International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 02 | Feb 2026 www.irjet.net p-ISSN: 2395-0072

Cyber security in the Age of Artificial Intelligence: Threat Intelligence, Automated Attacks, and Defensive AI Systems

Aksharbhai Miyani[1], Hardik Nakum[2], Ankit Vegad[3], Prof. Lata H. Butiya[4]

[1]Dept. of Computer Application Gyanmanjari Innovative University
[2]Dept. of Computer Application Gyanmanjari Innovative University Bhavnagar, Gujarat, India
[3]Dept. of Computer Application Gyanmanjari Innovative University
[4](Guide) Assistant Professor Dept. of Computer Application, GMIU Bhavnagar, Gujarat, India

---

Abstract – *The accelerating convergence of artificial intelligence (AI) and cyber security has fundamentally altered the threat landscape, creating both unprecedented defensive capabilities and novel attack vectors that challenge decades of established security assumptions. This review synthesizes research across machine learning–driven in- trusion detection, AI-powered offensive tooling, adversarial machine learning, and automated threat intelligence to map the current state of knowledge at this critical intersection. Drawing on over seventy peer-reviewed studies, framework documentation from MITRE ATT&CK and ATLAS, and threat-intelligence reports from ENISA, NIST, Crowd Strike, and Mandiant, we trace how deep learning supplanted signature-based detection, how generative models now craft polymorphic malware and hyper-personalized phishing at scale, and how reinforcement-learning agents autonomously probe network perimeters. We identify a persistent asymmetry: offensive applications of AI mature faster than their defensive counterparts, partly because adversaries face no regulatory or ethical constraints on model deployment. The review further examines adversarial machine learning—model poisoning, evasion, and extraction—as a meta-threat that undermines the very AI systems designed to protect digital infrastructure. Through comparative analysis of defensive architectures including AI-augmented Security Orchestration, Automation, and Response (SOAR) pipelines, Extended Detection and Response (XDR) platforms, and behavioral analytics engines, we evaluate the practical readiness of each approach. Real-world case studies of deep fake-enabled fraud, LLM- generated exploit code, and autonomous vulnerability discovery ground the discussion in operational reality. We conclude by charting critical research gaps—explain ability in security models, cross-domain transfer of adversarial robustness, and the governance vacuum surrounding dual-use AI—and propose a structured agenda for future investigation. This review is intended to serve as a unified reference for researchers, practitioners, and policymak- ers navigating the rapidly shifting terrain where artificial intelligence meets cyber security.*

Keywords – Artificial Intelligence, Cyber security, Threat Intelligence, Adversarial Machine Learning, Automated Attacks, Intrusion Detection, Deep fakes, Large Language Models, Cyber Defense, AI Governance

I. INTRODUCTION

Few developments in the history of information secu- rity have compressed timelines as violently as the main- streaming of large-scale artificial intelligence. Between 2020 and 2024, the cyber security community witnessed a qualitative shift: threat actors moved from experimenting with machine learning models to operationalizing them at scale [1, 2]. The 2023 Verizon Data Breach Investigations Report noted that social-engineering attacks—the category most amenable to AI-driven personalization—accounted for a record share of initial access vectors [3]. Concur- rental, defenders began deploying neural-network–based intrusion detection, behavioral analytics, and automated incident-response pipelines, creating an arms race whose dynamics are still poorly understood [4].

This convergence is not accidental. The same capabil- ities that make modern AI powerful—pattern recognition over massive corpora, language generation indistinguish- able from human output, policy optimization in adversarial environments—map directly onto both attack and defense requirements. A spear-phishing engine and a phishing-detection classifier share the same underlying transformer architecture; the difference lies in the objective function and the ethical frame of the operator [5, 6]. Recognizing this symmetry is essential to any serious analysis of the field.

1.1 The Weap onization of AI

The notion that AI could be weaponized was anticipated well before the current wave. Brundage et al. warned in 2018 that advances in AI would "expand the set of actors who are capable of carrying out [cyber] attacks, the rate at which these actors can carry them out, and the set of plau- sible targets" [7]. That forecast has been largely validated. Seymour and Tully demonstrated automated spear phishing on social media as early as 2016 [8]; by 2023, researchers showed that GPT-class

models could generate contextually convincing phishing emails at a fraction of the cost of hu- man operators [5]. The implications extend well beyond phishing. Pa Pa et al. examined ChatGPT's capacity to produce functional malware components, finding that while safety filters block direct requests, prompt-injection and in- direct methods frequently succeed [9]. Greshake et al. fur- ther demonstrated that LLM-integrated applications can be weaponized through indirect prompt injection, blurring the boundary between the model and the attack surface [10].

1.2 Automation of Cybercrime

Criminal ecosystems have long embraced automation— botnets, exploit kits, and phishing-as-a-service platforms predate the deep learning era—but AI introduces a quali- tative leap. Machine learning allows attack tools to adapt in real time: polymorphic malware that rewrites its sig- nature before each propagation attempt, credential-stuffing engines that learn login-form structures across heteroge- neous websites, and ransomware that selects encryption tar- gets based on inferred file value [11, 12]. Reinforcement- learning agents have been shown to discover viable pene- tration paths in simulated enterprise networks without hu- man guidance, reducing the expertise barrier for would-be attackers [13, 14].

The economic calculus is equally alarming. Genera- tive AI lowers the marginal cost of producing high-quality attack artifacts—emails, exploit code, fraudulent voice clips—toward zero, while the cost of defending against each new variant remains substantial [15]. This asymme- try suggests that the current period may represent a struc- tural shift in attacker–defender economics, not merely an incremental escalation.

1.3 The Imperative for AI-Driven Defense

If AI expands the attacker's toolkit, it also offers the most credible path to scalable defense. Signature-based detec- tion, the dominant paradigm for three decades, cannot keep pace with adversaries who generate novel artifacts per en-gagement [16]. Machine learning models trained on net- work telemetry, endpoint behavior, and log semantics have demonstrated superior detection of zero-day threats, lateral movement, and insider anomalies [17–19]. Deep learn- ing architectures—particularly recurrent and transformer- based models—have shown promise in processing sequen- tial security data at scale [20, 21].

Yet defensive AI introduces its own vulnerabilities. Ad- versarial machine learning research has shown that classi- fiers can be evaded with carefully crafted perturbations [22, 23], that training data can be poisoned to introduce back- doors [24, 25], and that model parameters can be extracted through query-based attacks [26]. The defender who re- lies on AI without understanding these failure modes risks a false sense of security—a concern that Taddeo et al. have called the "double-edged sword" of trusting AI in cyberse- curity [27].

1.4 Scope and Contributions of This Review

This paper offers a comprehensive, literature-driven re- view of the intersection between artificial intelligence and cybersecurity. Unlike narrowly scoped surveys that address individual sub-problems (e.g., ML-based intrusion detec- tion alone), we adopt a holistic perspective that encom- passes offensive AI, defensive AI, adversarial robustness, threat intelligence automation, and the governance chal- lenges that bind them.

The principal contributions of this review are as follows:

- **C1.** A systematic synthesis of over seventy peer-reviewed studies, framework documents, and industry reports spanning 2014–2024, organized around a unified tax- onomy of AI roles in cyber security.
- **C2.** A comparative analysis of AI-driven defensive architectures—IDS/IPS, behavioral analytics, SOAR, and XDR—grounded in empirical findings and opera- tional deployment evidence.
- **C3.** An in-depth examination of adversarial machine learn- ing as a *meta-threat* that jeopardizes the integrity of AI-based defenses themselves.
- **C4.** A critical evaluation of AI-enabled offensive capabili- ties, including automated exploit generation, deepfake social engineering, and LLM-assisted malware cre- ation, with real-world case studies.
- **C5.** Identification of key research gaps, contradictions in existing literature, and a structured agenda for future investigation.

The remainder of this paper is organized as follows. Sec- tion II establishes foundational concepts at the intersec- tion of AI and cyber security. Section III surveys the AI- augmented threat landscape. Section IV examines auto- mated attack methodologies. Section V reviews AI-driven defensive systems. Section VI discusses AI-enabled threat intelligence. Section VII treats adversarial machine learn-ing. Section VIII presents illustrative case studies. Sec- tion IX addresses ethics, policy, and governance. Section X projects future trends. Section XI maps research gaps, and Section XII closes with concluding remarks.

II. FOUNDATIONS OF AI IN CYBER SECURITY

Before examining how AI reshapes both offense and defense, it is worth revisiting the intellectual lineage that brought the field to its present state. The application of ma- chine learning to security problems is older than the current hype cycle suggests; what has changed is the scale of data available, the expressiveness of models, and the adversarial sophistication of the operating environment.

2.1 Machine Learning in Security: A Historical Arc

Early applications of ML in cyber security were over- whelmingly focused on intrusion detection. Buczak and Guven's 2016 survey documented decades of work ap- plying decision trees, random forests, support vector ma- chines, and naïve Bayes classifiers to network traffic clas- sification [17]. These methods worked well in controlled laboratory settings but struggled with concept drift, class imbalance, and the high false-positive rates that made oper- ational deployment impractical [16]. Sommer and Paxson's influential 2010 critique argued that the "closed-world" as- sumptions underlying most ML security research—fixed feature spaces, stationary distributions, balanced datasets— bore little resemblance to production environments [16]. Their work remains a useful caution against uncritical en- thusiasm.

By the mid-2010s, deep learning architectures began to address some of these limitations. Convolutional neural networks (CNNs) proved effective at extracting spatial fea- tures from raw byte sequences and system-call traces, while recurrent architectures (LSTMs, GRUs) captured temporal dependencies in network flows [18, 28]. The key advan- tage was representation learning: deep models could dis- cover relevant features without hand-engineering, reducing the domain expertise required for deployment [20].

2.2 Deep Learning vs. Classical ML: A Nuanced Com- parison

The narrative that deep learning universally outperforms classical methods oversimplifies a more textured reality. Apruzzese et al.'s 2023 evaluation found that gradient- boosted ensembles matched or exceeded deep models on several benchmark intrusion-detection tasks, particularly when training data were limited [29]. Shaukat et al. reached a similar conclusion in their 2020 survey, noting that lo- gistic regression and random forests remained competitive for binary classification tasks where interpretability wasvalued [19]. Deep learning's advantages tend to emerge in three specific conditions: (1) large and heterogeneous datasets, (2) tasks requiring raw-input processing without feature engineering, and (3) scenarios where sequential or spatial structure in the data carries discriminative informa- tion [21, 30].

