traffic as attacker tactics evolve, known as concept drift, progressively degrades model performance without triggering obvious failure signals [77]. Adversaries can deliberately amplify this effect through data poisoning, injecting deceptive patterns that bias learned representations.

Affected AI Paradigms: All paradigms in Figure 3 are affected, but supervised ML and DL are most sensitive to label quality and imbalance. Unsupervised methods alleviate the labeling burden but generate high false-positive rates unsuitable for production without human oversight [8,9]. FL partially mitigates data scarcity through distributed learning but introduces heterogeneous data distributions across participants that complicate aggregation and model convergence.

Mitigation Approaches: Data augmentation using GANs and VAEs can synthesize realistic minority-class attack samples to redress imbalance [31]. Transfer learning and domain adaptation reduce dependence on target-domain labels by leveraging pre-trained representations. Continual and online learning frameworks address concept drift through incremental model updates, though stability–plasticity trade-offs remain unresolved.

Insights: There is no universally accepted data-collection standard or labeling protocol for cybersecurity datasets. Datasets for APTs, encrypted malware, and AI-targeted supply-chain attacks remain critically scarce. The gap between published dataset performance and

real-world utility is rarely quantified in published evaluations, creating a reproducibility problem across the literature. Additionally, collection of labeled network traces raises significant data-protection obligations under GDPR and equivalent frameworks. Sharing sensitive operational logs across organizations to improve dataset diversity requires explicit consent.

Table 5. Network Intrusion Detection (IDS/IPS) Datasets.

Datasets	Attacks	Features	Limitation
KDD Cup [116]	DoS, Probe, R2L, U2R	41 TCP/IP handcrafted features	Outdated, redundant records, unrealistic traffic
NSL KDD [49]	DoS, Probe, R2L, U2R	Refined KDD99 feature set	Outdated, lacks modern attacks
UNSW-NB15 [13]	Fuzzers, DoS, Reconnaissance, Shellcode, Exploits, worms	49 flow-based network features	Limited encrypted traffic realism
CICIDS2017 [11]	DDoS, Brute Force, Botnet, PortScan, Web attacks, Infiltration	Flow statistics, packet metrics	Lab-generated environment
CSE-CIC-IDS2018 [12]	DDoS, Brute Force, DoS, Botnet, Infiltration	Netflow statistical Features	High processing complexity and controlled environment
CIC-DDoS2019 [44]	WebDDoS, Botnet, UDP Flood, Amplification, SYN Flood attacks	Flow-based statistical features	Focuses only on DDoS attacks and controlled environment
Kyoto 2006+ [117]	Malware, Scanning, unknown attacks	Honeypot logs and traffic flows	Limited attack labeling
HIKARI-2021 [118]	Malware traffic VPN and non-VPN attacks	Statistical flow features	Limited attack diversity

Table 6. IoT and IIoT Datasets.

Datasets	Attacks	Features	Limitation
ToN-IoT [67]	DDoS, Ransomware, Backdoor, Injection, Password attacks	Zeek logs and features	Complex, multi modal records and preprocessing
Bot-IoT [66]	DDoS, Dos, Scanning, Keylogging	Argus tool-based flow features	Synthetic traffic, lacks modern attacks
IoT-23 [119]	Mirai, Torii, Okiru, Botnets	Zeek logs and features	Labeling is based on zeek features
N-BaIoT [68]	Mirai, BASHLITE	Device traffic statistics	Limited attack families
MQTT-IoT-IDS2020 [120]	MQTT Flood, Malformed packets, brute force	MQTT protocol features	limited scope
EdgeIIoTset [121]	DDoS, MITM, Injection, Malware	Edge traffic telemetry logs	Relatively new benchmark
CICIoT2023 [69]	DoS, Spoofing, Botnet, Reconnaissance	HIoT Flow features	Realism and reproducibility concerns

4.1.2. Adversarial AI and Evasion

AI-driven security systems are structurally vulnerable to adversarial manipulation. Evasion attacks, presented in Table 3, craft inputs at inference time to misclassify malicious

traffic as benign; poisoning attacks corrupt training data to embed hidden backdoors; model-inversion attacks extract sensitive information from model outputs; prompt injection exploits the instruction-following behavior of LLM-based security tools; and jailbreaking circumvents alignment constraints in foundational models [33–35]. The fundamental tension is that the same statistical learning properties that enable AI to generalize across unseen benign data also render it sensitive to adversarially constructed distributional shifts.

Technical Severity: Critical. Adversarial attacks using imperceptibly small input perturbations can significantly reduce model detection accuracy. Prompt injection against LLM-based security copilots like Microsoft Security Copilot and SentinelOne Purple AI can redirect automated remediation actions, suppress alerts, or exfiltrate context-window contents, directly undermining operational security [34].

Affected AI Paradigms: DL models are most vulnerable to gradient-based evasion due to their high-dimensional, differentiable decision surfaces. Transformer-based and LLM-based systems are uniquely susceptible to prompt injection and jailbreaking. Figure 3 rates robustness for both ML and DL as moderate, reflecting the absence of provably robust architectures at production scale. Federated learning introduces additional poisoning surfaces at individual participant nodes, where a minority of malicious clients can corrupt the global model.

Mitigation Approaches: Adversarial training—augmenting training sets with adversarially perturbed examples—improves empirical robustness but does not provide formal guarantees and degrades clean-data accuracy. Certified defenses based on randomized smoothing provide provable robustness within bounded perturbation radii but scale poorly to large DL architectures. LLM firewalls, differential privacy applied to training gradients, and OWASP LLM security guidelines offer partial mitigations for prompt-based attacks [33].

Insights: Certified robustness for transformer-scale models remains computationally intractable. Adversarial robustness and detection accuracy are fundamentally in tension as no architecture has yet achieved both simultaneously at production threat levels. Robust federated aggregation mechanisms that tolerate Byzantine participants without sacrificing convergence are an active and unresolved research problem.

4.1.3. Privacy, Ethics, and Governance

AI-driven cybersecurity systems require continuous collection and analysis of network traffic, system logs, and user behavior to achieve operational effectiveness, creating a fundamental tension between security monitoring and personal data protection [122]. This surveillance–privacy trade-off is exacerbated by generative models trained on sensitive organizational data, which may unintentionally memorize and leak credentials, medical records, or personal identifiers. Authors in [123] argue that a Good AI Society requires strong data governance, transparency, and accountability as foundational requirements rather than post hoc add-ons. Lifecycle-based ethical frameworks that integrate transparency, fairness, and privacy throughout AI development and deployment have been proposed [124], yet practical implementation remains constrained by data quality limitations and the black-box nature of high-performing models [77,125].

