Letters in High Energy Physics Volume 2024
ISSN: 2632-2714

Adversarial AI in National Security: Understanding and Countering AI-Generated Cyber Threats

Chiranjeevi Kunaparaju

Principal Site Reliability Engineer at Palo Alto Networks, Santa Clara California, United States

Email: chiranjeevirajukr@gmail.com

ORCID NO: https://orcid.org/0009-0004-0528-6973

Abstract

The rapid integration of artificial intelligence into cyber operations has transformed the threat landscape facing national security systems. Adversarial and generative AI techniques now enable attackers to automate reconnaissance, craft highly personalized social engineering campaigns, evade detection mechanisms, and directly compromise AI-enabled defensive systems. These developments challenge the effectiveness of traditional cybersecurity approaches that were designed for human-driven or rule-based attacks. This study addresses the growing need for a structured understanding of AI-generated cyber threats and defensible strategies to counter them within national security contexts.

Methodologically, the article adopts a structured analytical approach that synthesizes established adversarial machine learning literature, institutional risk management frameworks, and threat intelligence models. A taxonomy of AI-generated cyber threats relevant to national security is developed, followed by a systematic mapping of these threats to known adversarial AI techniques and stages of the cyber attack lifecycle. Building on this analysis, the study proposes a layered defensive framework that integrates governance and risk management controls, technical safeguards, and operational response mechanisms across the AI system lifecycle.

The key contributions of this work are threefold. First, it provides a consolidated taxonomy that clarifies how adversarial AI manifests across military, intelligence, and critical infrastructure systems. Second, it links adversarial techniques to concrete national security impacts, highlighting critical points of defensive failure. Third, it advances an integrated defense framework aligned with internationally recognized AI and cybersecurity standards.

The findings underscore that effective national security defense against adversarial AI requires coordinated governance, robust technical resilience, and adaptive operational capabilities. The proposed framework offers practical guidance for policymakers, defense institutions, and security practitioners seeking to strengthen resilience against AI-generated cyber threats in an evolving strategic environment.

Keywords: *Adversarial artificial intelligence, AI-generated cyber threats, national security, adversarial machine learning, cybersecurity risk management, threat intelligence frameworks, AI governance, critical infrastructure protection*

1. Introduction

1.1 Background and motivation

Artificial intelligence has become a foundational component of modern national security infrastructures, supporting functions such as cyber defense automation, intelligence analysis, surveillance, decision support, and protection of critical infrastructure. At the same time, the rapid diffusion of advanced AI capabilities has fundamentally altered the cyber threat landscape. Threat actors are increasingly leveraging AI systems to automate, scale, and personalize cyber operations in ways that were previously infeasible using human-driven methods alone (Brundage et al., 2018; ENISA, 2024).

A defining feature of this evolution is the rise of AI-enabled cyber threats that target both conventional digital systems and AI models deployed within national security

environments. These threats include adversarial attacks against machine learning systems, AI-assisted social engineering and phishing, automated vulnerability discovery, deepfake-enabled deception, and model extraction or inversion attacks that compromise sensitive data and intellectual property (Barreno et al., 2010; Papernot et al., 2016; Tolosana et al., 2020). As governments increasingly integrate AI into mission-critical systems, the attack surface expands to include not only software and networks but also training data, model architectures, inference pipelines, and AI supply chains.

Concurrently, cyber operations are undergoing a structural shift from predominantly human-driven attack pipelines to AI-augmented and semiautonomous workflows. Advances in generative models, reinforcement learning, and agentic AI allow attackers to rapidly generate malware variants, craft highly targeted phishing campaigns, probe systems for weaknesses, and adapt attack strategies in near real time (Buczak & Guven, 2016; Madry et al., 2018). This shift significantly reduces the cost, skill barrier, and time required to execute sophisticated cyber operations, thereby amplifying risks to military networks, intelligence systems, and national critical infrastructure.

1.2 Problem statement

Despite decades of progress in cybersecurity, traditional defense mechanisms are increasingly misaligned with the threat dynamics introduced by adversarial and generative AI. Conventional security controls such as signature-based detection, rule-driven intrusion detection systems, and static risk assessments are largely designed to counter known threats and predictable attack patterns. Adversarial AI attacks, by contrast, are adaptive, datadriven, and capable of exploiting the internal behavior of defensive models themselves (Carlini & Wagner, 2017; Athalye et al., 2018).

One major limitation lies in the vulnerability of machine learning systems to adversarial manipulation. Research has demonstrated that carefully crafted inputs can cause misclassification, evade detection, or degrade model performance without triggering traditional security alerts (Goodfellow et al., 2015; Szegedy et al., 2014). These weaknesses are particularly concerning in national security contexts, where AI systems may support threat detection, biometric identification, intelligence prioritization, or autonomous decision-making.

In addition to technical shortcomings, significant gaps persist in governance, detection, and response mechanisms for AIenabled cyber threats. Existing cybersecurity governance frameworks were not originally designed to account for AIspecific risks such as data poisoning, model inversion, or AI supply chain compromise. While emerging standards such as the

National Institute of Standards and Technology Artificial Intelligence Risk Management Framework provide high-level guidance, operational integration within national security cyber defense remains uneven (NIST, 2023; NIST, 2024).