The trade-off landscape is further complicated by adver- sarial robustness. Simple linear models, while less accurate on clean data, can be more resistant to gradient-based eva- sion attacks because they present a simpler loss surface for the attacker to manipulate [22]. Deep networks, conversely, are notoriously vulnerable to small perturbations that are imperceptible to human analysts but cause confident mis- classification [23]. This creates a paradox for security ar- chitects: the most powerful detection models may also be the most fragile in adversarial settings.

2.3 Threat Intelligence Pipelines

Modern threat intelligence (TI) extends well beyond raw indicators of compromise (IOCs). Tounsi and Rais's 2018 survey distinguished three layers of intelligence— strategic, operational, and tactical—each requiring differ- ent analytical methods and serving different decision mak- ers [31]. Machine learning enters the TI pipeline at multiple points: natural- language processing (NLP) extracts struc- tured threat data from unstructured reports and dark-web forums; clustering algorithms group related IOCs into cam- paigns; and predictive models forecast likely next targets based on historical attack patterns [32, 33].

Wagner et al. examined the sharing dimension, finding that while platforms such as STIX/TAXII have standard- ized data exchange, the quality and timeliness of shared in- telligence varies widely across sectors [34]. The integration of AI into sharing platforms raises additional concerns: au- tomated enrichment can amplify errors if upstream data are poisoned, and model-generated IOCs may carry biases in- herited from training corpora.

2.4 MITRE ATT&CK and ATLAS: Structuring the Knowledge Base

Two frameworks have become indispensable for organiz- ing knowledge about adversary behavior in AI-influenced environments. MITRE ATT&CK provides a compre- hensive taxonomy of tactics, techniques, and procedures (TTPs) based on observed real-world intrusions [35]. Its structured format enables automated mapping of detection rules to adversary behaviors, a capability exploited by most modern SIEM and XDR platforms [36].

MITRE ATLAS (Adversarial Threat Landscape for AI Systems) extends this paradigm to AI-specific threats [37]. ATLAS documents techniques such as model evasion, data poisoning, model inversion, and supply-chain compromise of ML pipelines, filling a gap that ATT&CK—designed for conventional IT infrastructure—does not cover. The exis-tence of ATLAS reflects a growing recognition that AI sys- tems are not merely tools for security but are themselves attack surfaces requiring dedicated threat modeling. The interplay between ATT&CK and ATLAS is instructive: an attacker might use ATT&CK techniques to gain initial net- work access and then pivot to ATLAS techniques to com- promise the ML-based detection system that guards the en- vironment.

2.5 The Emerging Role of Large Language Models

The release of GPT-3 in 2020 and its successors marked a discontinuity in the AI-security relationship. Large lan- guage models (LLMs) possess capabilities—code genera- tion, natural-language reasoning, multi-step planning—that map onto both offensive and defensive workflows with min- imal adaptation [38]. Xu et al.'s 2024 systematic review catalogued LLM applications ranging from automated vul- nerability analysis and patch generation to phishing-email synthesis and social-engineering chatbots [39]. Yao et al. surveyed the security and privacy risks intrinsic to LLMs themselves, including prompt injection, training-data mem- orization, and model alignment failures [40].

The dual-use character of LLMs intensifies the gover- nance challenges discussed in Section IX. Unlike narrow ML models that require domain-specific training, a single foundation model can serve offensive, defensive, and neu- tral applications depending solely on the prompt. This ver- satility complicates regulatory efforts that attempt to con- trol AI misuse at the model level rather than the application level.

III. THE AI-AUGMENTED THREAT LANDSCAPE

The incorporation of artificial intelligence into the ad- versary's arsenal has not merely expanded the volume of attacks—it has changed their *character*. Historically, cy- ber threats scaled through automation of repetitive tasks: botnets replaying credential lists, exploit kits scanning for known vulnerabilities. AI introduces adaptivity, en- abling attacks that learn from defensive responses and ad- just in flight. This section surveys six domains where AI- augmented threats have materialized or are credibly antici- pated.

3.1 AI-Powered Phishing and Social Engineering

Phishing remains the single most common initial access vector, accounting for roughly 15% of breaches in the 2023 Verizon DBIR [3]. Traditional phishing relied on mass dis- tribution and statistical luck; AI dramatically shifts this cal- culus. Large language models can generate contextually tailored emails that reference the target's recent publica- tions, organizational role, or social-media activity, achiev- ing click-through rates that surpass even those of expert hu- man social engineers [5].Seymour and Tully's 2016 experiment on automated Twitter-based spear phishing was among the first empirical demonstrations [8]. Their system combined social-network reconnaissance with NLP-generated lure messages, achiev- ing a click-through rate exceeding 30%. The intervening years have only amplified the threat. Alkhalil et al. docu- mented the evolution from template-based phishing to AI- personalized campaigns that adapt their language to the victim's communication style, making them substantially harder for both humans and traditional filters to detect [44]. Basit et al. complemented this offensive view with a sur- vey of AI-enabled phishing detection techniques, noting an ironic arms race: the same NLP models used to generate convincing phishing are now deployed to detect it [6].

3.2 Deep fake-Enabled Social Engineering

Deepfakes represent perhaps the most viscerally alarm- ing application of generative AI to cybersecurity. Tolosana et al.'s 2020 survey catalogued the rapid maturation of face- swapping and voice-cloning technologies, driven primarily by generative adversarial networks (GANs) [45]. The se- curity implications are twofold. First, attackers can im- personate trusted individuals in video or audio commu- nications, bypassing authentication schemes that rely on voice-print or visual verification. Second, the mere *ex- istence* of deepfake technology erodes trust in legitimate communications—a phenomenon Westerlund termed the "liar's dividend" [46].

Mirsky and Lee's comprehensive survey traced the deep- fake creation and detection ecosystem, noting that detec- tion methods consistently lag behind generation methods by twelve to eighteen months [41]. The asymmetry is struc- tural: generators need only produce outputs indistinguish- able from real data in a single modality, while detectors must identify artifacts across all possible generation meth- ods. Industry reports from CrowdStrike and Mandiant have documented operational use of deepfake audio in business email compromise (BEC) schemes, with losses in individ- ual incidents exceeding $25 million [2, 47].

3.3 Malware Generation via Large Language Models

The capacity of LLMs to generate functional code ex- tends naturally to malicious payloads. Pa Pa et al.'s 2023 study systematically tested ChatGPT's ability to pro- duce malware components—keyloggers, reverse shells, file encryptors—finding that while direct requests were usu- ally blocked by safety filters, task decomposition and role- playing prompts circumvented protections in the majority of cases [9]. This finding exposes a fundamental limitation of alignment-based safety: filters optimized for natural- language instruction-following can be bypassed by refor- mulating malicious intent as benign sub-tasks.Yamin et al. examined the broader weaponization land- scape, classifying AI-generated threats into intelligence- gathering, payload-generation, and delivery-optimization

Table 1: Comparative Analysis of AI-Powered Cyber Threats

Threat Category	AI Technique	Primary Vector	Target	Maturity	Key Reference
AI-Powered Phishing	LLM text generation; NLP	Email, SMS, social media	Human users	High	Hazell [5]
Deepfake Social Eng.	GAN; voice cloning	Video/audio calls; BEC	Executives, finance	Medium	Mirsky [41]
LLM Malware Generation	Code-gen LLMs	Polymorphic payloads	Endpoints, servers	Medium	Pa Pa [9]
AI Vulnerability Discovery	DL code analysis; fuzzing	Zero-day exploits	Software, firmware	Medium	Li [42]
Autonomous Botnets	RL; GAN traffic disguise	DDoS, lateral movement	Network infra	Low	Truong [4]
AI Influence Operations	LLM content gen.; bots	Social media, news	Public opinion	High	ENISA [43]

categories [15]. Their analysis highlights that the barrier to entry for malware development has been substantially lowered: individuals without traditional programming ex- pertise can now produce functional exploits by iteratively prompting an LLM, effectively democratizing capabilities that were previously restricted to skilled operators.

3.4 AI-Assisted Vulnerability Discovery

Automated vulnerability discovery is not new—fuzzers like AFL have been staples of security testing for years— but AI augmentation introduces qualitative improvements. Deep-learning–based approaches such as VulDeePecker analyze source code semantics to predict vulnerability loca- tions with precision that exceeds pattern-matching heuris- tics [42]. Shoshitaishvili et al. surveyed the state of bi- nary analysis techniques, documenting the increasing role of symbolic execution guided by learned heuristics [48]. The offensive implication is straightforward: if ML mod- els can identify vulnerabilities more efficiently than hu- man analysts, attackers armed with similar models can discover zero-days faster than vendors can patch them. The DARPA Cyber Grand Challenge (2016) demonstrated that fully autonomous systems could discover, exploit, and patch vulnerabilities in real time, establishing a proof-of- concept for AI-driven cyber autonomy. The dual-use char- acter of vulnerability-discovery AI—essential for defensive hardening yet equally useful for offensive exploitation— exemplifies the governance challenges discussed in Sec- tion IX.

3.5 Autonomous Botnets and Adaptive Malware

Classical botnets operate on command-and-control (C2) architectures that create single points of failure. AI- augmented botnets could, hypothetically, operate with decentralized decision-making: each node running a lightweight reinforcement-learning agent that selects prop- agation strategies, evasion techniques, and C2 fallback channels based on the local network environment [4]. While fully autonomous AI botnets remain largely the- oretical, the building blocks—GAN-based network traf- fic disguise [49, 50], RL-driven lateral movement strate- gies [51]—have been individually demonstrated in research settings.

Gibert et al.'s survey on machine learning for malwareclassification implicitly highlights the offensive flip side: if classifiers learn to distinguish malware families by be- havioral signatures, adversaries can train GANs to gener- ate samples that mimic benign behavior distributions while retaining malicious functionality [52]. This adversarial co-evolution—attack models informing defense models, and vice versa—is a defining feature of the current threat land- scape.