Technical Severity: High. The deployment of GenAI tools—including LLM-based security copilots—without comprehensive risk assessment increases the probability of data leakage, unclear accountability chains, and security incidents [126]. AI models trained on biased or unrepresentative data may produce systematically unfair predictions, amplifying discrimination or failing to detect threats disproportionately affecting specific network segments or user groups.

Affected AI Paradigms: FL models explicitly address privacy by keeping training data local; however, they remain vulnerable to model poisoning, inversion, and gradient-leakage attacks that can reconstruct sensitive training samples from shared parameter updates. Differential privacy and homomorphic encryption provide formal privacy guarantees but impose significant computational overhead and model accuracy penalties.

Mitigation Approaches: Privacy-preserving techniques—FL, differential privacy, homomorphic encryption, and secure multi-party computation—enable collaborative learning without exposing raw data. Algorithmic bias auditing tools such as IBM AI Fairness 360, Aequitas support fairness assessment. Explainable AI techniques such as LIME, SHAP improve transparency, though their applicability to high-capacity transformer models remains limited.

Insights: Quantifying the privacy achieved by iterative FL training in realistic cybersecurity deployments remains an open problem. The EU AI Act, and emerging national AI regulations impose explicit obligations on explainability, data minimization, and accountability for automated decisions.

4.1.4. Explainability and Interpretability

Despite their superior detection performance, DL and transformer-based AI models operate as black boxes that produce outputs without human-interpretable reasoning chains. Security analysts require understandable justifications for automated alerts and remediation decisions before taking operational actions, yet the black-box nature of high-capacity models limits this capacity [77,125]. XAI techniques, particularly LIME and SHAP, have gained adoption in cybersecurity, as discussed in the context of the SOAR systems covered in Section 3.2 [82], and the Mistral-7B XAI- SOC system [127] demonstrates that LLMs can be prompted to generate interpretable security response narratives from complex event logs.

Technical Severity: High. Unexplainable decisions reduce analyst trust, impede post-incident forensic analysis, and create accountability gaps—particularly when automated AI responses have operational or legal consequences. In high-stakes environments (critical infrastructure, military networks), the inability to audit AI decision logic undermines certification pathways and regulatory compliance.

Affected AI Paradigms: Classical ML (logistic regression, decision trees) is inherently more interpretable; however, ensemble methods (Random Forest, XGBoost) and all DL architectures require post hoc XAI instrumentation. The moderate capability described in Figure 3 for both ML and DL can be achieved with the help of XAI techniques such as LIME and SHAP. Transformer-based models such as LLMs present the most severe explainability gap: attention mechanisms provide partial insights but do not constitute formal explanations. RL-based autonomous agents are particularly opaque, as their learned policies emerge from reward maximization over millions of environment interactions.

Mitigation Approaches and Insights: Post hoc XAI (LIME, SHAP) provides local feature-importance explanations for individual predictions. Concept-based explanations and prototype-based methods offer more semantically meaningful interpretations for domain experts. Architecture-level interventions like attention visualization, decision distillation into interpretable surrogate models show potential to improve transparency at the cost of fidelity. LLM-generated narrative explanations [127] represent an emerging approach that translates model outputs into analyst-readable security summaries.

4.1.5. Skilled Workforce

The effective deployment of AI-driven security systems requires a workforce with concurrent expertise in cybersecurity domain knowledge, machine learning engineering, and AI system operations—a rare interdisciplinary combination that is in severe global

short supply. AI security copilots are explicitly framed in the literature as a mechanism for alleviating the skilled workforce shortage in SOCs [128], yet their adoption introduces its own human factors challenges: over-reliance on AI-generated recommendations, deskilling of analysts through automation, miscalibrated trust, and reduced situational awareness when AI systems fail silently or generate confident but incorrect outputs [129].

Technical Severity: Medium–High (organizational, rather than purely technical). Misuse of AI tools by under-trained analysts—including incorrect tuning of detection thresholds, misinterpretation of XAI outputs, or acceptance of LLM hallucinations as authoritative—can introduce systematic security gaps that adversaries can exploit. Alert fatigue, documented extensively in SOAR deployment studies [76,78], is partly a workforce problem: AI generates high volumes of alerts that human analysts cannot triage effectively without well-calibrated AI support.

Affected AI Paradigms: LLM-based security copilots may require particular analyst training to recognize and mitigate hallucinations and prompt injection risks. RL-based autonomous systems require specialist understanding of reward design and policy safety verification to supervise safely. FL deployments require data governance expertise at participating nodes that most operational teams do not currently possess.

Mitigation Approaches: AI-augmented SOC frameworks combine AI automation with human-in-the-loop (HITL) oversight to preserve analyst engagement and contextual judgment [130]. Trust calibration mechanisms such as providing confidence scores and uncertainty quantification alongside AI recommendations can help analysts maintain appropriate reliance.

Insights: Human–AI teaming in SOC environments is inadequately studied. Studying analyst skill retention under AI-assisted workflows could be a possible research direction.

4.1.6. Cost and Resource Constraints

The computational costs of training, deploying, and maintaining high-capacity AI security models, particularly transformer-based and LLM architectures, are substantial and frequently underestimated in research evaluations. While cloud-based deployment models offer elastic scalability [92], the operational cost of large model inference, combined with data engineering infrastructure, labeling pipelines, and continuous retraining to address concept drift, creates significant barriers for small-to-medium enterprises and resource-constrained deployment contexts such as IoT endpoints, industrial edge nodes, military tactical networks. The deployment decision matrix in Table 2 acknowledges this explicitly: classical ML is assigned to IoT end-devices precisely because DL and LLM models cannot operate within their resource envelopes.

Technical Severity: Medium (varies strongly by deployment context). For cloud-hosted enterprise security platforms, cost is a budget constraint rather than a technical blocker. For edge computing, IoT, and Critical Infrastructure Protection (CIP) contexts, the energy, memory, and latency requirements of DL and transformer models often render them operationally infeasible without model compression, quantization, or knowledge distillation.

Affected AI Paradigms: QML and LLM architectures carry the highest computational overhead; their real-time deployment requires specialized hardware such as GPU clusters and quantum processors which are unavailable in most operational security environments. FL introduces communication overhead through iterative model-update exchanges that can degrade performance in bandwidth-constrained edge environments. Compact ML models (shallow networks, decision trees, lightweight SVMs) remain the only viable option for resource-constrained endpoint deployment.

