AI/ML in Cybersecurity

Opportunities and Risks for U.S. Federal Agencies

Brian McElyea, PhD & Gregory D. Miller Jr

6/24/25

Table of Contents

Abstract

Artificial Intelligence (AI) and Machine Learning (ML) are rapidly transforming U.S. federal cybersecurity operations, enabling rapid analysis of massive data streams and more adaptive defenses critical for national and economic security. This whitepaper discusses how U.S. federal agencies are leveraging AI/ML to enhance threat detection, user behavior analytics, and automated incident response, while also addressing the accompanying risks and challenges. We explain the AI/ML lifecycle – from model training and deployment to ongoing evaluation – in the context of cybersecurity, referencing common frameworks and tools used in federal environments. Concrete examples from agencies such as the Cybersecurity and Infrastructure Security Agency (CISA), the Department of Homeland Security (DHS), the National Security Agency (NSA), the Department of Defense (DoD), and the Government Accountability Office (GAO) illustrate current applications. We also examine risks, including adversarial machine learning (efforts to deceive or subvert AI systems), data poisoning, and model drift, and discuss hybrid architectures that integrate rule-based logic with ML models for robust cyber defense. Finally, we present strategic recommendations – grounded in technical rationale and best practices – to guide federal cybersecurity leaders in the responsible adoption of AI/ML. The goal is to inform decision-makers on balancing AI/ML's opportunities for stronger cyber defense with diligent risk management, in alignment with federal standards and mission needs.

Introduction

The digital frontier is the modern battlefield, and federal agencies are at the forefront. As adversaries weaponize automation, our cyber defense capabilities must undergo a radical transformation. Artificial Intelligence (AI) and Machine Learning (ML) are at the core of this transformation, providing the power to convert massive datasets into proactive defense and preempt threats far faster than human analysts.

The federal government has formally recognized this reality; Executive Order 14110 emphasizes the secure and trustworthy use of AI, underscoring its potential to strengthen national and economic security. This paper moves beyond policy to practical implementation, providing leaders with a framework for integrating these powerful tools. We analyze the proven opportunities and critical vulnerabilities of AI in a federal context, offering strategic recommendations aligned with emerging best practices, such as the NIST AI Risk Management Framework. This document is a guide for building a more intelligent, resilient, and effective national cyber defense.

AI/ML Lifecycle in Cybersecurity Applications

To navigate the complexities of federal AI deployment, a single, authoritative lifecycle model is required. While various models exist, such as the privacy-focused framework from the General Services Administration (GSA), drawing from guidance across CISA, NSA, and NIST—including the AI Risk Management Framework and AI security principles—this paper presents a synthesized six-stage lifecycle tailored to cybersecurity applications. The six stages are:

1. **Plan & Design**
2. **Collect & Process Data**
3. **Build & Use Model**
4. **Verify & Validate**
5. **Deploy & Use**
6. **Operate & Monitor**

The development of this model is a strategic choice. Comparing the GSA's process-oriented model ("Identify a Problem -> Gather Data -> Create & Test") with the CISA, NSA, and NIST frameworks reveals a fundamental evolution in the government's approach to AI. The GSA model is suited for general-purpose AI development, emphasizing stakeholder engagement and privacy considerations. The CISA and NSA model, however, is security-centric by design, structured to counter specific, identified threats to AI systems. This deliberate shift reflects a maturation in the federal understanding of AI, from a technology to be *managed* to a potential attack surface to be *defended*. The very structure of the lifecycle thus becomes primary security

control, making it indispensable for agencies developing or deploying AI for national security, critical infrastructure protection, and cybersecurity missions.

Table 1 provides a unified overview of the six stages of the Federal AI Lifecycle, integrating the technical stages with their corresponding security concerns, risk management functions, and overarching governance mandates.