3.6 AI in Influence Operations and Information War- fare

Beyond technical infrastructure attacks, AI enables so- phisticated information operations. Generative models can produce propaganda, disinformation, and fake personas at scale, blurring the line between cybersecurity and infor- mation security. The ENISA 2023 Threat Landscape re- port identified AI-generated disinformation as a top-ten threat, particularly in the context of election interference and geopolitical conflict [43]. Microsoft's Digital Defense Report 2023 similarly documented state-sponsored actors using AI-generated content to amplify influence campaigns across social-media platforms [53]. These operations do not directly compromise technical systems but erode the trust infrastructure on which digital societies depend—a form of damage that is harder to detect and harder to re- mediate than a traditional breach.

IV. AUTOMATED ATTACK METHODOLOGIES

While the previous section mapped the threat landscape, this section examines the technical mechanisms by which AI enables attack automation. The distinction matters: the landscape describes *what* is being attacked; this section ad- dresses *how* AI operationalizes those attacks at each stage of the kill chain.

4.1 AI-Driven Exploit Generation

Traditional exploit development requires deep under- standing of target architectures, memory layouts, and ex- ecution flows—skills concentrated among a small pool of specialists. AI threatens to commoditize this ex- pertise. Neural-network–based approaches to automatic exploit generation (AEG) learn patterns from historical vulnerability–exploit pairs and can synthesize novel ex-ploits for previously unseen vulnerabilities [54]. Li et al.'s VulDeePecker system demonstrated that deep learning models could not only detect vulnerabilities but also char- acterize them with sufficient granularity to guide exploit construction [42].

The integration of LLMs into exploit-development work- flows adds another dimension. A skilled attacker can use an LLM as a code-generation assistant, producing shell- code, ROP chains, or format-string payloads through iter- ative prompting [39]. Unlike classical AEG tools that re- quire formal specifications, LLMs can work from natural- language descriptions of the target vulnerability, lower- ing the expertise barrier further. Experimental evidence suggests that LLM-assisted exploit development reduces time-to-exploit by 40–60% for known vulnerability classes, though success rates drop significantly for novel or com- plex vulnerability types [38].

4.2 Automated Penetration Testing

Penetration testing—systematically probing networks for exploitable weaknesses—is a natural candidate for AI automation. The task can be formulated as a sequential de- cision problem: given a partially observed network topol- ogy, select the next action (scan, exploit, pivot) that maxi- mizes the probability of reaching a target asset. This formu- lation maps directly onto reinforcement learning (RL) [14].

Ghanem and Chen trained RL agents on simulated enter- prise networks and demonstrated that they could discover multi-step attack paths that matched or exceeded those identified by human penetration testers [13]. Schwartz and Kurniawati extended this work using partially ob- servable Markov decision processes (POMDPs) to model the attacker's incomplete knowledge of the network topol- ogy [14]. Nguyen and Reddi's survey documented the broader application of deep RL to cyber security, noting that while automated penetration testing is the most mature ap- plication, challenges remain in transferring policies trained in simulation to real-world environments [51].

The defensive corollary is equally important: the same RL-based penetration testing tools can be used by red teams and security auditors to stress-test defenses before real ad- versaries do. McKinnel et al. conducted a meta-analysis of AI in penetration testing and vulnerability assessment, finding that autonomous tools discovered 15–20% more vulnerabilities than manual testing alone in controlled ex- periments, though they also produced higher false-positive rates [54].

4.3 Credential Harvesting Automation

Credential theft has evolved from brute-force dictio- nary attacks to AI-optimized campaigns. Machine learn- ing models trained on leaked password corpora can gener- ate statistically likely passwords for specific demograph- ics, languages, or organizations, dramatically improvingthe efficiency of offline cracking [55]. Combined with AI- generated phishing pages that dynamically mimic target lo- gin portals, the entire credential-harvesting pipeline can op- erate with minimal human oversight.

The CrowdStrike 2024 Global Threat Report docu- mented a 75% year-over-year increase in identity-based at- tacks, attributing much of the growth to automation of cre- dential harvesting and replay [2]. Mandiant's M-Trends 2023 report identified compromised credentials as the lead- ing initial access vector in incidents they investigated, sur- passing exploitation of public-facing applications for the first time [47].

4.4 AI-Enhanced Ransomware

Ransomware has undergone a remarkable evolution from opportunistic screen lockers to targeted, data-exfiltrating operations. AI augmentation adds several capabilities: au- tomated target selection based on victim financials, intelli- gent file prioritization that encrypts high-value assets first to maximize leverage, and adaptive command-and-control that evades network-level detection [11].

Razaulla et al.'s 2023 survey documented the conver- gence of ransomware-as-a-service (RaaS) business mod- els with AI capabilities, noting that affiliate programs now offer AI-powered tooling as a competitive differentiator. The implications are significant: even unsophisticated affil- iates can deploy ransomware variants that exhibit adaptive behavior traditionally associated with advanced persistent threats.

4.5 The Automated Attack Lifecycle

The individual capabilities surveyed above— reconnaissance, weaponization, delivery, exploitation, and action-on-objectives—can be chained into end-to-end automated attack pipelines. Figure 1 illustrates how AI augments each phase of the attack lifecycle, from initial reconnaissance through exfiltration.

The key insight is that AI does not merely accelerate in- dividual phases—it enables *closed-loop* attack automation in which the output of one phase dynamically informs the next. An RL-based penetration agent, for example, uses the results of automated reconnaissance to select exploits, observes the outcome, updates its network model, and re- plans—all without human intervention [13]. This closed- loop character qualitatively distinguishes AI-driven attacks from script-based automation.

V. AI-Driven Defensive Cyber Security

If the preceding sections paint a sobering picture of AI- enabled offense, this section examines the defensive arse- nal. The core premise is straightforward: human analysts

Figure 1: AI-augmented cyber -attack lifecycle. Each phase is annotated with the AI techniques that enhance its automation, adaptively, and evasion capabilities. Traditional attack phases (outer ring) are mapped to their AI-enabled counterparts (inner annotations).

cannot scale to the volume, velocity, and variety of mod- ern threats, and AI offers the most viable path to bridg- ing this gap. The practical challenge—distinguishing gen- uine threats from background noise at machine speed with- out introducing unacceptable false-positive rates—remains formidable.

5.1 AI-Driven Intrusion Detection and Prevention Sys- tems

Intrusion detection has been the dominant proving ground for ML-based security research. Khraisat et al.'s 2019 survey identified over 200 published studies on ML- based IDS, covering anomaly detection, misuse detection, and hybrid approaches [56]. Liu and Lang categorized the methodological landscape into shallow methods (SVM, random forest, *k*-NN) and deep methods (CNN, LSTM, au- toencoder), noting that deep methods consistently outper- formed on datasets with high-dimensional feature spaces but showed marginal improvements on well-curated, low- dimensional features [57].

Ahmad et al.'s 2021 systematic review assessed 130 stud- ies spanning both classical and deep-learning IDS, conclud- ing that while detection accuracy on benchmark datasets (NSL-KDD, CICIDS 2017) routinely exceeds 99%, these figures are misleading indicators of real-world perfor- mance [58]. The disconnect arises from several sources: benchmark datasets contain synthetic or outdated traffic, class imbalance is artificially corrected, and evaluation rarely accounts for adversarial evasion. Bridges et al. ex- tended this critique to host-based IDS, documenting sim- ilar gaps between laboratory accuracy and operational ef- fectiveness [59].

5.2 Behavioral Analytics and User Entity Behavior An- alytics

Where network-level IDS monitors traffic, behavioral analytics focuses on the patterns of users and entities—login times, access patterns, data movement volumes— to identify anomalies indicative of compromise. Sarker's 2021 overview of deep cybersecurity positioned UEBA as a critical complement to perimeter-based detection, partic- ularly for detecting insider threats and compromised cre- dentials that bypassed network controls [30].

The technical challenge is establishing reliable base- lines in dynamic environments. Dasgupta et al. noted that behavioral models must continuously adapt to organiza- tional changes—new employees, role transitions, seasonal workload variations—without triggering false alerts [55]. Federated learning has been proposed as a solution for multi-tenant environments, allowing behavior models to be trained across organizational boundaries without sharing raw data [62], though the security of federated learning it- self is contested (see Section VII).

5.3 Extended Detection and Response (XDR)

XDR platforms represent an architectural evolution from siloed detection tools toward integrated threat manage- ment. By correlating telemetry across endpoints, networks, email, cloud workloads, and identity providers, XDR sys- tems aim to reconstruct full attack narratives from frag- mented signals [36]. The AI component in modern XDR is twofold: cross-domain correlation engines that use graph- based models to link seemingly unrelated events, and prior- itization algorithms that rank incidents by estimated impact and confidence [53].

Microsoft's Digital Defense Report 2023 disclosed that their XDR platform processes over 65 trillion signals daily, using AI to reduce the median detection-to-response gap from hours to minutes for automated response work- flows [53]. While these figures are vendor-reported and should be treated with appropriate skepticism, they suggest that AI-augmented XDR is moving from aspiration to op- erational reality at major cloud providers.

5.4 Security Orchestration, Automation, and Response (SOAR)

SOAR platforms automate incident-response work- flows through predefined playbooks triggered by detec- tion events. Islam et al.'s multi-vocal review found that SOAR adoption reduced mean time to respond (MTTR) by 60–80% in surveyed organizations, primarily by automat- ing tier-1 alert triage and standardizing containment ac- tions [60]. AI enhances SOAR in two ways: NLP-based alert enrichment that extracts actionable context from un- structured log data, and decision-support models that rec- ommend response actions based on historical incident out- comes.