Detection and response capabilities also lag behind the pace of adversarial innovation. AI-generated attacks often evolve dynamically, making them difficult to detect using static thresholds or predefined indicators of compromise. Moreover, incident response playbooks rarely account for scenarios in which AI models themselves are the attack target or attack vector. These gaps create systemic vulnerabilities that adversaries can exploit at scale.

1.3 Research objectives and contributions

In response to these challenges, this study aims to advance understanding of adversarial AI threats within national security cyber environments and to propose actionable countermeasures grounded in established standards and empirical research. The specific objectives of this research are threefold.

First, the study develops a structured taxonomy of AI-generated cyber threats relevant to national security systems. This taxonomy categorizes threat types based on attack objectives, AI techniques employed, targeted assets, and potential strategic impact, providing a common analytical language for researchers, practitioners, and policymakers.

Second, the research systematically maps adversarial AI techniques to national security contexts using recognized threat modeling frameworks, including the MITRE Adversarial Threat Landscape for ArtificialIntelligence Systems. This mapping clarifies how specific adversarial methods such as evasion, poisoning, model extraction,

and inference attacks manifest across military, intelligence, and critical infrastructure domains.

Third, the study proposes a layered defensive framework for countering AIgenerated cyber threats. The framework integrates governance and risk management controls, technical safeguards, and operational response mechanisms across the AI system lifecycle. By aligning cybersecurity practices with AI-specific risk management principles, the proposed approach seeks to enhance national security resilience against both current and emerging adversarial AI threats.

2. Conceptual and Theoretical Foundations

2.1 Adversarial Artificial Intelligence

Adversarial artificial intelligence refers to the study and exploitation of vulnerabilities in machine learning and artificial intelligence systems by malicious actors. In adversarial machine learning, attackers intentionally manipulate inputs, training data, model parameters, or system interfaces in order to cause erroneous, insecure, or unintended behavior. Unlike traditional software vulnerabilities, adversarial AI exploits statistical learning processes, making attacks difficult to detect using conventional security mechanisms (Barreno et al., 2010; Huang et al., 2011).

Adversarial AI security extends beyond model-level attacks to encompass the broader AI system, including data pipelines, deployment environments, application programming interfaces, and human decision interfaces. As emphasized in the National Institute of Standards and Technology guidance, AI systems must be understood as socio-technical systems rather than isolated algorithms, since failures can arise from interactions between models, data, users, and operational contexts (National Institute of Standards and Technology, 2023).

Attack surfaces in adversarial AI span the entire AI lifecycle. During data collection and preparation, attackers may conduct data poisoning by injecting malicious or biased samples that degrade model performance or introduce hidden behaviors. During training and validation, model manipulation and backdoor insertion can occur, particularly in outsourced or third-party model development pipelines (Gu et al., 2017). At deployment time, evasion attacks exploit carefully crafted inputs to bypass detection or induce misclassification, even when models perform well under standard evaluation conditions (Goodfellow et al., 2015; Carlini & Wagner, 2017). Postdeployment, attackers may exploit inference APIs to extract model parameters, infer sensitive training data, or reconstruct proprietary models, thereby undermining confidentiality and national security advantages (Tramèr et al., 2016; Shokri et al., 2017).

In national security contexts, these attack surfaces are particularly critical because AI systems are increasingly used for threat detection, intelligence analysis, surveillance, logistics, and decision support. Adversarial compromise of such systems can result in cascading operational failures, misinformed strategic decisions, or deliberate manipulation of military and intelligence processes.

2.2 AI Risk and Cybersecurity Governance

Effective governance of adversarial AI risks requires structured risk management approaches that integrate cybersecurity, safety, and organizational accountability.

The AI Risk Management Framework developed by the National Institute of Standards and Technology provides a lifecycle-based model for identifying, assessing, and mitigating risks associated with AI systems. The framework is organized around four core functions: govern, map, measure, and manage (National Institute of Standards and Technology, 2023).

Within this framework, adversarial AI risks are treated as both technical and systemic threats. Governance emphasizes accountability, role definition, and oversight mechanisms to ensure that AI security responsibilities are clearly assigned across organizations. The mapping function focuses on understanding AI system contexts, threat environments, and intended uses, which is essential for identifying adversarial exposure points. Measurement involves evaluating robustness, resilience, and vulnerability through testing, red teaming, and monitoring. Management focuses on risk response strategies, including mitigation, transfer, or acceptance, depending on mission criticality and threat severity.

Alignment with international standards further strengthens AI cybersecurity governance. ISO/IEC 23894 provides guidance on AI-specific risk management, emphasizing uncertainty, model opacity, and adaptive threat behavior. ISO/IEC 27001 complements this by establishing requirements for information security management systems, ensuring confidentiality, integrity, and availability of data and AI-enabled services. Together, these standards enable organizations to embed adversarial AI risk controls within existing cybersecurity and enterprise risk management structures rather than treating AI security as a standalone concern.

For national security institutions, such alignment supports interoperability, compliance, and assurance across defense agencies, intelligence organizations, and critical infrastructure operators. It also facilitates coordinated responses to crossborder AI-enabled cyber threats, which increasingly transcend national and organizational boundaries.