Mitigation Approaches: Model compression techniques such as pruning, quantization, and knowledge distillation—reduce inference costs by 10–100× with modest accuracy

penalties. Neuromorphic and specialized AI security accelerators such as Arm Ethos and Intel Movidius enable DL inference at the edge with significantly lower power consumption. Tiered security architectures that assign lightweight models to edge nodes and forward contextual summaries to cloud-hosted deep models represent a pragmatic hybrid.

Insights: Cost–performance trade-off benchmarks for AI security models under realistic operational constraints like memory limits, latency SLAs, energy budgets are limited in published literature.

4.2. Emerging Trends

4.2.1. AI-Augmented Security Operation Centers (SOC)

Modern SOCs increasingly deploy AI and ML to augment human analysts, automating repetitive alert triage, log analysis, and threat correlation to free cognitive capacity for strategic threat hunting and incident response. The integration of XAI methodologies is exemplified by the Mistral-7B-based system [127] that generates interpretable security responses from complex event logs by satisfying the dual requirements of detection performance and analyst trust. AI copilots in Cybersecurity Operations Centers (CSOCs) are framed as a structural solution to the skilled-workforce shortage, enabling analysts to maintain operational effectiveness under high alert loads [128]. Systematic reviews of AI-enabled cyber situation awareness (CSA) tools [130] document improved decision-making across multiple threat-severity levels, while agent-based architectures incorporating LLM assistants demonstrate SOC integration with commercial vendors like IBM, Microsoft, and Palo Alto Networks [106].

4.2.2. Autonomous Cyber Defense Systems

Autonomous cyber defense (ACD) systems aim to detect, analyze, and mitigate threats in real time with minimal or zero human intervention. RL has emerged as the leading paradigm for ACD agent development, motivated by its success in adaptive sequential decision-making under uncertainty [16,105]. The Cyber Autonomy Gym for Experimentation (CAGE) competition framework provides standardized training environments for RL-based defense agents. Multi-agent RL (MARL) architectures have been extended to military network contexts by Loevenich et al. [131], integrating LLMs, rule-based systems, and RL to automate monitoring, detection, and attack mitigation in tactical environments. LLM-assisted agentic frameworks (IBM, Microsoft, Palo Alto Networks) demonstrate integration of autonomous reasoning with conventional security tooling [106].

4.2.3. Human-AI Collaboration

Human–AI collaboration frameworks represent the integration model that most directly addresses the workforce challenge in Section 4.1.5 and the governance requirements of AI deployment section discussed in Section 4.1.3. While AI systems automate pattern recognition, alert correlation, and routine remediation, human analysts contribute contextual judgment, ethical reasoning, and accountability that AI systems cannot independently provide. HITL architectures in SOC environments have been shown to reduce cognitive workload while maintaining analyst engagement and situational awareness [130]. The risk of excessive automation—leading to deskilling, alert blindness, and over-reliance on AI-generated conclusions—is documented in the literature on AI copilots [128,129] and underscores the need for carefully calibrated automation boundaries.

4.2.4. Zero-Trust Architecture

Zero-trust architecture, formalized by NIST SP 800-207 [132], eliminates implicit trust from network design: every user, device, and system must be continuously verified regardless of network position. AI plays a central enabling role in ZTA by providing the

dynamic risk scoring, behavioral biometrics, and continuous authentication that static rule-based systems cannot deliver at scale. Ramezanpour and Jagannath [133] demonstrate RL-based dynamic authorization for 5G/6G networks under ZTA. Hussain et al. [134] deploy FL for privacy- preserving continuous monitoring within a zero-trust framework, directly linking ZTA to the data-privacy challenge in Section 4.1.3. Javeed et al. [135] combine CNN and BiLSTM for intrusion detection in zero-trust contexts, while blockchain-enabled FL (BFL) frameworks [132] address ZTA in distributed remote-work environments. AI-driven UEBA integrated with ZTA provides continuous post-authentication behavioral monitoring that detects compromised credentials and insider threats in real time [136].

4.2.5. Agentic AI for Cybersecurity and Automation

Agentic AI—systems that combine perception, planning, tool use, and autonomous action execution through LLM reasoning—represents the most transformative and least mature capability on Figure 2's emerging trends axis. Agentic security systems integrate LLM reasoning with external tooling (SIEM queries, firewall APIs, vulnerability scanners, threat intelligence feeds) to autonomously orchestrate multi-step security workflows: threat hunting, vulnerability chain analysis, automated penetration testing, and incident response playbook execution. The paper [106] analyzes vendor implementations and finds that current agentic deployments combine modular AI agents with LLM assistants in SOC workflows. On the other hand, GenAI-driven platforms—Microsoft Security Copilot, SentinelOne Purple AI, Recorded Future AI—demonstrate early agentic capability in commercial products [29]. The dual-use concern is particularly acute for agentic AI: the same tool-using, multi-step planning capabilities that enable autonomous defense can be weaponized for autonomous attack campaign orchestration.

4.3. Open Research Problems and Future Directions

Building from the challenge analysis in Section 4.1 and the trends in Section 4.2, the following research priorities are identified as most critical for the enhancement of AI-driven cybersecurity.

Certified Robustness at Scale: Adversarial robustness remains unresolved for transformer-scale models. Future research must develop formal certification methods that scale to large architectures and provide meaningful guarantees under realistic threat models—moving beyond empirical robustness evaluations that are invalidated by adaptive adversaries.

Privacy-Preserving Collaborative Intelligence: Federated learning frameworks for cross-organization threat intelligence sharing must be hardened against Byzantine poisoning and gradient-inversion attacks. Secure aggregation mechanisms and differential privacy budgeting for iterative FL training in production cybersecurity environments represent high-priority open problems.

Trustworthy and Auditable Autonomy: RL-based and agentic AI systems require formal policy verification, constrained decision boundaries, and standardized explainability frameworks before they can be safely deployed in critical infrastructure and military contexts. HITL architectures must be designed with cognitive ergonomics evidence to ensure that human oversight remains meaningful rather than nominal.

Adaptive and Continual Learning: Concept drift and the rapid evolution of attack tactics require AI security models that can adapt continuously without catastrophic forgetting of known threat patterns. Continual learning under distribution shift, with provable detection of model degradation, is an open challenge across all AI paradigms in Figure 2.

Governance-Aligned AI Design: Technical AI development must be co-designed with regulatory and governance requirements from the outset—not adapted post hoc. This requires interdisciplinary research combining AI engineering, cybersecurity domain expertise,

legal scholarship, and ethics to produce AI security systems that are compliant-by-design with emerging regulatory frameworks such as EU AI Act [137], NIST AI RMF [138], ISO/IEC 42001 [139].