The integration of SOAR with AI-based detection cre- ates a tempting but risky feedback loop. If the detection model generates a false positive, the SOAR playbook may execute an automated containment action (e.g., quarantin-

Table 2: Comparative Analysis of AI-Driven Defensive Approache

Approach	Core ML Technique	Strengths	Limitations	Key References
Network IDS (Anomaly)	Autoencoder, LSTM, Isolation Forest	Unknown-threat detection; no signature updates needed	High FPR; concept drift; adversarial evasion	[56 –58]
Network IDS (Signature + ML)	Hybrid CNN-RF, gradient boosting	Low FPR; interpretable decisions	Misses novel attacks; rule maintenance overhead	[17, 29]
Host-based IDS	System-call sequence RNN, file-access graph analysis	Detects fileless & living-off-the-land attacks	Performance overhead; OS dependence	[12, 59]
Behavioral Analytics (UEBA)	Bayesian networks, clustering, temporal CNN	Insider threat detection; credential abuse discovery	Requires long baselines; privacy concerns	[30, 55]
SOAR / Playbook Automation	Decision trees, NLP for alert triage	Reduces analyst workload 60–80%; standardizes response	Brittle to novel attack patterns; integration complexity	[60, 61]
XDR Platforms	Cross-domain correlation, graph neural networks	Holistic visibility; cross-telemetry correlation	Vendor lock-in; data-volume challenges	[36, 53]

ing a host or blocking an IP), causing operational disruption without human review. Uetz et al. developed the SOCBED testbed precisely to evaluate such end-to-end automation pipelines under realistic conditions, finding that without careful threshold calibration, automated response amplifies the cost of false positives rather than reducing it [61].

5.5 AI-Powered Threat Hunting

Proactive threat hunting—searching for evidence of compromise that evades automated detection—has tradi- tionally been a purely human activity, reliant on the hunter's intuition and domain expertise. AI is beginning to augment this process. Machine learning models trained on historical hunting outcomes can suggest high-yield hypotheses, prior- itize assets for investigation, and automate the tedious data- retrieval steps that consume most of a hunter's time [63].

The AI security architecture integrating these defensive layers is depicted in **Figure 2**.

VI. AI-Driven Threat Intelligence

Threat intelligence (TI) has evolved from a manual, report-driven discipline into a data-intensive pipeline where machine learning plays an increasingly central role. The volume of threat data—vulnerability disclosures, malware samples, dark-web chatter, network telemetry—has grown beyond the capacity of human analysts to process, creating a natural entry point for AI augmentation.

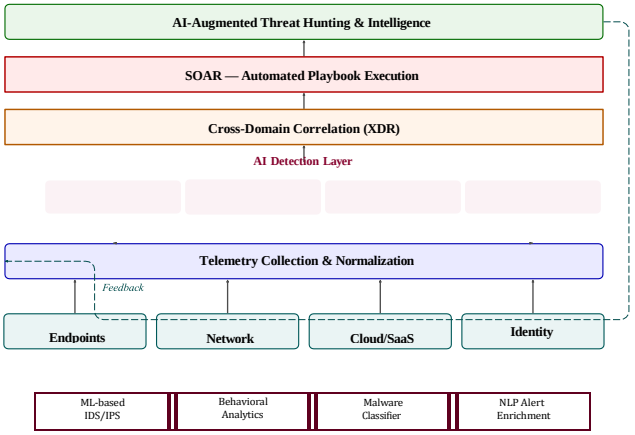

Figure 2: Integrated AI security architecture showing the relationship between AI-powered detection (IDS/IPS, UEBA), correlation (XDR), automation (SOAR), and proactive hunting layers. Arrows indicate data and decision flows.

6.1 Predictive Threat Intelligence

The most ambitious application of AI to TI is prediction: forecasting attacks before they occur. Sun et al.'s survey distinguished between short-horizon prediction (e.g., fore- casting the next target of an active campaign within hours) and strategic prediction (e.g., identifying which vulnerabil- ity classes will be exploited in the coming quarter) [32]. The former has shown promising results using time-series models over IOC streams, while the latter remains largely aspirational due to the chaotic dynamics of the threat land- scape.

Schlette et al. examined the intelligence cycle from a se- curity incident response perspective, finding that AI-basedenrichment (automated IOC scoring, entity resolution, re- port summarization) significantly reduced analyst work- load during the processing and analysis phases [33]. How- ever, they cautioned that over-reliance on automated enrich- ment risks creating a monoculture: if multiple organiza- tions use the same enrichment models, an adversary who understands the model's blind spots can craft threats that evade detection uniformly across all subscribers.

6.2 Big Data Analytics for Security

The scale of security telemetry in large organizations— billions of log events per day across endpoints, fire- walls, cloud services, and identity providers—demands dis- tributed analytics frameworks. Zeadally et al. surveyed the application of big-data techniques (MapReduce, stream processing, graph analytics) to cybersecurity, noting that the primary challenge is not computation but data quality: noisy, incomplete, and inconsistent logs degrade model per- formance regardless of algorithmic sophistication [64].

Li et al.'s cross-disciplinary survey argued that the inte- gration of AI with security big-data analytics is hampered by a talent gap: security analysts lack ML expertise, while data scientists lack security domain knowledge [65]. This observation helps explain why many AI-security tools re- main underutilized in operational settings despite strong benchmark results in research contexts.

6.3 Intelligence Sharing Platforms and Standards

Effective defense requires not just generating intelli- gence but sharing it across organizational boundaries. Wag- ner et al.'s survey documented the evolution of sharing plat- forms from informal mailing lists to structured frameworks such as STIX (Structured Threat Information eXpression) and TAXII (Trusted Automated eXchange of Indicator In- formation) [34]. AI enhances sharing in two ways: auto- mated translation of unstructured threat reports into struc- tured STIX bundles using NLP, and anomaly detection ap- plied to incoming feeds to filter unreliable or outdated in- telligence.

Tounsi and Rais's analysis identified a persistent ten- sion between sharing breadth and intelligence quality [31]. Broad sharing increases the chance that a relevant IOC reaches the right defender, but also increases the noise floor and the risk of intelligence poisoning—a variant of data poisoning in which an adversary injects false IOCs into shared feeds to either overwhelm analysts or calibrate per- ceptions of which threat vectors are active.

The AI threat intelligence pipeline synthesizing these components is depicted in Figure 3.

VII. Adversarial Machine Learning

Adversarial machine learning (AML) occupies a unique position in this review: it is simultaneously a tool for at-

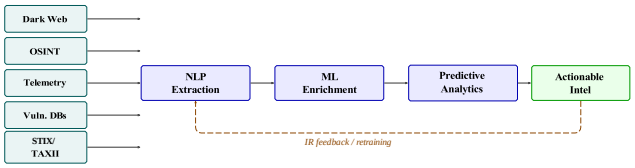

Figure 3: AI-enabled threat intelligence pipeline. Raw data from multiple sources (dark web, OSINT, telemetry) flows through NLP-based extraction, ML-driven enrichment, and predictive analytics before producing actionable intelligence. Feedback loops from incident response continuously retrain models.

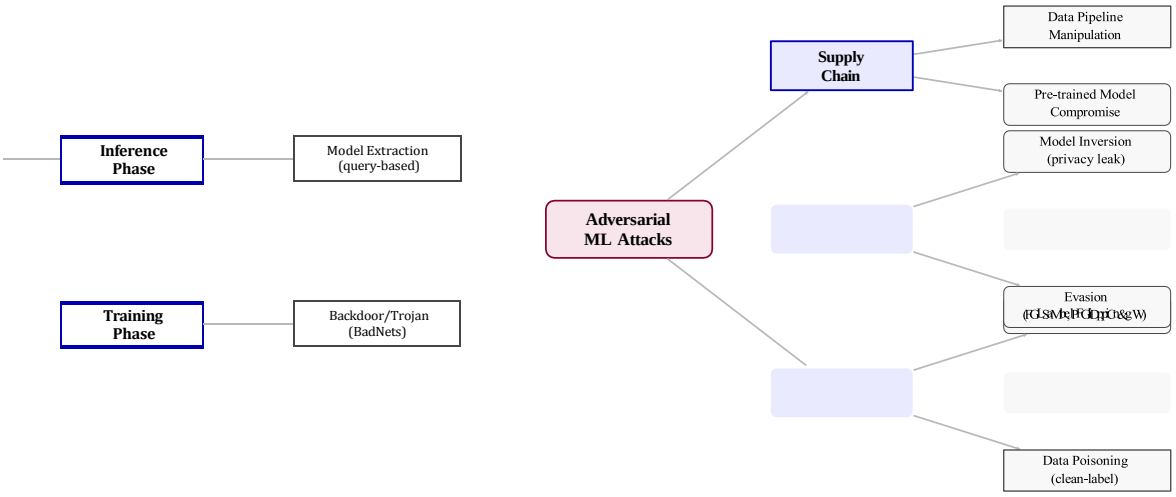

Figure 4: Taxonomy of adversarial machine learning attacks organized by phase, goal, and attacker knowledge. Each leaf node corresponds to a specific attack class discussed in the text.

tackers seeking to evade AI-based defenses and a research discipline aimed at hardening those defenses. The ten- sion between these roles makes AML the most intellec- tually demanding—and practically consequential—topic at the AI-security intersection.

7.1 Taxonomy of Adversarial Threats

Following the structure of the MITRE ATLAS frame- work [37] and the taxonomies proposed by Biggio and Roli [66] and Chakraborty et al. [67], adversarial threats to ML systems can be organized along three axes: the *at- tack phase* (training vs. inference), the *attacker's goal* (in- tegrity, availability, or confidentiality), and the *attacker's knowledge* (white-box, grey-box, or black-box). Figure 4 illustrates this taxonomy.

7.2 Evasion Attacks

Evasion attacks—crafting inputs that cause a deployed model to misclassify—are the most extensively studied class of AML threats. Goodfellow et al.'s Fast Gradi- ent Sign Method (FGSM) demonstrated that even minimal, human-imperceptible perturbations to input data could flip classifier outputs with high confidence [22]. Subsequent work by Carlini and Wagner introduced optimization-based

Table 3: Summary of Adversarial Machine Learning Attack Types

Attack Type	Phase	Knowledge	Goal	Primary Defense	Key Work
FGSM	Inference	White-box	Evasion	Adversarial training	Goodfellow [22]
PGD	Inference	White-box	Evasion	Certified defenses	Madry [68]
C&W Attack	Inference	White-box	Evasion	Detection + ensembling	Carlini [23]
IDSGAN	Inference	Black-box	Evasion	Feature hardening	Lin [50]
Clean-Label Poison	Training	Grey-box	Backdoor	Data sanitization	Shafahi [25]
BadNets (Trojan)	Training	White-box	Backdoor	Neural cleanse	Gu [24]
Model Extraction	Inference	Black-box	Theft	Query limiting; watermarking	Tramèr [26]
Data Poisoning	Training	Grey-box	Integrity	Robust statistics	Goldblum [69]

attacks that produce perturbations closer to the impercepti- bility boundary, establishing what remains the gold stan- dard for evaluating robustness [23]. Madry et al. proposed projected gradient descent (PGD) as a principled adversar- ial training defense and showed that models trained against PGD attacks achieve measurably higher robustness, though at a significant cost in clean-data accuracy [68].