2.3 Threat Intelligence Frameworks for AI Systems

Threat intelligence frameworks play a central role in understanding, categorizing, and responding to adversarial AI threats. The MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems, commonly referred to as MITRE ATLAS, provides a structured knowledge base of adversarial tactics, techniques, and procedures targeting AI systems.

MITRE ATLAS extends traditional cyber threat modeling by focusing specifically on AI-related attack vectors, including data poisoning, model evasion, model extraction, and AI supply chain compromise. The framework organizes adversarial behavior into tactical phases such as reconnaissance, resource development, model manipulation, exploitation, and impact, enabling defenders to systematically map threats to controls and countermeasures.

In national security cyber operations, MITRE ATLAS is particularly relevant because it bridges the gap between abstract adversarial AI research and operational threat intelligence. Intelligence agencies and defense organizations can use ATLAS to align adversarial AI threats with missioncritical systems, assess adversary capabilities, and prioritize defensive investments. By integrating ATLAS with existing cyber frameworks and risk management processes, national security institutions gain a common language for analyzing AI-enabled threats and coordinating defensive actions across technical and organizational domains.

3. Taxonomy of AI-Generated Cyber Threats in National Security

3.1 Categories of AI-Enabled Threats

AI-generated cyber threats encompass a diverse set of attack categories that leverage artificial intelligence to enhance scale, precision, and adaptability. One prominent category is AI-assisted phishing and social engineering, where generative models are used to produce highly personalized and linguistically convincing messages that bypass traditional detection mechanisms. These attacks pose significant risks to military personnel, intelligence analysts, and government officials by increasing the likelihood of credential compromise and unauthorized access.

Another category involves automated vulnerability discovery and exploit generation. Machine learning models can be applied to code analysis, fuzzing, and vulnerability pattern recognition, enabling attackers to identify and weaponize software weaknesses more rapidly than manual methods. In national security environments, such automation can accelerate attacks against classified systems, defense supply chains, and critical infrastructure.

Adversarial attacks on defensive AI models represent a direct threat to AI-enabled security systems. These include evasion attacks against intrusion detection models, poisoning attacks against threat intelligence pipelines, and manipulation of decisionsupport systems used in intelligence and military planning. Successful attacks can degrade situational awareness and undermine trust in automated defense capabilities.

Deepfakes and synthetic identity threats constitute a growing class of AI-enabled cyber operations. Advanced generative models can create realistic audio, video, and biometric artifacts that impersonate trusted individuals, enabling disinformation, espionage, and unauthorized system access. These threats are particularly concerning for national security due to their potential impact on

command integrity, public trust, and strategic communication.

Finally, model extraction, inversion, and data poisoning attacks target the confidentiality and integrity of AI systems themselves. By exploiting inference interfaces or training data pipelines, adversaries can steal proprietary models, infer sensitive training data, or implant malicious behaviors. Such attacks threaten national security by eroding technological advantages and exposing sensitive operational data.

Table 1: Taxonomy of AI-Generated Cyber Threats Affecting National Security

Threat category	Attack objective	AI technique used	Targeted systems	National security impact
AI-assisted phishing and social engineering	Credential theft, unauthorized access	Natural language generation, user profiling	Government email systems, defense networks	Compromise of classified accounts and operational access
Automated vulnerability discovery and exploit generation	Rapid exploitation of software weaknesses	ML-based code analysis, reinforcement learning	Military software, critical infrastructure systems	Accelerated cyber intrusions and system disruption
Adversarial attacks on defensive AI models	Degrade detection or decision accuracy	Evasion attacks, data poisoning, adversarial examples	Intrusion detection systems, intelligence analytics	Loss of situational awareness and false threat assessments
Deepfakes and synthetic identity threats	Impersonation, deception, influence	Generative adversarial networks, diffusion models	Biometric systems, command communications	Erosion of trust, strategic misinformation
Model extraction and inversion attacks	Theft of models or sensitive data	API probing, inference attacks	Deployed AI services, cloud-based defense tools	Loss of technological advantage and data leakage
Data poisoning and AI supply chain compromise	Implant hidden behaviors or biases	Training data manipulation, backdoor insertion	AI training pipelines, third-party models	Long-term system compromise and covert manipulation

4. Adversarial AI Attack Lifecycle

Adversarial AI attacks in national security contexts follow a structured lifecycle in which artificial intelligence

amplifies attacker capability at every stage. Unlike conventional cyber operations, AI-enabled attacks scale rapidly, adapt dynamically, and exploit both technical and human vulnerabilities. Understanding this lifecycle is critical for designing effective countermeasures.

4.1 Reconnaissance and Targeting

The reconnaissance phase involves identifying high-value national security targets such as military command systems, intelligence databases, critical infrastructure control systems, and public communication platforms. AI significantly enhances this stage by automating data collection and analysis across open-source intelligence, social media, leaked datasets, and network telemetry.

Machine learning models are used to profile organizations, map digital infrastructure, identify exposed services, and infer user behavior patterns. Natural language processing enables large-scale analysis of documents and communications to extract sensitive operational details. In national security environments, this AI-driven reconnaissance increases precision in target selection while reducing the time and cost traditionally associated with intelligence gathering (Barreno et al., 2010; Brundage et al., 2018).