Benchmarks for Real-World Effectiveness: The gap between benchmark-dataset performance and production operational effectiveness is a persistent and critical credibility problem for the field. Standardized evaluation frameworks that measure AI security system performance under adversarial conditions, realistic traffic distributions, and concept drift—with comparable, reproducible methodology—are needed to enable evidence-based deployment decisions.

5. Conclusions

This survey examined the role of artificial intelligence in cyber defense through the five-dimensional taxonomy in Figure 2, covering AI Techniques (Dimension 1), Security Capabilities (Dimension 2), Application Domains (Dimension 3), Deployment and Operational Challenges (Dimension 4), and Emerging Trends (Dimension 5). The critical analyses in Sections 3 and 4 reveal a field in fundamental transition—from reactive, signature-based defense toward adaptive, intelligence-driven, and increasingly autonomous security architectures—but one where operational readiness and governance preparedness consistently trail research capability.

Across the five security capabilities, no single AI paradigm is universally optimal. Supervised ML and DL deliver high detection accuracy but trade off explainability, robustness, and data efficiency, as capability matrix shown in Figure 3 makes explicit. Reinforcement learning enables adaptive autonomous defense yet carries critically low explainability—a structural liability wherever automated actions have operational or legal consequences. Federated learning addresses cross-organizational privacy constraints but remains vulnerable to Byzantine poisoning and gradient-inversion attacks. LLMs are transforming threat intelligence and incident response through semantic understanding of unstructured security data, while simultaneously introducing hallucination, prompt injection, and jailbreaking vulnerabilities at the same security functions they are intended to protect. XAI techniques bridge performance and accountability across all paradigms, but formal verification of their faithfulness remains an open theoretical problem.

Analysis of Dimension 3 confirms that consequence severity, threat model, and regulatory context differ sharply across application domains. DL-based NIDS in enterprise and critical infrastructure settings are limited by benchmark optimism bias and unsupervised detection's high false-positive rates. CPS security requires AI models that couple cyber threat detection with physical-process safety constraints. IoT deployments are bounded by device resource envelopes that restrict viable AI to lightweight ML at the endpoint. Healthcare and military contexts impose the field's most demanding governance requirements—HIPAA and GDPR in the former; international humanitarian law, meaningful human control, and the unresolved responsibility gap in the latter.

The six Dimension 4 challenges form a mutually reinforcing constraint system: data quality caps model performance, adversarial AI degrades robustness at every threat level, the privacy concerns, the explainability deficit undermines regulatory compliance, workforce limitations reduce supervisory capacity, and cost constraints widen the security equity gap between large and small organizations. Dimension 5's emerging trends describe a maturity gradient from production-ready (AI-augmented SOCs, zero-trust architecture) through transitional (human–AI collaboration, autonomous cyber defense) to frontier-stage (agentic AI), with the dual-use character of each capability intensifying the adversarial arms race documented across the entire taxonomy.

Future research must prioritize certified robustness at scale, privacy-preserving collaborative intelligence, trustworthy and auditable autonomy, adaptive continual learning, governance-aligned AI design, and standardized real-world effectiveness benchmarks. Realizing the potential of AI-driven cyber defense safely and equitably will require sustained interdisciplinary collaboration among researchers, engineers, legal scholars, and policymakers—ensuring that technological advancement remains aligned with operational accountability and the governance frameworks society requires.

Author Contributions: Conceptualization, A.M. (Amitabh Mishra) and V.V.; formal analysis, A.M. (Amitabh Mishra) and V.V.; investigation, A.M. (Amitabh Mishra), V.V., A.M. (Asmi Mishra) and S.S.; resources, A.M. (Amitabh Mishra), V.V., A.M. (Asmi Mishra) and S.S.; writing—original draft preparation A.M. (Amitabh Mishra), V.V., A.M. (Asmi Mishra) and S.S.; writing—review and editing, A.M. (Amitabh Mishra); visualization, A.M. (Amitabh Mishra) and V.V.; supervision, A.M. (Amitabh Mishra) and V.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript, the authors utilized Grammarly v9.99.0 and and Claude Haiku 4.5to assist with language refinement, grammar correction, writing improvements, and paraphrasing to suit academic writing. These tools were employed exclusively to improve the readability and clarity of the manuscript and did not contribute to the study design, data collection, analysis, interpretation of results, or scientific conclusions. The authors carefully reviewed and verified all generated content and assume full responsibility for the accuracy, integrity, and originality of the work presented. All figures and tables presented in this manuscript are original and were created by the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. ISACA. AI-Driven Cyber Threats Are the Biggest Concern for Cybersecurity Professionals Going into 2026, Finds New ISACA Research. Available online: https://www.businesswire.com/news/home/20251020612551/en/AI-Driven-Cyber-Threats-Are-the-Biggest-Concern-for-Cybersecurity-Professionals-Going-Into-2026-Finds-New-ISACA-Research (accessed on 20 October 2025).
2. Ofusori, L.; Bokaba, T.; Mhlongo, S. Artificial intelligence in cybersecurity: A comprehensive review and future direction. *Appl. Artif. Intell.* **2024**, *38*, 2439609. [CrossRef]
3. Ali, R.; Ali, A.; Iqbal, F.; Khattak, A.M.; Aleem, S. A systematic review of artificial intelligence and machine learning techniques for cyber security. In *Proceedings of the International Conference on Big Data and Security*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 584–593.
4. Mohamed, N. Current trends in AI and ML for cybersecurity: A state-of-the-art survey. *Cogent Eng.* **2023**, *10*, 2272358. [CrossRef]
5. Randieri, C.; Fiani, F.; Lubrano, K.; Napoli, C. Innovations and Future Perspectives in the Use of Artificial Intelligence for Cybersecurity: A Scoping Review. *Technologies* **2025**, *13*, 584. [CrossRef]
6. Ferrag, M.A.; Friha, O.; Maglaras, L.; Janicke, H.; Shu, L. Federated deep learning for cyber security in the internet of things: Concepts, applications, and experimental analysis. *IEEE Access* **2021**, *9*, 138509–138542. [CrossRef]
7. Kilincer, I.F.; Ertam, F.; Sengur, A. Machine learning methods for cyber security intrusion detection: Datasets and comparative study. *Comput. Netw.* **2021**, *188*, 107840. [CrossRef]
8. Lansky, J.; Ali, S.; Mohammadi, M.; Majeed, M.K.; Karim, S.H.T.; Rashidi, S.; Hosseinzadeh, M.; Rahmani, A.M. Deep learning-based intrusion detection systems: A systematic review. *IEEE Access* **2021**, *9*, 101574–101599. [CrossRef]
9. Sarker, I.H. CyberLearning: Effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks. *Internet Things* **2021**, *14*, 100393. [CrossRef]
10. ISCX-2012. Available online: https://www.unb.ca/cic/datasets/ids.html (accessed on 22 July 2026).
11. CICDDoS-17. Available online: https://www.unb.ca/cic/datasets/ids-2017.html (accessed on 22 July 2026).
12. CICIDS-18. Available online: https://www.unb.ca/cic/datasets/ids-2018.html (accessed on 22 July 2026).
13. UNSW-NB15. Available online: https://research.unsw.edu.au/projects/unsw-nb15-dataset (accessed on 22 July 2026).