In the cybersecurity context, evasion attacks target intru- sion detection systems, malware classifiers, and spam fil- ters. Lin et al.'s IDSGAN used generative adversarial net- works to modify malicious traffic features such that they were classified as benign by black-box IDS models while preserving attack functionality [50]. Rigaki and Garcia demonstrated a similar technique for disguising malware C2 traffic as legitimate web browsing [49]. These results challenge the assumption that ML-based detection is inher- ently superior to signature-based methods: if an adversary with knowledge of the detection model's architecture can reliably evade it, the investment in ML may yield a false sense of security [29].

A persistent debate in the literature concerns the *prac- tical feasibility* of evasion attacks. Papernot et al. ar- gued that the computational cost and access require- ments of white-box attacks limit their operational deploy- ment [70], while subsequent work on transferability—the phenomenon whereby adversarial examples crafted against one model also fool a different model—suggests that black- box evasion is more practical than initially assumed [22]. The resolution of this debate has significant implications for defensive resource allocation.

7.3 Data Poisoning

Data poisoning attacks target the training pipeline rather than the deployed model. By injecting carefully crafted samples into the training dataset, an attacker can bias the learned decision boundary to create backdoors or degrade overall accuracy. Shafahi et al. demonstrated "clean-label" poisoning, in which the injected samples carry correct la- bels and are individually benign, making them difficult to detect through manual inspection [25].

Goldblum et al.'s comprehensive treatment of dataset security documented the expanding attack surface as ML practitioners increasingly rely on web-scraped, crowd-sourced, or third-party training data [69]. In cybersecu- rity applications, poisoning risks are particularly acute for models trained on shared threat-intelligence feeds (see Sec- tion VI): an adversary who contributes false IOCs to a shared platform can degrade the detection models of all downstream consumers simultaneously. Gu et al.'s BadNets work established the concept of Neu- ral Trojan attacks, where a poisoned model behaves nor- mally on clean inputs but exhibits attacker-chosen behavior when a specific trigger pattern is present in the input [24]. The supply-chain implications are severe: organizations that deploy pre-trained or fine-tuned models from untrusted sources risk inheriting embedded backdoors that are unde- tectable through standard evaluation metrics.

7.4 Model Stealing and Extraction

Model stealing attacks aim to reconstruct a proprietary model's parameters or decision boundary through query ac- cess alone. Tramèr et al.'s seminal work demonstrated that linear models, decision trees, and shallow neural networks deployed behind prediction APIs could be replicated with high fidelity using a modest number of queries [26]. While deeper networks present greater extraction challenges, sub- sequent research has shown that knowledge distillation techniques can approximate even large models to within operationally useful accuracy.

In cybersecurity, model stealing has a particularly per- nicious implication: if an attacker can extract a copy of a deployed IDS classifier, they can mount white-box eva- sion attacks against it offline, then deploy evasion payloads against the production system with high confidence of suc- cess [66]. This two-stage attack—extraction followed by evasion—invalidates the common assumption that black- box deployment provides sufficient protection for security models.

7.5 Defenses Against Adversarial Attacks

Defensive strategies span several categories. *Adversar- ial training* augments the training set with adversarial ex- amples, improving robustness at the cost of clean-data ac- curacy [68]. *Certified defenses* provide provable robust- ness guarantees within bounded perturbation regions but scale poorly to high-dimensional inputs. *Input preprocess-ing* (randomized smoothing, input transformation) attempts to neutralize adversarial perturbations before they reach the model [67]. *Model ensembling* diversifies the decision sur- face, making it harder for a single adversarial example to fool all ensemble members. No single defense currently provides satisfactory protec- tion across all attack types. Carlini and Wagner's evalu- ation methodology revealed that many published defenses were circumvented by adaptive attacks [23], establishing an important methodological principle: defenses must be eval- uated against adversaries that are aware of and adapted to the specific defense mechanism, not merely against static attack methods.

VIII. Case Studies

Abstract analysis benefits from grounding in concrete in- cidents. This section presents four case studies that illus- trate how AI capabilities—both offensive and defensive— have manifested in operational environments. Each case is reconstructed from published reports, academic analyses, and industry threat intelligence.

8.1 Case Study 1: AI-Enhanced Spear Phishing Cam- paigns

The evolution from template-based phishing to AI- personalized campaigns reached a notable inflection point in 2023, when threat intelligence firms began attribut- ing specific campaigns to LLM-assisted content genera- tion. CrowdStrike's 2024 Global Threat Report docu- mented multiple campaigns by state-affiliated actors using AI-generated lure emails that incorporated recently pub- lished information about targets scraped from LinkedIn profiles, academic publications, and organizational web- sites [2].

The distinguishing feature of these campaigns was lin- guistic sophistication. Unlike earlier phishing operations that were detectable through grammatical errors, awkward phrasing, or cultural incongruities, the AI-generated emails exhibited native-level fluency in multiple languages and adapted their register to match the target's professional con- text [5]. Defensive teams reported that traditional content- based phishing filters—trained on corpora of older, cruder phishing attempts—failed to flag the new generation of AI- crafted lures, necessitating retraining on updated datasets and the development of stylometric detection methods that analyze writing-style consistency rather than content alone.

8.2 Case Study 2: Deepfake Audio in Business Email Compromise

In a widely reported 2023 incident documented by Man- diant, attackers used AI-cloned voice samples to imper- sonate a multinational corporation's CFO during a video conference, authorizing a $25 million transfer to attacker-

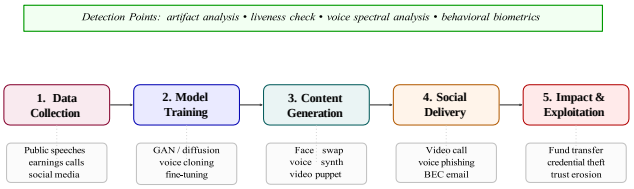

Figure 5: Deepfake attack lifecycle: from public data collection through model training, content generation, and social delivery to exploitation, with detection intervention points highlighted.

controlled accounts [47]. The attack exploited two vulner- abilities: the availability of voice training data (recorded earnings calls and media interviews are publicly accessi- ble for senior executives) and the absence of multi-factor authentication for high-value financial authorizations con- ducted via video call.

The incident catalyzed industry interest in deepfake de- tection for real-time communications. Tolosana et al. noted that while frame-level deepfake detection in pre- recorded video achieved accuracy above 95% on bench- mark datasets, real-time detection during live video calls remained unreliable due to compression artifacts, network jitter, and the need for sub-second latency [45]. The case underscores a recurring theme in this review: detection technology lags behind generation technology, and the gap is widest in operational conditions.

8.3 Case Study 3: LLM-Assisted Malware Develop- ment

The practical capability of LLMs to assist in malware de- velopment moved from theoretical concern to documented reality in 2023. Pa Pa et al.'s systematic evaluation showed that ChatGPT could generate functional components— reverse shells, keyloggers, file encryption routines—when prompted through task decomposition, even though direct malware-generation requests were blocked by safety fil- ters [9].

Underground forums subsequently documented threat actors sharing "jailbreak" prompts and workflow guides for using commercial LLMs to accelerate malware develop- ment [40]. The Microsoft Digital Defense Report 2023 identified LLM-assisted code generation as an emerging threat vector, noting that while current safety guardrails prevent the most straightforward misuse, "sufficiently mo- tivated actors with moderate technical skill can circumvent them" [53].

What makes this case study particularly instructive is the dual-use nature of the underlying capability. The same LLM features that enable malware generation—code completion, debugging assistance, explanation of low-level constructs—are also invaluable for legitimate security re- search, red teaming, and education. This duality defies simple regulatory solutions and argues for defense-in-depth approaches that do not rely solely on model-level content filtering.

8.4 Case Study 4: Autonomous Vulnerability Discovery

The DARPA Cyber Grand Challenge (CGC) of 2016 remains the most vivid demonstration of AI-driven au- tonomous vulnerability discovery. Seven fully autonomous systems competed in a Capture-the-Flag tournament, dis- covering, exploiting, and patching vulnerabilities in real time without human intervention [48]. The winning sys- tem, Mayhem (developed by ForAllSecure), combined symbolic execution, fuzzing, and binary analysis to iden- tify and patch vulnerabilities faster than most human com- petitors.

Post-CGC developments have extended this capability. VulDeePecker applied deep learning to source-code vulner- ability detection, achieving precision and recall rates that exceeded rule-based static analyzers [42]. The integration of LLMs into vulnerability triage—using natural-language models to reason about code semantics, generate proof-of- concept exploits, and propose patches—has further com- pressed the vulnerability lifecycle [39].

For defenders, these tools offer the prospect of finding and fixing vulnerabilities before adversaries do. For attack- ers, the same tools provide an automated path from vulner- ability discovery to weaponization, reducing the time and expertise required to develop operational exploits [54]. The net security effect depends on a race condition: whether defenders adopt these tools faster and more broadly than attackers, a question that remains empirically unresolved.

IX. Ethics, Policy, and Governance

The technical analysis in preceding sections reveals a re- curring theme: the same AI capabilities serve both attack- ers and defenders, and the line between legitimate security research and offensive capability development is uncom- fortably thin. This dual-use character elevates governance from a peripheral concern to a central challenge.

9.1 The Dual-Use Dilemma

Brundage et al.'s 2018 report on the malicious use of AI articulated the dual-use problem with unusual clarity: restricting access to powerful AI models impedes defen- sive research just as effectively as it constrains attack- ers [7]. The subsequent release of increasingly capable open-source models (LLaMA, Mistral, Falcon) has made access-control strategies largely moot for foundation mod- els, shifting the governance debate toward application-layer controls and post-deployment monitoring.