4.2 Weaponization and Automation

During weaponization, AI tools are employed to transform reconnaissance outputs into operational attack assets. Generative models enable the automated creation of highly convincing phishing content, malware variants, and exploit scripts tailored to specific targets.

Adversarial machine learning techniques are also used to design inputs that exploit weaknesses in defensive AI systems. For example, attackers may generate adversarial examples to bypass intrusion detection systems or manipulate AI-based authentication mechanisms. Automation allows attackers to test and refine attack payloads at scale, selecting the most effective variants before deployment (Goodfellow et al., 2015; Athalye et al., 2018).

4.3 Delivery and Exploitation

AI enhances the delivery phase by optimizing timing, channel selection, and payload customization. Reinforcement learning can be used to identify the most effective delivery strategies based on real-time feedback. Social engineering campaigns benefit from AI-generated personalization, increasing the likelihood of successful exploitation.

Exploitation may involve traditional software vulnerabilities or weaknesses in AI-enabled systems themselves. Examples include prompt injection in large language models, evasion of AI-based malware detectors, or exploitation of poorly secured AI APIs. In national security systems, successful exploitation can lead to unauthorized access to sensitive data or operational disruption (Carlini & Wagner, 2017; Papernot et al., 2016).

4.4 Persistence and Evasion

Once access is obtained, attackers use AI to maintain persistence while avoiding detection. Adaptive malware can modify its behavior based on observed defensive responses, making static detection methods ineffective. Adversarial techniques such as model evasion and inference attacks allow malicious activity to remain hidden within normal system behavior.

AI-driven evasion is particularly dangerous in environments that rely heavily on automated monitoring. Attackers can probe defensive models, learn their decision boundaries, and adjust attack strategies accordingly. This continuous adaptation complicates incident response and increases dwell time within critical systems (Madry et al., 2018; Shokri et al., 2017).

4.5 Impact and Strategic Consequences

The final stage involves achieving strategic objectives that extend beyond technical compromise. In national security contexts, impacts may include intelligence leakage, disruption of military operations, degradation of critical infrastructure, or erosion of public trust through misinformation and deepfakes.

AI amplifies the scale and speed of these impacts, enabling coordinated campaigns that combine cyber intrusion with information operations. The strategic consequences may include escalation risks, loss of deterrence credibility, and long-term damage to institutional resilience (Scharre, 2018; Tolosana et al., 2020).

Figure 1: Diagram of AI-Enabled Cyber Attack Lifecycle for National Security Systems

This diagram presents a structured lifecycle of AI-enabled cyber attacks targeting national security systems, encompassing reconnaissance and targeting, weaponization and automation, delivery and exploitation, persistence and evasion, and impact and strategic consequences. It highlights how artificial intelligence amplifies attacker capabilities at each stage by increasing automation, adaptability, and scale, thereby intensifying operational complexity and strategic risk for military, intelligence, and critical infrastructure environments.

5. Mapping Threats to Adversarial Techniques and Impact

5.1 Adversarial Techniques Across the AI Lifecycle

Adversarial techniques can be categorized based on how they interact with AI systems during different stages of the attack lifecycle. These include evasion attacks that manipulate inputs at inference time, poisoning attacks that corrupt training data, model extraction attacks that steal proprietary models, and inference attacks that reveal sensitive information about training data or system behavior.

In national security systems, these techniques target both offensive and defensive AI models, undermining trust in automated decision-making and situational awareness.

5.2 Alignment With Known Adversarial ML Attacks

Established adversarial machine learning research provides a framework for understanding how AI-generated cyber threats operate. Evasion attacks align with adversarial examples that cause misclassification. Poisoning attacks compromise model integrity during training or updates. Model extraction attacks enable adversaries to replicate sensitive defense models, while inference attacks threaten data confidentiality.

Mapping real-world AI-enabled cyber threats to these categories allows security practitioners to systematically assess risk and design targeted defenses (Huang et al., 2011; Tramèr et al., 2016; Fredrikson et al., 2015).

5.3 Implications for Military, Intelligence, and Critical Infrastructure Systems

Military systems face risks related to compromised decision-support tools and autonomous platforms. Intelligence agencies risk exposure of sensitive data and analytical methods. Critical infrastructure operators face threats to availability, safety, and public confidence.

The convergence of adversarial AI and cyber operations increases the likelihood of cascading failures across sectors, making cross-domain risk assessment and coordination essential. Frameworks such as those developed by National Institute of Standards and Technology and MITRE provide structured approaches for managing these risks at scale.