14. Markevych, M.; Dawson, M. A review of enhancing intrusion detection systems for cybersecurity using artificial intelligence (ai). In *Proceedings of the International Conference Knowledge-Based Organization*; Nicolae Balcescu Land Forces Academy: Sibiu, Romania, 2023; Volume 29, pp. 30–37.
15. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; Volume 1.
16. Nguyen, T.T.; Reddi, V.J. Deep reinforcement learning for cyber security. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *34*, 3779–3795.
17. Halbouni, A.; Gunawan, T.S.; Habaebi, M.H.; Halbouni, M.; Kartiwi, M.; Ahmad, R. CNN-LSTM: Hybrid deep neural network for network intrusion detection system. *IEEE Access* **2022**, *10*, 99837–99849. [CrossRef]
18. Zainel, H.; Koçak, C. LAN intrusion detection using convolutional neural networks. *Appl. Sci.* **2022**, *12*, 6645. [CrossRef]
19. Alzahrani, A.I.; Ayadi, M.; Asiri, M.M.; Al-Rasheed, A.; Ksibi, A. Detecting the presence of malware and identifying the type of cyber attack using deep learning and VGG-16 techniques. *Electronics* **2022**, *11*, 3665. [CrossRef]
20. Nataraj, L.; Karthikeyan, S.; Jacob, G.; Manjunath, B.S. Malware images: Visualization and automatic classification. In Proceedings of the 8th International Symposium on Visualization for Cyber Security, Pittsburgh, PA, USA, 20 July 2011; pp. 1–7.
21. Wang, W.; Zhu, M.; Wang, J.; Zeng, X.; Yang, Z. End-to-end encrypted traffic classification with one-dimensional convolution neural networks. In Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 22–24 July 2017; pp. 43–48.
22. Kim, J.; Kim, J.; Thu, H.L.T.; Kim, H. Long short term memory recurrent neural network classifier for intrusion detection. In Proceedings of the 2016 International Conference on Platform Technology and Service (PlatCon), Jeju, Republic of Korea, 15–17 February 2016; pp. 1–5.
23. Yin, C.; Zhu, Y.; Fei, J.; He, X. A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access* **2017**, *5*, 21954–21961. [CrossRef]
24. Vedula, V.; Lama, P.; Boppana, R.V.; Trejo, L.A. On the detection of low-rate denial of service attacks at transport and application layers. *Electronics* **2021**, *10*, 2105. [CrossRef]
25. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
26. Mirsky, Y.; Doitshman, T.; Elovici, Y.; Shabtai, A. Kitsune: An ensemble of autoencoders for online network intrusion detection. *arXiv* **2018**, arXiv:1802.09089.
27. Jain, S. Advancing cybersecurity with artificial intelligence and machine learning: Architectures, algorithms, and future directions in threat detection and mitigation. *World J. Adv. Eng. Technol. Sci.* **2025**, *14*, 273–290. [CrossRef]
28. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
29. Sai, S.; Yashvardhan, U.; Chamola, V.; Sikdar, B. Generative AI for cyber security: Analyzing the potential of ChatGPT, DALL-E, and other models for enhancing the security space. *IEEE Access* **2024**, *12*, 53497–53516. [CrossRef]
30. Xu, H.; Wang, S.; Li, N.; Wang, K.; Zhao, Y.; Chen, K.; Yu, T.; Liu, Y.; Wang, H. Large language models for cyber security: A systematic literature review. In *ACM Transactions on Software Engineering and Methodology*; ACM: New York, NY, USA, 2024.
31. Andreoni, M.; Lunardi, W.T.; Lawton, G.; Thakkar, S. Enhancing autonomous system security and resilience with generative AI: A comprehensive survey. *IEEE Access* **2024**, *12*, 109470–109493. [CrossRef]
32. Wei, Y.; Shangguan, M. A review of deep learning-based intrusion detection systems. *Highlights Sci. Eng. Technol.* **2023**, *56*, 188–199. [CrossRef]
33. Rathod, V.; Nabavirazavi, S.; Zad, S.; Iyengar, S.S. Privacy and security challenges in large language models. In Proceedings of the 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 5–8 January 2025; pp. 00746–00752.
34. Shayegani, E.; Mamun, M.A.A.; Fu, Y.; Zaree, P.; Dong, Y.; Abu-Ghazaleh, N. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv* **2023**, arXiv:2310.10844.
35. Sha, Z.; Zhang, Y. Prompt stealing attacks against large language models. *arXiv* **2024**, arXiv:2402.12959.
36. Penmetsa, M.; Bhumireddy, J.R.; Chalasani, R.; Tyagadurgam, M.S.V.; Gangineni, V.N.; Pabbineedi, S. Next-Generation Cybersecurity: The Role of AI and Quantum Computing in Threat Detection. *Int. J. Emerg. Trends Comput. Sci. Inf. Technol.* **2021**, *2*, 54–61. [CrossRef]
37. Kalinin, M.; Krundyshev, V. Security intrusion detection using quantum machine learning techniques. *J. Comput. Virol. Hacking Tech.* **2023**, *19*, 125–136.
38. Abreu, D.; Rothenberg, C.E.; Abelém, A. Qml-ids: Quantum machine learning intrusion detection system. In Proceedings of the 2024 IEEE Symposium on Computers and Communications (ISCC), Paris, France, 26–29 June 2024; pp. 1–6.
39. Kukliansky, A.; Orescanin, M.; Bollmann, C.; Huffmire, T. Network anomaly detection using quantum neural networks on noisy quantum computers. *IEEE Trans. Quantum Eng.* **2024**, *5*, 3100611. [CrossRef]
40. Mercaldo, F.; Ciaramella, G.; Iadarola, G.; Storto, M.; Martinelli, F.; Santone, A. Towards explainable quantum machine learning for mobile malware detection and classification. *Appl. Sci.* **2022**, *12*, 12025. [CrossRef]