The cybersecurity community faces a version of this dilemma that is more acute than most other domains. Pub- lishing vulnerability research enables defenders to patch but also provides attackers with exploitation guides. Pub- lishing adversarial ML techniques enables defenders to test robustness but also provides attackers with evasion blueprints. This tension is not new —responsible disclo-sure has been debated for decades—but AI amplifies its consequences by lowering the expertise required to oper- ationalize published knowledge [27].

9.2 Responsible AI in Cyber security

The concept of responsible AI—encompassing fairness, transparency, accountability, and safety—has been exten- sively discussed in general contexts [71, 72] but has re- ceived comparatively limited attention in cyber security ap- plications. Capuano et al.'s survey on explainable AI in cyber security highlighted a fundamental tension: secu- rity models must be interpretable enough for analysts to trust their outputs, yet the most interpretable models tend to be the easiest for adversaries to reverse-engineer and evade [73].

Google's Secure AI Framework (SAIF) represents an in- dustry attempt to operationalize responsible AI principles specifically for security applications [74]. SAIF extends conventional AI safety frameworks with security-specific guidance on model supply-chain integrity, adversarial ro- bustness testing, and least-privilege access to model ca- pabilities. Its adoption by a major cloud provider sig- nals growing recognition that AI security requires purpose- built governance frameworks rather than generic AI ethics guidelines.

9.3 Regulatory Landscape

The regulatory environment for AI in cybersecurity is fragmented and rapidly evolving. The European Union's AI Act classifies AI systems by risk level and imposes requirements ranging from transparency obligations for limited-risk systems to outright prohibition of certain high- risk applications [75]. Cybersecurity applications occupy an ambiguous position in this taxonomy: defensive AI tools are generally considered beneficial, but the same tools re- purposed offensively could fall under high-risk or prohib- ited categories.

In the United States, the NIST AI Risk Management Framework provides voluntary guidance for organizations deploying AI systems, with specific attention to security implications [76]. The framework emphasizes contextual evaluation: the risk profile of an AI system depends not only on its technical characteristics but on the deployment environment, threat model, and organizational governance capacity. This contextual approach contrasts with the EU's more prescriptive classification-based regime.

Cath et al. compared US, EU, and UK approaches to AI governance, finding significant disagreements on the appro- priate balance between innovation incentives and precau- tionary regulation [77]. In cybersecurity specifically, overly restrictive regulation risks pushing AI security research and development to less regulated jurisdictions, creating geo- graphic asymmetries in defensive capability.

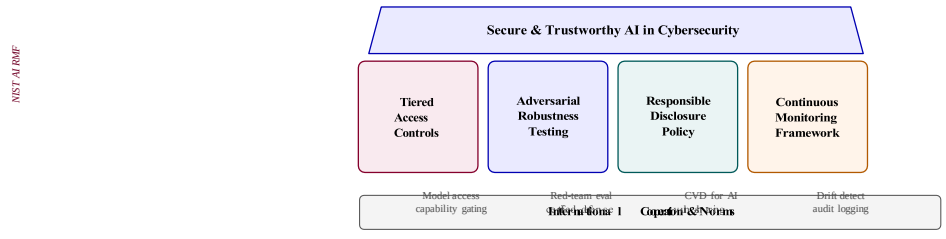

Figure 6: Proposed governance framework for AI in cyber security: four pillars resting on international cooperation, guided by NIST AI RMF and EU AI Act principles.

9.4 International Cyber Policy and AI

The geopolitical dimension cannot be ignored. Nation- states are investing heavily in both offensive and defensive AI capabilities, with limited international norms governing their use [4]. The ENISA AI cyber security challenges re- port identified the absence of international agreements on AI-enabled cyber operations as a critical governance gap, noting that existing frameworks (Tallinn Manual, Budapest Convention) predate the AI era and lack provisions for au- tonomous or semi-autonomous cyber weapons [78].

The prospect of AI vs. AI cyber conflict (discussed fur- ther in Section X) introduces additional policy challenges. Autonomous systems that can detect, attribute, and respond to cyber -attacks at machine speed raise questions about es- calation dynamics, proportionality, and accountability that existing international law is ill-equipped to address [7].

9.5 Toward a Governance Framework

Drawing on the literature surveyed above, we identify four pillars for an effective governance framework at the AI-cyber security intersection:

1. **Tiered access controls**: Differentiated access to high- risk capabilities (e.g., autonomous exploit generation) based on demonstrated need and accountability struc- tures.
2. **Mandatory adversarial robustness testing**: Requiring AI systems deployed in security-critical roles to undergo standardized adversarial evaluation before deployment.
3. **Transparency with limits**: Publishing model architec- tures and evaluation methodologies while restricting ac- cess to pre-trained weights optimized for offensive ap- plications.
4. **International coordination**: Establishing norms for re- sponsible disclosure of AI-security vulnerabilities and limits on autonomous cyber weapons, analogous to arms-control frameworks.

X. Future Trends and EMERGING DIRECTIONS

Extrapolating from the trajectories documented in this review, several emerging directions are likely to define the next phase of the AI-cyber security relationship. We ground these projections in published research while acknowledge- ing the inherent uncertainty of technological forecasting.

10.1 Autonomous Cyber Warfare

The convergence of reinforcement learning, automated exploit generation, and AI-driven command and control points toward a future in which cyber operations are conducted with minimal human oversight. Nguyen and Reddi's survey documented the maturation of RL-based cy- ber agents from proof-of-concept demonstrations to sys- tems capable of multi-step attack execution in realistic sim- ulated environments [51]. Sewak et al. projected that within a decade, autonomous agents will handle routine penetra- tion testing, incident response, and threat hunting with hu- man oversight limited to strategic decisions [63].

The policy implications are profound. Autonomous of- fensive cyber tools compress the decision loop from hours or days to seconds, raising concerns about inadvertent es- calation, attribution uncertainty, and the erosion of human judgment in contexts where the consequences of error are severe [7]. International law has yet to grapple seriously with the prospect of AI systems that can independently de- cide to launch, escalate, or de-escalate cyber operations.

10.2 AI vs. AI: The Coming Adversarial Equilibrium

As both attackers and defenders deploy increasingly so- phisticated AI systems, the cybersecurity landscape may evolve toward an adversarial equilibrium in which the pri- mary competitive axis is model quality rather than human expertise. GAN-based attack evasion versus adversarially trained detection [50, 68], LLM-generated phishing versus LLM-powered content analysis [5, 6], and RL-driven pen- etration versus RL-hardened perimeter defense [13, 14] al- ready prefigure this dynamic.

Truong et al.'s analysis of AI in offense and defense spec- ulated that this arms race may stabilize around an equi- librium where neither side achieves persistent advantage, analogous to the outcome of adversarial training dynam- ics in GAN theory [4]. However, the analogy is imper- fect: unlike GAN training, the cybersecurity arms race involves heterogeneous actors with asymmetric resources, objectives, and risk tolerances.

10.3 Quantum Computing and AI Security

The intersection of quantum computing and AI-driven cybersecurity introduces both threats and opportunities. Mosca's 2018 analysis estimated that quantum comput- ers capable of breaking RSA-2048 and ECC cryptography

Table 4: Comparison of Major AI Governance Frameworks for Cyber security

Framework	Issuing Body	Scope	Approach	Cyber Focus
AI RMF 1.0 [76]	NIST (USA)	All AI systems	Voluntary, risk-based	Moderate — security mentioned
EU AI Act [75]	European Parliament	AI in EU market	Mandatory, tiered risk	Limited — dual-use provisions
MITRE ATLAS [37]	MITRE Corp.	AI/ML systems	Threat taxonomy	High — AI attack-focused
Google SAIF [74]	Google	Cloud AI services	Best-practices guide	High — supply-chain, robustness
ENISA AI Threats [78]	ENISA (EU)	AI threat landscape	Advisory report	High — threat cataloguing

could emerge within 15–20 years, necessitating a proactive transition to post-quantum cryptographic standards [79]. AI has a role to play in this transition: machine learn- ing models can optimize post-quantum key exchange pro- tocols, identify quantum-vulnerable cryptographic imple- mentations in legacy codebases, and accelerate the design of quantum-resistant algorithms.

Conversely, quantum machine learning—leveraging quantum computing to train more powerful models—could amplify both offensive and defensive AI capabilities. Pi- randola et al.'s survey on quantum cryptography noted that quantum key distribution (QKD) promises information- theoretically secure communication channels [80], but the practical deployment of QKD networks remains limited by distance constraints, cost, and integration complexity with classical infrastructure.

10.4 Self-Healing Networks and Autonomous Re- silience

The concept of self-healing networks—systems that au- tonomously detect, isolate, and remediate security incidents without human intervention—represents the logical end- point of defensive AI automation. Liang et al. surveyed the application of ML to IoT security, noting that the scale and heterogeneity of IoT deployments make manual inci- dent response impractical and autonomous remediation es- sential [81].

Key technical challenges include ensuring that au- tonomous remediation actions do not cause greater opera- tional disruption than the incidents they address, maintain- ing resilience against adversarial manipulation of the re- mediation system itself, and establishing accountability for autonomous decisions that affect system availability [82]. The NIST AI Risk Management Framework acknowledges these challenges but provides limited operational guidance for their resolution [76].

10.5 Explain ability as a Security Requirement

As AI systems assume greater autonomy in security- critical decisions, explainability transitions from a de- sirable property to a functional requirement. Arrieta et al.'s comprehensive survey on explainable AI (XAI) catalogued methods ranging from post-hoc interpretation (SHAP, LIME) to inherently interpretable architectures (at- tention mechanisms, decision lists) [72]. In cybersecurity,

Explain ability serves dual purposes: enabling analysts to validate model outputs (reducing over-reliance on opaque classifiers) and supporting forensic investigation when au- tomated systems take consequential actions [73].

The tension between explainability and adversarial robustness—interpretable models are easier for attackers to reverse-engineer—remains an active area of research with no satisfactory resolution to date.

XI. Research Gaps and Open Problems

The literature reviewed in this paper reveals not only sub- stantial progress but also significant lacunae—areas where fundamental questions remain unanswered, where empiri- cal evidence contradicts prevailing assumptions, or where the research community's attention has been disproportion- ately focused on tractable sub-problems at the expense of more consequential ones. This section organizes these gaps into thematic clusters.