Table 2: Mapping AI-Generated Cyber Threats to Adversarial AI Techniques

Threat type	Evasion	Poisoning	Model extraction	Inference attacks
AI-generated phishing	✓			✓
Adversarial malware	✓			
Compromised defensive AI	✓	✓	✓	
Deepfake identity attacks	✓			
AI model supply chain attacks				✓
Surveillance and intelligence inference				✓

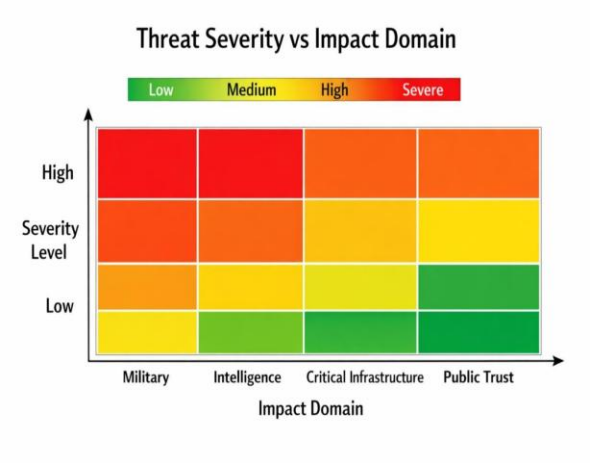

Figure 2: Threat Severity vs Impact Domain

This chart presents a comparative assessment of the severity of AI-generated cyber threats across four critical impact domains: military, intelligence, critical infrastructure, and public trust. The visualization highlights that military and intelligence systems experience consistently higher severity levels due to their strategic value and reliance on automated decision-support systems. In contrast, impacts on public trust escalate rapidly through AIenabled misinformation and influence operations, demonstrating how non-kinetic effects can generate significant national security consequences even without direct system

compromise

6. Defensive and Mitigation Strategies

The growing integration of artificial intelligence into national security systems requires a comprehensive defensive posture that addresses not only conventional cyber risks, but also threats arising from adversarial manipulation of AI models and AI-generated attack capabilities. Effective mitigation therefore depends on a multilayered strategy combining governance mechanisms, technical safeguards, and operational preparedness across the entire AI lifecycle.

6.1 Governance and Risk Management Controls

- **AI RMF lifecycle alignment:**

Governance forms the foundation of any national security response to adversarial AI. The Artificial Intelligence Risk Management Framework developed by National Institute of Standards and Technology provides a structured lifecycle approach that is directly applicable to adversarial threat mitigation. The framework emphasizes governance, mapping, measurement, and management of AI risks throughout system design, development, deployment, and operation. In the context of national security, lifecycle alignment ensures that adversarial risks such as data poisoning, model evasion, and unauthorized model reuse are identified early and reassessed continuously. Risk assessments are not treated as static compliance exercises, but as iterative processes that evolve alongside threat intelligence and operational feedback. This approach supports traceability, accountability, and defensibility of AI-enabled security decisions.

- **Procurement and assurance requirements:** Procurement policies represent a critical governance control for reducing exposure to adversarial AI threats. National security institutions increasingly rely on third-party AI components, pretrained models, and data pipelines. Without formal assurance requirements, these dependencies introduce opaque risks into mission-critical systems. Governance controls therefore require that AI procurement processes mandate documented model provenance, training data integrity assurances, and independent evaluation against adversarial threat scenarios. Alignment with international standards such as ISO/IEC 23894 and ISO/IEC 27001 further strengthens institutional oversight by embedding AI risk management within existing information security governance structures. These measures reduce the likelihood of supply-chain compromise and establish accountability across vendors and integrators.

6.2 Technical Defenses

- **Adversarial training:** Adversarial training is a core technical defense aimed at improving model robustness against evasion and manipulation attacks. By systematically exposing AI models to adversarial inputs during training and validation, defenders can reduce sensitivity to carefully crafted perturbations and malicious prompt

manipulation. While adversarial training does not eliminate all vulnerabilities, it increases the cost and complexity of successful attacks, particularly in high-stakes national security environments where reliability and predictability are essential. Its effectiveness is maximized when combined with threat-specific testing informed by known adversarial techniques documented in frameworks such as MITRE ATLAS.

- **Robust model evaluation:** Robust evaluation extends beyond standard accuracy metrics to include stress testing under adversarial conditions. National security AI systems must be evaluated against worst-case scenarios, including adaptive attackers who iteratively probe model behavior. Techniques such as redteam testing, adversarial benchmarking, and robustness audits allow defenders to identify brittle decision boundaries and unintended behaviors before deployment. These evaluations should be conducted periodically and after any system update, recognizing that changes to data, models, or operational context can introduce new vulnerabilities.

- **Monitoring and logging:** Continuous monitoring and comprehensive logging are essential for detecting adversarial activity targeting AI systems. Behavioral anomalies, distribution shifts, and unexpected output patterns can serve as early indicators of model exploitation or data integrity compromise. For national security applications, monitoring infrastructures must be tightly integrated with security operations centers, enabling rapid correlation between AI system telemetry and broader cyber threat intelligence. Logging also supports post-incident analysis, accountability, and legal defensibility in high-impact security incidents.

- **Secure AI supply chains:** AI supply chain security addresses risks introduced through datasets, pretrained models, development tools, and deployment infrastructure. Adversarial actors may exploit weak controls to insert backdoors, poisoned data, or malicious dependencies into AI pipelines. Mitigation strategies include strict access controls, cryptographic integrity checks, version control, and segregation of development and production environments. These controls mirror established software supply-chain security practices, adapted to the unique characteristics of AI systems and model artifacts.