41. Hdaib, M.; Rajasegarar, S.; Pan, L. Quantum deep learning-based anomaly detection for enhanced network security. *Quantum Mach. Intell.* **2024**, *6*, 26. [CrossRef]
42. Buczak, A.L.; Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Commun. Surv. Tutor.* **2015**, *18*, 1153–1176.
43. Ahmad, Z.; Shahid Khan, A.; Wai Shiang, C.; Abdullah, J.; Ahmad, F. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Trans. Emerg. Telecommun. Technol.* **2021**, *32*, e4150.
44. CICDDoS-19. Available online: https://www.unb.ca/cic/datasets/ddos-2019.html (accessed on 22 July 2026).
45. Sommer, R.; Paxson, V. Outside the closed world: On using machine learning for network intrusion detection. In Proceedings of the 2010 IEEE Symposium on Security and Privacy, Oakland, FL, USA, 16–19 May 2010; pp. 305–316.
46. Kostyuk, N.; Gartzke, E. Why cyber dogs have yet to bark loudly in Russia's invasion of Ukraine. *Tex. Natl. Secur. Rev.* **2022**, *5*, 113–126. [CrossRef] [PubMed]
47. Sánchez-Zas, C.; Larriva-Novo, X.; Villagrá, V.A.; Rodrigo, M.S.; Moreno, J.I. Design and evaluation of unsupervised machine learning models for anomaly detection in streaming cybersecurity logs. *Mathematics* **2022**, *10*, 4043. [CrossRef]
48. Apache Spark. Apache Spark Documentation. 2026. Available online: https://spark.apache.org/docs/latest/ml-guide.html (accessed on 5 May 2026).
49. NSLKDD. Available online: https://www.kaggle.com/datasets/hassan06/nslkdd (accessed on 22 July 2026).
50. NERC. 2026. Available online: https://www.nerc.com/standards/reliability-standards/cip (accessed on 11 July 2026).
51. ISA/IEC. 2026. Available online: https://www.fortinet.com/resources/cyberglossary/iec-62443 (accessed on 11 July 2026).
52. Humayed, A.; Lin, J.; Li, F.; Luo, B. Cyber-physical systems security—A survey. *IEEE Internet Things J.* **2017**, *4*, 1802–1831. [CrossRef]
53. Cárdenas, A.A.; Amin, S.; Lin, Z.S.; Huang, Y.L.; Huang, C.Y.; Sastry, S. Attacks against process control systems: Risk assessment, detection, and response. In Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security, Hong Kong, China, 22–24 March 2011; pp. 355–366.
54. Cardenas, A.A.; Amin, S.; Sastry, S. Secure control: Towards survivable cyber-physical systems. In Proceedings of the 2008 the 28th International Conference on Distributed Computing Systems Workshops, Beijing, China, 17–20 June 2008; pp. 495–500.
55. Langner, R. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Secur. Priv.* **2011**, *9*, 49–51. [CrossRef]
56. SWaT. Available online: https://www.kaggle.com/datasets/vishala28/swat-dataset-secure-water-treatment-system (accessed on 22 July 2026).
57. BATADAL. Available online: https://www.batadal.net/data.html (accessed on 22 July 2026).
58. Liu, Y.; Wang, J.; Li, J.; Niu, S.; Song, H. Machine learning for the detection and identification of Internet of Things devices: A survey. *IEEE Internet Things J.* **2021**, *9*, 298–320.
59. Sánchez, P.M.S.; Celdrán, A.H.; Bovet, G.; Pérez, G.M. Adversarial attacks and defenses on ML-and hardware-based IoT device fingerprinting and identification. *Future Gener. Comput. Syst.* **2024**, *152*, 30–42. [CrossRef]
60. Chen, D.D.; Woo, M.; Brumley, D.; Egele, M. Towards automated dynamic analysis for linux-based embedded firmware. In Proceedings of the NDSS, San Diego, CA, USA, 21–24 February 2016; Volume 1, pp. 1–16.
61. Warden, P.; Situnayake, D. *Tinyml: Machine Learning with Tensorflow Lite on Arduino and Ultra-Low-Power Microcontrollers*; O'Reilly Media: Sebastopol, CA, USA, 2019.
62. Lane, N.D.; Bhattacharya, S.; Georgiev, P.; Forlivesi, C.; Jiao, L.; Qendro, L.; Kawsar, F. Deepx: A software accelerator for low-power deep learning inference on mobile devices. In Proceedings of the 2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), Vienna, Austria, 11–14 April 2016; pp. 1–12.
63. Palaniappan, K.; Duraipandi, B.; Balasubramanian, U.M. Dynamic behavioral profiling for anomaly detection in software-defined IoT networks: A machine learning approach. *Peer-to-Peer Netw. Appl.* **2024**, *17*, 2450–2469. [CrossRef]
64. Wu, H.; Han, H.; Wang, X.; Sun, S. Research on artificial intelligence enhancing internet of things security: A survey. *IEEE Access* **2020**, *8*, 153826–153848. [CrossRef]
65. Akter, S.S.; Ahmed, R.; Khan, F.H.; Rahman, M.S. Cyber-Physical Energy Systems Security: Attacks, Vulnerabilities and Risk Management. In *Cyber Security Using Modern Technologies*; CRC Press: Boca Raton, FL, USA, 2023; pp. 155–182.
66. BoTIoT. Available online: https://research.unsw.edu.au/projects/bot-iot-dataset (accessed on 22 July 2026).
67. TonIoT. Available online: https://research.unsw.edu.au/projects/toniot-datasets (accessed on 22 July 2026).
68. N-BaIoT. Available online: https://www.kaggle.com/datasets/mkashifn/nbaiot-dataset (accessed on 22 July 2026).
69. CICIoT-23. Available online: https://www.unb.ca/cic/datasets/iotdataset-2023.html (accessed on 22 July 2026).
70. Alabdulatif, A.; Khalil, I.; Saidur Rahman, M. Security of blockchain and AI-empowered smart healthcare: Application-based analysis. *Appl. Sci.* **2022**, *12*, 11039. [CrossRef]
71. Subramanian, G.; Thampy, A.S. Implementation of blockchain consortium to prioritize diabetes patients' healthcare in pandemic situations. *IEEE Access* **2021**, *9*, 162459–162475. [CrossRef]