11.1 The Evaluation Gap: Laboratory vs. Operational Performance

The most persistent gap in AI-security research is the dis- connect between benchmark performance and real-world effectiveness. The surveys by Ahmad et al. [58], Sommer and Paxson [16], and Apruzzese et al. [29] collectively doc- ument a field in which 99%+ detection accuracy on stan- dardized datasets coexists with widespread dissatisfaction among practitioners who deploy these systems.

The root causes are well understood: benchmark datasets are static, balanced, and often synthetic, while produc- tion environments are dynamic, imbalanced, and adver- sarial. What is less understood is how to close this gap. The SOCBED testbed [61] and similar initiatives represent steps toward more realistic evaluation, but the community lacks consensus on standardized evaluation protocols that account for adversarial evasion, concept drift, and opera- tional constraints.

Open problem: Developing evaluation methodologies for AI-based security tools that are predictive of real- world performance, with standardized adversarial bench- marks that evolve alongside the threat landscape.

11.2 Adversarial Robustness Under Realistic Threat Models

Adversarial ML research has produced a rich theoret- ical understanding of model vulnerabilities, but much of this work operates under idealized threat models. White- box access, unlimited query budgets, and perfect knowl- edge of the data distribution are standard assumptions that rarely hold in practice [70]. Conversely, the most practical attacks—those that operate with limited access and noisy feedback—are understudied relative to their real-world rel- evance [66].

Contradiction in literature: Good fellow et al. and sub- sequent work suggest that adversarial transferability makes black-box attacks practical, while Papernot et al. argue that practical constraints significantly limit real-world applica- bility. This contradiction remains unresolved and has direct implications for how much organizations should invest in adversarial robustness versus other defensive measures.

11.3 Cross-Domain Transfer and Generalization

Most AI-security models are trained and evaluated on a single data modality (network traffic, system calls, mal- ware binaries). Real-world attacks, however, span multiple modalities: a phishing email (NLP) delivers a malicious attachment (binary analysis) that establishes network per- sistence (traffic analysis). XDR platforms attempt cross- domain correlation, but the ML models underlying them rarely generalize across domains [36, 53].

Open problem: Developing multi-modal security mod- els that can learn representations transferable across net- work, endpoint, and application-layer data, analogous to foundation models in NLP and computer vision.

11.4 The Governance Vacuum

Despite the frameworks discussed in Section IX, there is no internationally recognized standard for evaluating or regulating AI systems used in cyber security. The NIST AI RMF [76] and EU AI Act [75] provide general guid- ance but lack cyber security-specific provisions. MITRE ATLAS [37] documents AI-specific threats but does not prescribe defensive standards.

Open problem: Creating a governance framework that addresses the unique characteristics of AI in cyber security—dual-use capabilities, rapid capability evo- lution, asymmetric attacker-defender dynamics—without stifling legitimate research.

11.5 Data Scarcity and Sharing Barriers

High-quality labeled security data is scarce. Organiza- tions are reluctant to share incident data due to legal li- ability, reputational risk, and competitive concerns. Fed- erated learning has been proposed as a privacy-preserving alternative, but its security guarantees are themselves con- tested [62]. The result is a field where most models are trained on a small number of public datasets that may not represent the diversity of real-world threat environments.

Open problem: Developing practical mechanisms for collaborative model training across organizations that pre- serve privacy, resist poisoning, and produce models gener- alizable to diverse environments.

11.6 Explain ability-Security Trade-off

The tension between model interpretability and adver- sarial robustness (discussed in Sections VII and X) remains without a satisfactory theoretical resolution. Existing work approaches explainability and robustness as separate op- timization objectives [72, 73]; a unified framework that characterizes the fundamental trade-off curve would signif- icantly advance both fields.

11.7 Proposed Research Agenda

Based on the gaps identified above, we propose a struc- tured research agenda organized into immediate (1–3 year), medium-term (3–7 year), and long-term (7+ year) priori- ties:

1. **Immediate**: Standardized adversarial evaluation bench- marks; cross-organizational data-sharing frameworks with differential privacy guarantees; explainability stan- dards for security-critical AI deployments.

2. **Medium-term**: Multi-modal security foundation mod- els; international norms for AI-enabled cyber opera- tions; provably robust detection under realistic threat models.

3. **Long-term**: Autonomous defensive agents with formal safety guarantees; unified theory of the explainability-robustness trade-off; quantum-resilient AI security ar- chitectures.

XII. CONCLUSION

This review has traversed the AI-cyber security landscape from foundational methods to frontier challenges, synthe- sizing over seventy studies to construct a coherent picture of a field in rapid, consequential evolution. Several conclu- sions emerge with particular force.

First, the asymmetry between offensive and defensive AI adoption is structural, not incidental. Attackers operate without regulatory constraints, ethical review, or account- ability requirements, enabling faster iteration and deploy- ment of AI capabilities. Defenders, by contrast, must nav- igate compliance frameworks, organizational inertia, and the practical challenge of integrating AI into legacy security architectures. Closing this gap requires not only tech innovation but institutional reforms that accelerate defen- sive AI deployment without compromising safety.

Second, adversarial machine learning is not a niche con- cern but a *meta-threat* that conditions the reliability of every AI-based security tool. The literature consistently shows that models achieving near-perfect accuracy on clean data can be systematically degraded through evasion, poi- soning, and extraction attacks. Any organization deploy- ing AI-based defenses without adversarial robustness test- ing is, in effect, building its security on assumptions that a motivated adversary will invalidate. The MITRE ATLAS framework provides a starting point for threat modeling AI- specific attack surfaces, but operationalizing its guidance remains a work in progress for most organizations.

Third, the governance landscape is conspicuously under- developed relative to the technical capabilities it must reg- ulate. The EU AI Act, NIST AI RMF, and SAIF frame- work represent important steps, but none adequately ad- dresses the unique characteristics of AI in cybersecurity: the dual-use nature of capabilities, the speed of capabil- ity evolution, the asymmetric attacker–defender dynamic, and the geopolitical dimensions of AI-enabled cyber opera- tions. International coordination on norms for autonomous cyber weapons and responsible disclosure of AI-security vulnerabilities is not merely desirable—it is urgent.

Fourth, the field suffers from a persistent evaluation gap. Benchmark accuracy does not predict operational effective- ness, and the research community's reliance on a small number of aging, static datasets produces a distorted picture of practical readiness. The development of standardized, adversarially robust evaluation protocols is a prerequisite for translating laboratory advances into real-world security improvements.

Looking forward, the trajectory is toward greater au- tonomy on both sides of the attacker-defender divide. Reinforcement-learning agents that autonomously plan and execute multi-step attacks, AI systems that detect, con- tain, and remediate intrusions without human intervention, and generative models that produce novel attack artifacts at marginal cost near zero collectively define a landscape in which human judgment remains essential for strategic de- cisions but is increasingly augmented —and in some cases, replaced—by machine intelligence at the operational level.

The challenge for the cybersecurity community is to har- ness AI's defensive potential without being undone by its offensive applications, to build robustness into AI-based defenses without sacrificing the interpretability that en- ables human oversight, and to develop governance frame- works that are agile enough to keep pace with a technol- ogy whose capabilities evolve faster than any regulatory cy- cle. This review has mapped the terrain; navigating it will require sustained collaboration across academia, industry, and government—a task as much institutional as it is tech- nical.

References

[1] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez- Sanz, and V. Pospelova, "The emerging threat of AI-driven cyber- attacks: A review," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2037254, 2022.

[2] CrowdStrike, "2024 global threat report," CrowdStrike, Tech. Rep., 2024. [Online]. Available: https://www.crowdstrike.com/global-threat-report/

[3] Verizon, "2023 data breach investigations report," Verizon Business, Tech. Rep., 2023. [Online]. Available: https://www.verizon.com/business/resources/reports/dbir/

[4] T. C. Truong, Q. B. Diep, and I. Zelinka, "Artificial intelli- gence in the cyber domain: Offense and defense," *Symmetry*, vol. 12, no. 3, p. 410, 2020.

[5] J. Hazell, "Spear phishing with large language models,"*arXiv preprint arXiv:2305.06972*, 2023.

[6] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Ki- fayat, "A comprehensive survey of AI-enabled phishing at- tacks detection techniques," *Telecommunication Systems*, vol. 76, pp. 139–154, 2021.

[7] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley,B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, and Others, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228*, 2018.

[8] J. Seymour and P. Tully, "Weaponizing data science for so- cial engineering: Automated E2E spear phishing on Twitter," in *Proceedings of Black Hat USA*, 2016.

[9] Y. M. T. Pa Pa, S. Tanizaki, T. Kou, M. Van Eeten, K. Yosh- ioka, and T. Matsumoto, "An attacker's dream? exploring the capabilities of ChatGPT for developing malware," *arXiv preprint arXiv:2308.06857*, 2023.

[10] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compro- mising real-world LLM-integrated applications with indirect prompt injection," in *Proceedings of the ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023.

[11] S. Razaulla, C. Fachkha, C. Markarian, A. Gawanmeh,W. Mansoor, S. Taha, and M. Asim, "The age of ran- somware: A survey on the evolution, taxonomy, and re- search directions," *IEEE Access*, vol. 11, pp. 40 698–40 723, 2023.

[12] D. Ucci, L. Aniello, and R. Baldoni, "Survey of machine learning techniques for malware analysis," *Computers & Security*, vol. 81, pp. 123–147, 2019.

[13] M. C. Ghanem and T. M. Chen, "Reinforcement learning for efficient network penetration testing," in *Information*, vol. 11, no. 1, 2020, p. 6.

[14] J. Schwartz and H. Kurniawati, "Autonomous penetration testing using reinforcement learning," in *Proceedings of the Australasian Joint Conference on Artificial Intelligence*, 2019, pp. 2–14.

[15] M. M. Yamin, B. Katt, and V. Gkioulos, "Weaponized AI for cyber- attacks," *Journal of Information Security and Applications*, vol. 57, p. 102722, 2021.

[16] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proceedings of the IEEE Symposium on Security and Pri- vacy*, 2010, pp. 305–316.

[17] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[18] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao,H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35 365– 35 381, 2018.