6.3 Operational and Organizational Controls

- **Incident response adaptation:** Traditional incident response frameworks must be adapted to account for AI-specific failure modes. Adversarial AI incidents may not manifest as system outages, but as subtle degradation in decision quality or trustworthiness. Incident response plans therefore need explicit procedures for AI model isolation, rollback, retraining, and forensic analysis. Integrating AI expertise into incident response teams ensures that operational decisions consider both cyber and algorithmic dimensions of an attack, reducing response time and limiting downstream impact.

- **Human-AI teaming:** Human-AI teaming is a critical safeguard against over-reliance on automated decision systems. In national security contexts, human oversight provides contextual judgment, ethical reasoning, and adaptive thinking that remain difficult to replicate through automation alone. Well-designed teaming structures ensure that AI outputs inform, rather than replace, human decision-making. This reduces the risk of adversarial manipulation cascading directly into strategic or operational errors.

- **Workforce readiness:** Effective defense against adversarial AI requires a workforce that understands both cybersecurity and AI system behavior. Training programs must therefore extend beyond traditional cyber hygiene to include adversarial machine learning concepts, model evaluation

techniques, and AI risk governance. early detection of emerging threats Workforce readiness strengthens and informed response to AI-related institutional resilience by enabling incidents.

Table 3: Defense Matrix for Countering AI-Generated Cyber Threats

Function	Technical Controls	Governance Controls	Operational Controls
Prevention	Adversarial training, secure model design, supply-chain integrity checks	AI risk policies, procurement assurance requirements	Secure development practices, workforce training
Detection	Monitoring, anomaly detection, model performance auditing	Risk assessment reviews, compliance reporting	SOC integration, threat intelligence analysis
Response	Model isolation, rollback, retraining mechanisms	Incident governance procedures, accountability lines	AI-aware incident response teams
Recovery	Model revalidation, system hardening	Post-incident review, policy updates	Lessons-learned integration and readiness drills

7. Integrated National Security Defense Framework

7.1 Layered defense concept

An integrated defense against adversarial AI threats is best conceptualized as a layered model in which governance, technical, and operational controls reinforce one another. No single layer is sufficient on its own. Instead, resilience emerges from the interaction between policy oversight, system-level safeguards, and humancentered operational practices.

7.2 Integration of governance, technical, and operational layers

Governance mechanisms define acceptable risk thresholds, accountability structures, and assurance requirements. Technical safeguards translate these policies into enforceable system behaviors, while operational controls ensure that human actors can without abandoning existing security architectures. effectively interpret, manage, and respond to AI-driven threats.

This integration reduces gaps between policy intent and operational reality, ensuring that adversarial risks are addressed consistently across organizational boundaries.

7.3 Strategic alignment with national cybersecurity doctrine

The layered defense framework aligns with established national cybersecurity doctrines that emphasize defense-in-depth, resilience, and coordinated response. By extending these principles to AI systems, national security institutions can incorporate adversarial AI considerations