72. Abdellatif, A.A.; Samara, L.; Mohamed, A.; Erbad, A.; Chiasserini, C.F.; Guizani, M.; O'Connor, M.D.; Laughton, J. Medge-chain: Leveraging edge computing and blockchain for efficient medical data exchange. *IEEE Internet Things J.* **2021**, *8*, 15762–15775. [CrossRef]
73. Farhan, M. Empowering healthcare: Symbiotic innovations of AI and blockchain technology. In *Blockchain and AI*; CRC Press: Boca Raton, FL, USA, 2024; pp. 23–57.
74. McGhin, T.; Choo, K.K.R.; Liu, C.Z.; He, D. Blockchain in healthcare applications: Research challenges and opportunities. *J. Netw. Comput. Appl.* **2019**, *135*, 62–75. [CrossRef]
75. Shaked, A.; Cherdantseva, Y.; Burnap, P.; Maynard, P. Operations-informed incident response playbooks. *Comput. Secur.* **2023**, *134*, 103454. [CrossRef]
76. Tariq, S.; Baruwal Chhetri, M.; Nepal, S.; Paris, C. Alert fatigue in security operations centres: Research challenges and opportunities. *ACM Comput. Surv.* **2025**, *57*, 1–38. [CrossRef]
77. Aramide, O.O. AI-driven automated incident response and remediation in networks. *Int. J. Technol. Manag. Humanit.* **2025**, *11*, 1–9. [CrossRef]
78. Waelchli, S.; Walter, Y. Reducing the risk of social engineering attacks using SOAR measures in a real world environment: A case study. *Comput. Secur.* **2025**, *148*, 104137. [CrossRef]
79. Kinyua, J.; Awuah, L. AI/ML in Security Orchestration, Automation and Response: Future Research Directions. *Intell. Autom. Soft Comput.* **2021**, *28*, 527–545. [CrossRef]
80. IBM. Explainable AI. 2026. Available online: https://www.ibm.com/think/topics/explainable-ai (accessed on 27 April 2026).
81. Ferrara, E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci* **2024**, *6*, 3. [CrossRef]
82. Alliouche, R.; Chenni, H. Toward Intelligent Cyber Defense: A Comprehensive Study of SOAR Technologies for AI-Based DDoS Detection. Available online: https://www.researchgate.net/publication/394967622_Toward_Intelligent_Cyber_Defense_A_Comprehensive_Study_of_SOAR_Technologies_for_AI-Based_DDoS_Detection (accessed on 30 December 2025).
83. Matthias, A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* **2004**, *6*, 175–183. [CrossRef]
84. ISA/IEC. 2026. Available online: https://www.nis-2-directive.com (accessed on 11 July 2026).
85. Sun, N.; Ding, M.; Jiang, J.; Xu, W.; Mo, X.; Tai, Y.; Zhang, J. Cyber threat intelligence mining for proactive cybersecurity defense: A survey and new perspectives. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 1748–1774. [CrossRef]
86. Santos, P.; Abreu, R.; Reis, M.J.; Serôdio, C.; Branco, F. A systematic review of cyber threat intelligence: The effectiveness of technologies, strategies, and collaborations in combating modern threats. *Sensors* **2025**, *25*, 4272. [CrossRef] [PubMed]
87. Balasubramanian, P.; Liyana, S.; Sankaran, H.; Sivaramakrishnan, S.; Pusuluri, S.; Pirttikangas, S.; Peltonen, E. Generative AI for cyber threat intelligence: Applications, challenges, and analysis of real-world case studies. *Artif. Intell. Rev.* **2025**, *58*, 336. [CrossRef]
88. Roy, K. *A Global History of Warfare and Technology: From Slings to Robots*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 163–173.
89. Zweibelson, B. *Understanding the Military Design Movement: War, Change and Innovation*; Routledge: Abingdon, UK, 2023; pp. 63–72.
90. Najžer, B. The Hybrid Age. 2020. Available online: https://api.pageplace.de/preview/DT0400.9780755602537_A39699286/preview-9780755602537_A39699286.pdf (accessed on 28 December 2025).
91. Chen, J.Q.; Dinerman, A. Cyber capabilities in modern warfare. In *Cyber Security: Power and Technology*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 21–30.
92. Demchak, C.C. Cybered conflict, hybrid war, and informatization wars. In *Routledge Handbook of International Cybersecurity*; Routledge: Abingdon, UK, 2020; pp. 36–51.
93. Singh, A.; Gupta, S.; Jain, M. Adaptation of modern technologies and challenges in the defense sectors. *Res Mil.* **2022**, *12*, 1547–1556.
94. Werkhoven, P.; Kester, L.; Neerincx, M. Telling autonomous systems what to do. In Proceedings of the 36th European Conference on Cognitive Ergonomics, Utrecht, The Netherlands, 5–7 September 2018; pp. 1–8.
95. Zurek, T.; Kwik, J.; Van Engers, T. Model of a military autonomous device following International Humanitarian Law. *Ethics Inf. Technol.* **2023**, *25*, 15. [CrossRef]
96. Abaimov, S.; Martellini, M. Artificial intelligence in autonomous weapon systems. In *21st Century Prometheus: Managing CBRN Safety and Security Affected by Cutting-Edge Technologies*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 141–177.
97. Konert, A.; Balcerzak, T. Military autonomous drones (UAVs)-from fantasy to reality. Legal and Ethical implications. *Transp. Res. Procedia* **2021**, *59*, 292–299. [CrossRef]
98. Munir, M. Autonomous Weapons Systems: Taking the Human Out of the Loop. 2022. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4074072# (accessed on 28 December 2025).