Federal Lifecycle Stage	Key Activities & Objectives	Primary Security Concerns (per CISA/NSA)	Applicable NIST RMF Functions	Key Federal Mandates & Governance
1. Plan & Design	This stage establishes the strategic, ethical, and security foundation for the AI system. A cross-functional team defines mission objectives, conducts initial risk assessments, and, most critically, determines if the system qualifies as "high-impact." This classification dictates the level of rigor, testing, and oversight required for all subsequent stages. Establish a governance team.	Data Supply Chain	Govern, Map	High-Impact AI determination (OMB M-25-21[1]). AI Impact Assessment. Establishment of AI Governance Board and CAIO (OMB M-24-10).
2. Collect & Process Data	This stage is an active defense against data-centric threats, ensuring data integrity, security, and fairness. Core activities include sourcing data from trusted origins, tracking provenance with secure ledgers, and applying technical controls like encryption and access management. This phase requires an adversarial mindset to defend the data supply chain against threats like data poisoning. Encrypt and classify data.	Data Supply Chain, Maliciously Modified Data	Govern, Map, Manage	Secure data sourcing and provenance tracking. Compliance with data privacy and protection laws (OMB M-22-18/EO 14028).

[1] Note: Agencies should consult the latest from OMB, OSTP, and NIST for authoritative mandates.

Federal Lifecycle Stage	Key Activities & Objectives	Primary Security Concerns (per CISA/NSA)	Applicable NIST RMF Functions	Key Federal Mandates & Governance
3. Build & Use Model	Here, prepared data is used to build and refine the AI model within a secure, controlled environment. This stage represents a convergence of MLOps and DevSecOps into a "Secure MLOps" framework, emphasizing secure development environments (e.g., GovCloud, air-gapped systems), vetted tools (e.g., TensorFlow, PyTorch), and integrated defenses against adversarial attacks. Implement defenses against adversarial ML.	Data Supply Chain, Maliciously Modified Data	Govern, Manage	Secure development practices. Supply chain risk management for ML libraries (OMB M-22-18).
4. Verify & Validate	This stage is the final quality and security gate before deployment, implementing the Measure function of the NIST AI RMF. It involves comprehensive Test, Evaluation, Verification, and Validation (TEVV) to assess accuracy, robustness, and fairness. For high-impact systems, this includes mandatory pre-deployment testing in realistic environments and AI Red Teaming—a structured, adversarial effort to find flaws and vulnerabilities. Conduct AI red teaming.	Data Supply Chain, Maliciously Modified Data	Govern, Measure, Manage	Mandatory pre-deployment testing for high-impact AI. Red teaming requirements (EO 14110).
5. Deploy & Use	This stage transitions the validated model into the live operational environment. It requires implementing secure deployment patterns like containerization on trusted infrastructure, which may include GovCloud, on-premises servers, or edge devices for real-time processing. The entire system must comply with federal IT security controls (e.g., NIST SP 800-53) and data protection mandates. Protect government data from misuse.	Data Supply Chain, Maliciously Modified Data, Data Drift	Govern, Manage	Secure deployment on trusted infrastructure. Contractual clauses protecting government data (OMB M-25-21).

Federal Lifecycle Stage	Key Activities & Objectives	Primary Security Concerns (per CISA/NSA)	Applicable NIST RMF Functions	Key Federal Mandates & Governance
6. Operate, Monitor & Maintain (MLOps)	The AI lifecycle does not end at deployment. This continuous phase, known as MLOps, focuses on sustaining model performance and trustworthiness. Key activities include monitoring for model drift (performance degradation) and data drift (changes in input data), using AI observability tools to understand *why* performance changes, and retraining models as needed. This operationalizes federal mandates for ongoing monitoring and human oversight. Update AI Use Case Inventory and retrain models as needed.	Data Supply Chain, Maliciously Modified Data, Data Drift	Govern, Map, Measure, Manage	Ongoing monitoring for high-impact AI (OMB M-25-21). Annual AI Use Case Inventory reporting (EO 13960).

Table 1: Six Stages of the Federal AI Lifecycle for Cybersecurity

Risks and Challenges of AI/ML in Cybersecurity

While AI/ML offers significant advantages in cybersecurity, it also presents substantial risks and challenges that federal agencies must address. These risks range from inherent vulnerabilities in machine learning to operational and governance issues, such as a lack of transparency and data integrity concerns. As Executive Order 14110 warns, irresponsible AI use can "exacerbate societal harms and pose risks to national security." Cybersecurity leaders must understand these challenges to effectively mitigate them.

Adversarial Machine Learning and Evasion Attacks

Adversaries can directly target AI/ML systems to evade detection or cause them to fail. This field, adversarial machine learning (AML), is particularly critical in cybersecurity, where machine learning (ML) models face intelligent, adaptive adversaries.