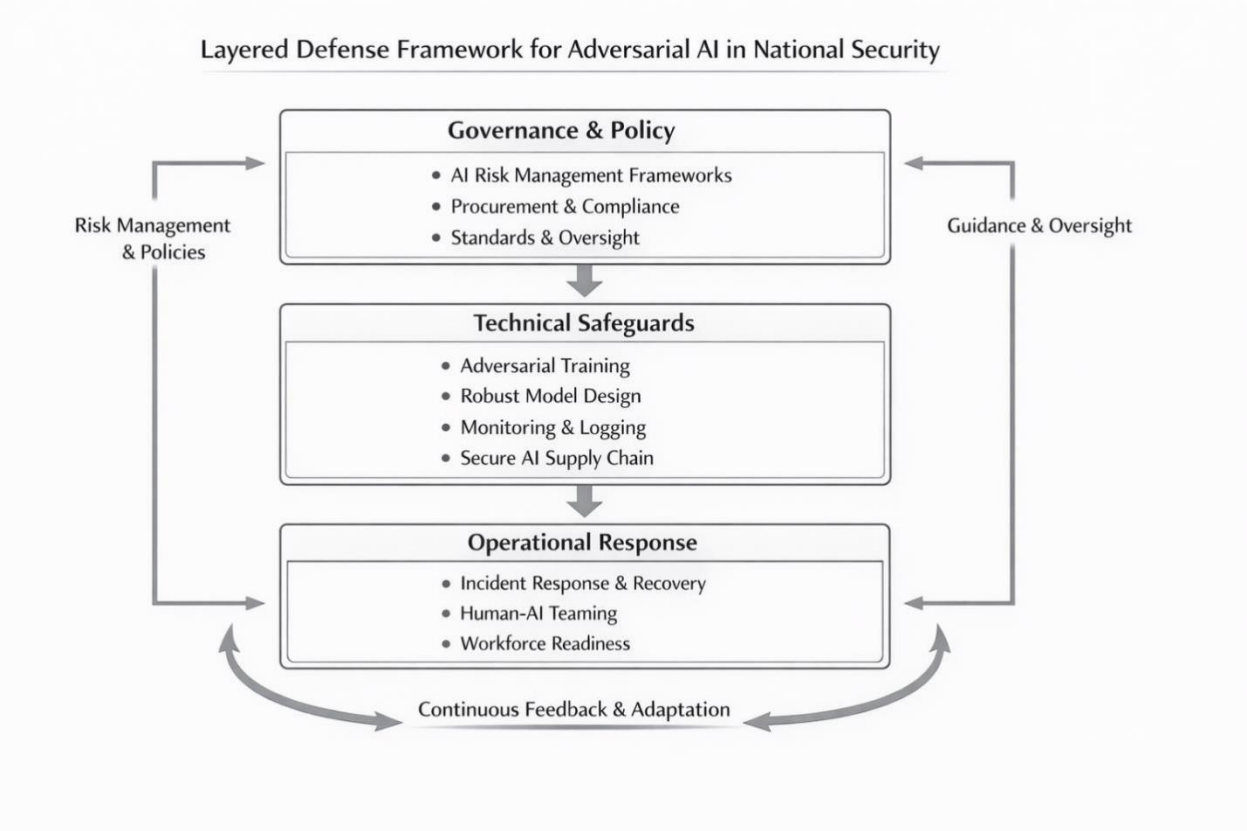

Figure 2: Layered Defense Framework for Adversarial AI in National Security

This diagram presents a three-layer, defense-in-depth framework for mitigating adversarial AI threats in national security contexts. The governance and policy layer establishes risk management, standards compliance, and oversight mechanisms that guide AI use. The technical safeguards layer translates these policies into robust system-level protections, including adversarial training, monitoring, and secure AI supply chains. The operational response layer focuses on real-world implementation through incident response, human-AI teaming, and workforce readiness. Continuous feedback loops across layers enable adaptive improvement and sustained resilience against evolving AI-generated cyber threats.

8. Policy, Ethical, and Strategic Implications

8.1 Dual-use risks of AI technologies

Artificial intelligence technologies deployed for national security and cybersecurity purposes exhibit strong dual-use characteristics. Techniques designed to enhance defensive capabilities, such as automated vulnerability discovery, threat intelligence correlation, and behavioral anomaly detection, can be repurposed by adversaries to scale cyber attacks, automate reconnaissance, and evade detection. This dual-use nature complicates governance, as restrictions intended to limit malicious exploitation may simultaneously constrain legitimate defensive innovation (Brundage et al., 2018).

From a national security perspective, the accessibility of advanced AI models and open research dissemination lowers the barrier to entry for state and non-state actors seeking to conduct sophisticated cyber operations. Adversarial machine learning techniques, including model evasion, poisoning, and extraction, further exacerbate this risk by enabling attackers to directly undermine defensive AI systems themselves (Barreno et al., 2010; Huang et al., 2011). As a result, policymakers face the challenge of balancing openness

and collaboration with the need for strategic control over sensitive AI capabilities.

Risk-based governance frameworks, such as the NIST Artificial Intelligence Risk Management Framework, provide a structured approach to addressing dual-use concerns by emphasizing contextual risk assessment, impact analysis, and continuous monitoring across the AI lifecycle (National Institute of Standards and Technology, 2023, 2024). However, effective implementation requires sector-specific interpretation for national security environments, where threat tolerance and consequences differ substantially from civilian domains.

8.2 Accountability and oversight challenges

The integration of AI into national security cyber operations introduces significant accountability and oversight challenges. AIenabled systems often operate with a high degree of autonomy and complexity, making it difficult to attribute decisions, errors, or failures to specific actors or processes. This opacity complicates traditional mechanisms of responsibility, particularly in cases where AI-driven defenses or countermeasures produce unintended consequences.

Oversight challenges are further amplified when AI systems are trained on large, heterogeneous datasets sourced from multiple agencies or external partners. Issues such as data provenance, model versioning, and supply chain integrity become critical, especially given documented risks associated with data poisoning and backdoored models (Gu et al., 2017; Fredrikson et al., 2015). Without robust auditability and documentation, assessing compliance with legal and ethical standards becomes increasingly difficult.

Institutional frameworks, including national cybersecurity strategies and international standards such as ISO/IEC 23894 and ISO/IEC 27001, emphasize the importance of governance structures, documentation, and risk ownership (International Organization for Standardization, 2022, 2023). For national security applications, these mechanisms must be complemented by independent oversight bodies, clear chains of command, and predefined escalation procedures to ensure that AIenabled cyber operations remain aligned with democratic accountability and the rule of law.

8.3 International coordination and norms

Adversarial AI threats transcend national boundaries, making international coordination a strategic necessity rather than an option. AI-generated cyber threats can propagate rapidly across interconnected systems, targeting critical infrastructure, defense networks, and public trust simultaneously. Unilateral approaches to AI governance risk creating regulatory fragmentation that adversaries can exploit.

International efforts to establish norms for responsible AI use in security contexts remain fragmented. While global initiatives and standards bodies have begun addressing AI risk management, there is no comprehensive, binding framework governing the military and intelligence use of AI in cyberspace (Chinen, 2025). This absence of consensus increases the risk of escalation, misinterpretation, and unintended conflict in cyberspace.

Strategic coordination among allied states is therefore essential. Shared threat intelligence, harmonized risk assessment methodologies, and joint research initiatives can improve collective resilience against adversarial AI threats. Aligning national frameworks with internationally recognized standards and threat taxonomies, such as those provided by NIST and ENISA, offers a practical pathway toward convergence while preserving national sovereignty (European Union Agency for Cybersecurity, 2024).