99. Fisher, B.A. How International Humanitarian Law Will Constrain the Use of Autonomous Weapon Systems in the Conduct of Hostilities. Ph.D. Thesis, Murdoch University, Perth, Australia, 2022.
100. Watts, T.F.; Bode, I. *Automation and Autonomy in Loitering Munitions Catalogue (v. 1)*; Zenodo: Geneva, Switzerland, 2023. [CrossRef]
101. Bianconi, G.; Arenas, A.; Biamonte, J.; Carr, L.D.; Kahng, B.; Kertesz, J.; Kurths, J.; Lü, L.; Masoller, C.; Motter, A.E.; et al. Complex systems in the spotlight: Next steps after the 2021 Nobel Prize in Physics. *J. Phys. Complex.* **2023**, *4*, 010201. [CrossRef]
102. Dresp-Langley, B. The weaponization of artificial intelligence: What the public needs to be aware of. *Front. Artif. Intell.* **2023**, *6*, 1154184. [CrossRef] [PubMed]
103. Klare, M.T. *Assessing the Dangers: Emerging Military Technologies and Nuclear (In) Stability*; Arms Control Association: Washington, DC, USA, 2023.
104. Armitage, R. We must oppose lethal autonomous weapons systems. *Br. J. Gen. Pract.* **2019**, *69*, 510. [CrossRef] [PubMed]
105. Lohn, A.; Knack, A.; Burke, A.; Jackson, K. *Autonomous Cyber Defence: A Roadmap from Lab to Ops*; Center for Emerging Technology and Security: Washington, DC, USA, 2023.
106. Mareedu, A. Autonomous Security Operations Centers (SOC): AI Agents for Threat Triage, Response, and Orchestration. *Int. J. Emerg. Res. Eng. Technol.* **2025**, *6*, 63–70. [CrossRef]
107. Santoni de Sio, F.; Van den Hoven, J. Meaningful human control over autonomous systems: A philosophical account. *Front. Robot. AI* **2018**, *5*, 323836. [CrossRef] [PubMed]
108. EMBER. Available online: https://www.kaggle.com/datasets/trinhvanquynh/ember-for-static-malware-analysis (accessed on 22 July 2026).
109. SOREL-20M. Available online: https://github.com/sophos/SOREL-20M (accessed on 22 July 2026).
110. MSMalChallenge. Available online: https://www.kaggle.com/competitions/malware-classification/data (accessed on 22 July 2026).
111. Malimg. Available online: https://www.kaggle.com/datasets/manmandes/malimg (accessed on 22 July 2026).
112. Malnet. Available online: http://malnet.cc.gatech.edu/image-data/ (accessed on 22 July 2026).
113. VirusShare. Available online: https://virusshare.com (accessed on 22 July 2026).
114. CICMalMem22. Available online: https://www.unb.ca/cic/datasets/malmem-2022.html (accessed on 22 July 2026).
115. APICallSeq. Available online: https://www.kaggle.com/datasets/ang3loliveira/malware-analysis-datasets-api-call-sequences (accessed on 22 July 2026).
116. KDDCUP. Available online: https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (accessed on 22 July 2026).
117. Kyoto2006+. Available online: https://www.kaggle.com/datasets/harshwardhanbhangale/kyoto-2006 (accessed on 22 July 2026).
118. Hikari2021. Available online: https://zenodo.org/records/5199540 (accessed on 22 July 2026).
119. IoT-23. Available online: https://www.stratosphereips.org/datasets-iot23 (accessed on 22 July 2026).
120. mqtt-2020. Available online: https://ieee-dataport.org/open-access/mqtt-iot-ids2020-mqtt-internet-things-intrusion-detection-dataset (accessed on 22 July 2026).
121. EdgeIoT. Available online: https://www.kaggle.com/datasets/mohamedamineferrag/edgeiiotset-cyber-security-dataset-of-iot-iiot?select=Edge-IIoTset+dataset (accessed on 22 July 2026).
122. Mittelstadt, B.D.; Allo, P.; Taddeo, M.; Wachter, S.; Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **2016**, *3*, 2053951716679679. [CrossRef]
123. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
124. Radanliev, P. AI ethics: Integrating transparency, fairness, and privacy in AI development. *Appl. Artif. Intell.* **2025**, *39*, 2463722. [CrossRef]
125. Cadet, E.; Etim, E.; Essien, I.; Ajayi, J.; Erigha, E. Ethical challenges in AI-driven cybersecurity decision-making. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2024**, *10*, 1031–1064. [CrossRef]
126. Humphreys, D.; Koay, A.; Desmond, D.; Mealy, E. AI hype as a cyber security risk: The moral responsibility of implementing generative AI in business. *AI Ethics* **2024**, *4*, 791–804. [CrossRef]
127. Saurabh, B.; Utkrisht, S.; Sandeep, S.; Kumar, D.; Rajkumar, U. Generative AI Enabled Actionable Decision Support in Cyber Security Operations for Enterprise Security. In Proceedings of the 2024 ITU Kaleidoscope: Innovation and Digital Transformation for a Sustainable World (ITU K), New Delhi, India, 21–23 October 2024; pp. 1–8.
128. Rajgopal, P.R. SOC Talent Multiplication: AI Copilots as Force Multipliers in Short-Staffed Teams. *Int. J. Comput. Appl.* **2025**, *187*, 46–62. [CrossRef]

129. Akre, V.; Kobbaey, T.; Lazarov, G.; Abdulsalam, K.; Al-Sit, W.; Diab, J. From Alert Fatigue to Augmented Defense: A Case for AI Copilots in Cybersecurity Operation Centers. In Proceedings of the 2025 10th International Conference on Information Technology Trends (ITT), Dubai, United Arab Emirates, 6–7 November 2025; pp. 282–287.
130. Karunasingha, N.; Chhetri, M.B.; Nepal, S.; Paris, C.; Kanhere, S.S. SoK: AI Support for Analyst Situation Awareness in Security Operation Centres. In Proceedings of the 2025 European Symposium on Usable Security (EuroUSEC), Manchester, UK, 10–11 September 2025; pp. 151–163.
131. Loevenich, J.F.; Adler, E.; Bécue, A.; Velazquez, A.; Wrona, K.; Boshnakov, V.; Falkcrona, J.; Nordbotten, N.; Worthington, O.L.; Röning, J.; et al. Training autonomous cyber defense agents: Challenges & opportunities in military networks. In Proceedings of the MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM), Washington, DC, USA, 28 October–1 November 2024; pp. 158–163.
132. Pokhrel, S.R.; Yang, L.; Rajasegarar, S.; Li, G. Robust zero trust architecture: Joint blockchain-based federated learning and anomaly detection-based framework. In Proceedings of the SIGCOMM Workshop on Zero Trust Architecture for Next Generation Communications, Sydney, Australia, 4–8 August 2024; pp. 7–12.
133. Ramezanpour, K.; Jagannath, J. Intelligent zero trust architecture for 5G/6G networks: Principles, challenges, and the role of machine learning in the context of O-RAN. *Comput. Netw.* **2022**, *217*, 109358. [CrossRef]
134. Hussain, M.; Pal, S.; Jadidi, Z.; Foo, E.; Kanhere, S. Federated zero trust architecture using artificial intelligence. *IEEE Wirel. Commun.* **2024**, *31*, 30–35. [CrossRef]
135. Javeed, D.; Saeed, M.S.; Adil, M.; Kumar, P.; Jolfaei, A. A federated learning-based zero trust intrusion detection system for Internet of Things. *Ad Hoc Netw.* **2024**, *162*, 103540. [CrossRef]
136. Ajish, D. The significance of artificial intelligence in zero trust technologies: A comprehensive review. *J. Electr. Syst. Inf. Technol.* **2024**, *11*, 30. [CrossRef]
137. EU Artificial Intelligence ACT. 2024. Available online: https://artificialintelligenceact.eu/high-level-summary/ (accessed on 11 July 2026).
138. NIST AI Risk Management Framework. 2023. Available online: https://www.nist.gov/itl/ai-risk-management-framework (accessed on 11 July 2026).
139. *ISO/IEC 42001*; Information Technology-Artificial Intelligence-Management System. ISO: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/42001 (accessed on 11 July 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.