Evasion attacks are a common threat, where attackers craft inputs to fool an ML model into misclassifying malicious data as benign. For example, a slightly modified malware sample, undetectable to humans, could cause an AI-based classifier to ignore it. Similarly, subtle changes to a phishing email's text or metadata could prevent an ML spam filter from flagging it. NIST researchers note that AI systems can "malfunction when exposed to untrustworthy data," and attackers actively seek ways to confuse AI decision-making.

Another evasion technique is model probing, where attackers send numerous queries to an AI system to learn its decision boundaries. This allows them to tailor attacks to fall just below the detection threshold. The NSA's AI Security Center emphasizes that adversaries are trying to "exploit AI technologies to compete with us and do us harm," highlighting the need to defend AI models from such manipulation. An undetected evasion in cybersecurity could lead to a breach that AI was intended to prevent, undermining trust in the system.

Data Poisoning and Data Integrity Risks

Data poisoning involves adversaries manipulating training or input data over time to corrupt a model's behavior. If malicious data is inserted into an agency's training pipeline or continuous learning process, the model can be "poisoned" to mislearn patterns. For instance, feeding crafted fake network logs could teach a threat detection model to ignore certain malicious activities. NSA and CISA guidance identifies "maliciously modified ('poisoned') data" as a key risk to AI data security.

Poisoning can occur during training (if datasets are insecure or come from manipulated sources, such as public threat feeds) or at runtime (if the model learns online). In either case, the model's integrity is compromised, leading to systematically incorrect outputs that benefit the attacker. A poisoned cybersecurity model might consistently fail to detect an attacker's malware or incorrectly classify their IP addresses as safe.

Detecting poisoning attacks can be difficult due to subtle, distributed changes in data. NIST's taxonomy warns that "the data itself may not be trustworthy" and highlights numerous opportunities for bad actors to corrupt data throughout an AI system's lifecycle. A real-world scenario involves attackers polluting threat intelligence feeds or logging systems to mislead AI-driven analysis. The consequences are severe, potentially creating blind spots in agency defenses or eroding the accuracy of automated threat scoring, leading to a false sense of security. Securing the AI data supply chain through measures like data validation, provenance tracking, and robust training techniques is paramount. Best practices from NSA and CISA include using digital signatures and trusted pipelines for data authenticity and integrity.

Model Drift and Erosion of Efficacy

Even without malicious interference, ML models face model drift (also known as concept drift or data drift). This refers to the gradual input of data change over time, resulting in a decline in the model's predictive performance. In cybersecurity, the threat landscape and "normal" behavior evolve rapidly, with new network traffic types, shifting user work patterns (e.g., remote work), and constant attacker innovation. A model trained using past data may not perform well on current realities.

For example, an intrusion detection model trained on old network usage profiles might flag many false positives or miss new attack techniques when an agency moves applications to the cloud, significantly changing regular traffic. Experts consider drift "a natural and expected occurrence" over an AI system's life. If left unaddressed, small degradations in performance can snowball into substantial accuracy reductions, meaning the model's performance quietly worsens. In a

security context, this could lead to steadily declining detection rates or an increase in false alarms, undermining the system's effectiveness.

Distinguishing benign drift from malicious poisoning is crucial. NSA and CISA guidance suggests using continuous monitoring of accuracy and performance: slow, gradual changes likely indicate drift, while abrupt, dramatic changes may suggest a deliberate attack. Agencies should constantly evaluate models on recent data and set thresholds for retraining. Regular retraining on fresh data is a primary mitigation for drift, but it must be done carefully to avoid ingesting poisoned data.

The risk of drift also routine updates, highlighting the importance of regular model updates. A model cannot be deployed and forgotten; agencies need processes to audit model outputs, recalibrate models, and incorporate new threat information. Combating drift is an ongoing battle, but it's winnable through vigilant monitoring and maintenance.