9. Limitations and Future Research Directions

9.1 Data availability and measurement challenges

A key limitation in the study of adversarial AI in national security contexts is the scarcity of reliable, high-quality data. Many AI-enabled cyber incidents involve classified systems or sensitive operations, limiting public disclosure and empirical analysis. As a result, existing research often relies on simulated environments, proof-of-concept attacks, or partial datasets that may not fully capture real-world threat dynamics.

Measurement challenges also arise from the evolving nature of AI-generated threats.

Traditional cybersecurity metrics, such as intrusion detection accuracy or incident frequency, may be

insufficient to evaluate adaptive, learning-based adversaries. Developing standardized metrics that account for adversarial adaptation, system resilience, and strategic impact remains an open research problem.

9.2 Benchmarking adversarial AI threats

The absence of standardized benchmarks for adversarial AI threats presents a significant obstacle to comparative evaluation and policy decision-making. While the machine learning community has proposed benchmarks for adversarial robustness, these are often narrowly focused on model performance and do not reflect operational constraints or national security requirements (Carlini & Wagner, 2017; Madry et al., 2018).

Future research should focus on developing benchmark frameworks that integrate technical robustness with system-level considerations, such as mission criticality, cascading failures, and human-machine interaction. Incorporating threat taxonomies and lifecycle models into benchmarking efforts could improve their relevance for defense and intelligence organizations.

9.3 Need for real-world national security deployment studies

There is a clear need for empirical studies examining the deployment of AI-based cybersecurity systems in real national security environments. Most existing literature focuses on laboratory experiments or commercial settings, leaving a gap in understanding how adversarial AI behaves under operational conditions, including resource constraints, legacy systems, and complex organizational structures.

Collaborative research between academia, government, and industry is essential to address this gap. Controlled pilot deployments, red-team exercises, and postincident analyses can provide valuable insights into the effectiveness and limitations of AI-enabled defenses, while respecting security and confidentiality requirements.

10. Conclusion

This study examined the growing role of adversarial artificial intelligence in shaping national security cyber threats and defenses. By synthesizing insights from adversarial machine learning research, cybersecurity governance frameworks, and threat intelligence literature, the article highlighted how AI both amplifies offensive cyber capabilities and introduces new vulnerabilities into defensive systems. The primary contribution of this work lies in its integrated perspective. Rather than treating AI-generated cyber threats as isolated technical problems, the study framed them as socio-technical and strategic challenges that require coordinated responses across governance, technology, and operations. The analysis underscores the importance of risk-based AI governance, robust accountability mechanisms, and international cooperation in mitigating adversarial AI risks.

From a policy and strategic standpoint, the findings suggest that national security institutions should prioritize lifecycle-based AI risk management, invest in adversarial testing and monitoring capabilities, and actively engage in international norm-setting initiatives. As AI continues to evolve, proactive and coordinated action will be essential to ensure that its integration into national security systems enhances resilience rather than creating new strategic vulnerabilities.

References

1. AI, N. (2023). Artificial intelligence risk management framework (AI RMF 1.0). URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai, 100-1.
2. Athalye, A., Carlini, N., & Wagner, D. (2018, July). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In International conference on machine learning (pp. 274-283). PMLR.
3. Barreno, M., Nelson, B., Joseph, A.D., & Tygar, J. D. (2010). The security of machine learning. Machine learning, 81(2), 121-148.
4. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D.

(2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

5. Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials, 18(2), 1153-1176.
6. Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp) (pp. 39-57). Ieee.
7. Carlini, N., & Wagner, D. (2017, November). Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM workshop on artificial intelligence and security (pp. 3-14).
8. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physicalworld attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 16251634).
9. Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1322-1333).
10. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
11. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733.
12. Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011, October). Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence (pp. 43-58).
13. International Organization for Standardization/International Electrotechnical Commission. (2023). ISO/IEC 23894: 2023—Information technology—Artificial intelligence— Guidance on risk management.
14. Landscape, E. T. (2021). European Union Agency for Cybersecurity. URL: https://www. enisa. europa. eu/topics/cyber-threats/threats-andtrends.
15. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
16. Oprea, A., & Vassilev, A. (2023). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (No. NIST Artificial Intelligence (AI) 100-2 E2023 (Withdrawn)). National Institute of Standards and Technology.
17. OWASP, T. (2023). OWASP top 10 for large language model applications. 27. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings. In 2016 IEEE European symposium on security and privacy (EuroS&P) (pp. 372-387). IEEE.
18. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.
19. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
20. Taorui Guan, “Evidence-Based Patent Damages,” 28 Journal of Intellectual Property Law (2020), 161.
21. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & OrtegaGarcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.
22. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction {APIs}. In 25th USENIX

security symposium (USENIX Security 16) (pp. 601-618).

23. Uppuluri, V. (2019). The Role of Natural Language Processing (NLP) in Business Intelligence (BI) for Clinical Decision Support. ISCSITR-INTERNATIONAL JOURNAL OF BUSINESS INTELLIGENCE (ISCSITR-IJBI), 1(2), 1-21.