[19] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, andM. Xu, "A survey on machine learning techniques for cy- ber security in the last decade," *IEEE Access*, vol. 8, pp. 222 310–222 354, 2020.

[20] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Jan- icke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *Journal of Information Security and Applications*, vol. 50, p. 102419, 2020.

[21] S. Gamage and J. Samarabandu, "Deep learning methods in network intrusion detection: A survey and an objective com- parison," *Journal of Network and Computer Applications*, vol. 169, p. 102767, 2020.

[22] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explain- ing and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2015.

[23] N. Carlini and D. Wagner, "Towards evaluating the robust- ness of neural networks," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2017, pp. 39–57.

[24] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," in *Proceedings of the Machine Learning and Computer Se- curity Workshop*, 2017.

[25] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Pro- ceedings of the International Conference on Neural In- formation Processing Systems (NeurIPS)*, 2018, pp. 6103– 6113.

[26] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ris- tenpart, "Stealing machine learning models via prediction APIs," in *Proceedings of the USENIX Security Symposium*, 2016, pp. 601–618.

[27] M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting artifi- cial intelligence in cybersecurity is a double-edged sword," *Nature Machine Intelligence*, vol. 1, no. 12, pp. 557–560, 2019.

[28] S. Mahdavifar and A. A. Ghorbani, "Application of deep learning to cybersecurity: A survey," *Neurocomputing*, vol. 347, pp. 149–176, 2019.

[29] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, "On the effectiveness of machine and deep learning for cyber security," *Computers & Security*, vol. 123, p. 102867, 2023.

[30] I. H. Sarker, "Deep cyber security: A comprehensive overview from neural network and deep learning perspec- tive," *SN Computer Science*, vol. 2, p. 154, 2021.

[31] W. Tounsi and H. Rais, "A survey on technical threat intelli- gence in the age of sophisticated cyber -attacks," *Computers & Security*, vol. 72, pp. 212–233, 2018.

[32] N. Sun, J. Zhang, P. Rimba, S. Gao, L. Y. Zhang, and Y. Xi- ang, "Data-driven cybersecurity incident prediction: A sur- vey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1744–1772, 2019.

[33] D. Schlette, M. Caselli, and G. Pernul, "A comparative study on cyber threat intelligence: The security incident response perspective," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2525–2556, 2021.

[34] T. D. Wagner, K. Mahbub, E. Palber, and A. Strainovic, "Cy- ber threat intelligence sharing: Survey and research directions," *Computers & Security*, vol. 87, p. 101589, 2019.

[35] MITRE Corporation, "MITRE ATT&CK: A globally- accessible knowledge base of adversary tactics and techniques," MITRE, Tech. Rep., 2023. [Online]. Available: https://attack.mitre.org

[36] C. Lawson and P. Firstbrook, "Innovation insight for ex- tended detection and response," *Gartner Research*, 2020, report ID: G00719984.

[37] MITRE Corporation, "MITRE ATLAS: Adversarial threat landscape for AI systems," MITRE, Tech. Rep., 2022. [Online].

Available: https://atlas.mitre.org

[38] F. N. Motlagh, M. Hajizadeh, M. Majd, P. Rashidi,T. Hwang, and A. M. Rahmani, "Large language mod- els in cybersecurity: State-of-the-art," *arXiv preprint arXiv:2402.00891*, 2024.

[39] H. Xu and Others, "Large language models for cyber security: A systematic literature review," *arXiv preprint arXiv:2405.04760*, 2024.

[40] J. Yao and Others, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, vol. 4, no. 2, p. 100211,2024.

[41] Y. Mirsky and W. Lee, "The creation and detection of deep- fakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1,pp. 1–41, 2022.

[42] Z. Li, D. Zou, S. Xu, X. Ou, H. Jin, S. Wang,Z. Deng, and Y. Zhong, "VulDeePecker: A deep learning- based system for vulnerability detection," *arXiv preprint arXiv:1801.01681*, 2018.

[43] European Union Agency for Cybersecurity (ENISA), "ENISA threat landscape 2023," ENISA, Tech. Rep., 2023. [Online]. Available: https://www.enisa.europa.eu/ publications/enisa-threat-landscape-2023

[44] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers in Computer Science*, vol. 3, p. 563060, 2021.

[45] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[46] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, 2019.

[47] Mandiant, "M-Trends 2023: Special report," Mandiant (Google Cloud), Tech. Rep., 2023. [Online]. Available: https://www.mandiant.com/m-trends

[48] Y. Shoshitaishvili, R. Wang, C. Salls, N. Stephens, M. Polino, A. Dutcher, J. Grosen, S. Feng, C. Hauser, C. Kruegel, and G. Vigna, "SOK: (State of) The Art of War: Offensive techniques in binary analysis," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2016, pp. 138–157.

[49] M. Rigaki and S. Garcia, "Bringing a GAN to a knife- fight: Adapting malware communication to avoid detec- tion," *arXiv preprint arXiv:1711.02442*, 2018.

[50] Z. Lin, Y. Shi, and Z. Xue, "IDSGAN: Generative adversar- ial networks for attack generation against intrusion detec- tion," *arXiv preprint arXiv:1809.02077*, 2018.

[51] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 8, pp. 3779–3795, 2023.

[52] D. Gibert, C. Matú, and J. Planes, "The rise of machine learning for detection and classification of malware: Re- search developments, trends and challenges," *Journal of Network and Computer Applications*, vol. 153, p. 102526, 2020.

[53] Microsoft, "Microsoft digital defense report 2023," Microsoft, Tech. Rep., 2023. [Online]. Available:https://www.microsoft.com/en-us/security/ security-insider/microsoft-digital-defense-report-2023

[54] D. R. McKinnel, T. Dargahi, A. Dehghantanha, and K.-K. R. Choo, "A systematic literature review and meta-analysis on artificial intelligence in penetration testing and vulnerability assessment," *Computers & Electrical Engineering*, vol. 75,pp. 175–188, 2019.

[55] D. Dasgupta, Z. Akhtar, and S. Sen, "Machine learning in cybersecurity: A comprehensive survey," *The Journal of Defense Modeling and Simulation*, vol. 19, no. 1, pp. 57–106, 2022.

[56] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, p. 20, 2019.

[57] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," *Applied Sciences*, vol. 9, no. 20, p. 4396, 2019.

[58] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning ap- proaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, p. e4150, 2021.

[59] R. A. Bridges, T. R. Glass-Vanderlan, M. D. Iannacone,M. S. Vincent, and Q. Chen, "A survey of intrusion detec- tion systems leveraging host data," *ACM Computing Sur- veys*, vol. 52, no. 6, pp. 1–35, 2019.

[60] C. Islam, M. A. Babar, and S. Nepal, "A multi-vocal review of security orchestration," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–45, 2021.

[61] R. Uetz, C. Hemminghaus, L. Hacklaender, P. Schlipper, andM. Henze, "SOCBED: A self-contained open-source cyber-attack experimentation testbed," *ACM Digital Threats: Re- search and Practice*, vol. 2, no. 3, pp. 1–17, 2021.

[62] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. De- hghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[63] M. Sewak, S. K. Sahay, and H. Rathore, "Deep reinforce- ment learning in the advanced cybersecurity threat detection and protection," *Information Systems Frontiers*, vol. 25, pp. 2039–2057, 2023.

[64] S. Zeadally, E. Adi, Z. Baig, and I. A. Khan, "Harnessing artificial intelligence capabilities to improve cybersecurity," *IEEE Access*, vol. 8, pp. 23 817–23 837, 2020.

[65] J.-h. Li, "Cyber security meets artificial intelligence: A sur- vey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 12, pp. 1462–1474, 2018.

[66] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[67] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, andD. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[68] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, andA. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the International Con- ference on Learning Representations (ICLR)*, 2018.

[69] M. Goldblum, D. Tsipras, C. Xie, X. Chen,A. Schwarzschild, D. Song, A. Madry, B. Li, and T. Goldstein, "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1563–1580, 2023.

[70] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Ce- lik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings of the IEEE European Symposium on Security and Privacy*, 2016, pp. 372–387.

[71] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila,P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pa- gallo, F. Rossi, and Others, "AI4People—an ethical frame- work for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4,pp. 689–707, 2018.

[72] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot,S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, and Others, "Explainable artificial intelli- gence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[73] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explain- able artificial intelligence in cybersecurity: A survey," *IEEE Access*, vol. 10, pp. 93 575–93 600, 2022.

[74] Google, "Secure AI framework (SAIF): A concep- tual framework for secure AI systems," Google, Tech. Rep., 2023. [Online]. Available: https://safety.google/ cybersecurity-advancements/saif/

[75] European Parliament, "Regulation (EU) 2024/1689 — the AI act," European Union, Tech. Rep., 2024. [Online].

Available: https://eur-lex.europa.eu/eli/reg/2024/1689

[76] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST, Tech. Rep. NIST AI 100-1, 2023.

[77] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, andL. Floridi, "Artificial intelligence and the 'good society': The US, EU, and UK approach," *Science and Engineering Ethics*, vol. 24, pp. 505–528, 2018.

[78] European Union Agency for Cybersecurity (ENISA), "AI cybersecurity challenges: Threat landscape for artificial intelligence," ENISA, Tech. Rep., 2021. [On- line]. Available: https://www.enisa.europa.eu/publications/ artificial-intelligence-cybersecurity-challenges

[79] M. Mosca, "Cybersecurity in an era with quantum comput- ers: Will we be ready?" *IEEE Security & Privacy*, vol. 16, no. 5, pp. 38–41, 2018.

[80] S. Pirandola, U. L. Andersen, L. Banchi, M. Berta,D. Bunandar, R. Colbeck, D. Englund, T. Gehring, C. Lupo, C. Ottaviani, and Others, "Advances in quantum cryptogra- phy," *Advances in Optics and Photonics*, vol. 12, no. 4, pp. 1012–1236, 2020.

[81] F. Liang, W. G. Hatcher, W. Liao, W. Gao, and W. Yu, "Machine learning for security and the internet of things: The good, the bad, and the ugly," *IEEE Access*, vol. 7, pp. 158 126–158 147, 2019.

[82] A. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-physical sys- tems security — a survey," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1802–1831, 2017.