Other Challenges: Explainability, Bias, and Over-Reliance

Several other significant challenges merit attention:

- **Lack of Explainability:** Many AI/ML models, especially deep learning models, act as "black boxes," providing little insight into why a specific decision was made. In federal agencies, this opacity is problematic, as decision-makers and oversight bodies often demand justification for actions, particularly in high-impact decisions (e.g., isolating a critical system). The inability to explain AI decisions can erode trust, complicate incident investigations, and hinder model debugging. For example, if an AI falsely flags normal user behavior as an insider threat, without explainability, it's hard to identify the cause. This risk can lead to either over-trusting a flawed model or ignoring a useful one due to a lack of transparency. Efforts are underway to improve explainability, but it remains a challenge.
- **Bias and False Positives/Negatives:** If training data or model design is biased, the AI system might systematically under-protect or over-enforce in certain areas. In cybersecurity, bias could mean a model is highly tuned to specific attack types but blind to others (perhaps due to their absence in the training set). Or a model might disproportionately flag activities from specific network segments as malicious due to coincidental historical patterns, creating security gaps or unnecessary work. Ensuring diverse and comprehensive training data, as well as regularly evaluating model performance across different scenarios, is crucial for mitigation. Federal AI principles emphasize avoiding bias that could impact fairness or civil liberties, as erroneous threat labeling could have career implications.
- **Over-Reliance and Automation Bias:** There's a risk that human operators become overly reliant on AI, assuming its infallibility. Over-reliance can lead to lapses in human vigilance; analysts might stop independent analysis or ignore obvious attack signs because "the AI didn’t raise an alert." This is dangerous because clever attackers might bypass AI, leaving attacks unnoticed if humans aren't engaged. Conversely, automation bias occurs if humans take AI alerts at face value without proper verification, potentially causing disruptions (e.g., wrongly disconnecting a system). A balanced approach is

needed: AI as an assistant, not the sole authority. Federal guidance generally advises that AI output should be subject to human judgment, especially in high-risk applications.

- **Security of AI Systems Themselves:** AI systems introduce new and expanded attack surfaces. The ML models and endpoints can be compromised via adversarial queries, and data stores and pipelines can be targeted through poisoning or manipulation. Model files could be stolen (raising IP and security concerns) or tampered with. Concerns have been raised about the need to "protect the AI development lifecycle," including training data, models, and ML code, as an extension of secure software development. If an attacker gains access to an agency's AI model, they might extract sensitive information or use it to plan evasions. Thus, securing AI components with strong access controls and encryption is a foundational aspect of proper AI hygiene.

In summary, the adoption of AI/ML in cybersecurity is accompanied by a complex risk landscape. Adversaries may attempt to confuse, poison, or evade AI systems; models can become stale or biased; and misuse or overconfidence can cause failures. Federal agencies must deploy AI with a clear understanding of these challenges. Fortunately, emerging best practices and architectural strategies can mitigate many of these issues. As NIST guidance concludes, there is "no foolproof defense" yet; a combination of robust design, continuous monitoring, and human vigilance is needed.

Strategic Recommendations for Federal Leaders

To guide federal cybersecurity leaders and decision-makers, this framework presents eight strategic recommendations for harnessing AI/ML effectively. These recommendations are not a checklist of discrete actions but an integrated, mutually dependent system for building a resilient, AI-enabled cyber defense posture.

1. **Establish Robust AI Governance and Accountability:** Agencies must implement comprehensive governance frameworks to ensure AI/ML use is transparent, auditable, and aligned with mission outcomes. This includes maintaining a comprehensive AI inventory, as required by Executive Order 13960, and applying risk management principles from frameworks such as the NIST AI Risk Management Framework to protect privacy and civil liberties.
2. **Invest in Data Quality, Security, and Availability:** As data is the lifeblood of AI, agencies must treat it as a strategic asset. This requires investing in high-quality, representative datasets and securing the entire data pipeline against tampering and poisoning attacks, following best practices from the National Security Agency (NSA) and CISA.
3. **Integrate AI/ML into Existing Cybersecurity Operations:** To be effective, AI tools must be embedded within, not isolated from, existing Security Operations Center (SOC) workflows, SIEM/SOAR platforms, and incident response processes. This integration builds analyst trust and ensures AI-driven insights are actionable.
4. **Develop a Skilled AI-Cyber Workforce and Provide Training:** The most advanced tools are ineffective without skilled personnel. Agencies must invest in recruiting, upskilling, and reskilling their cybersecurity workforce to be adept at developing, managing, and interpreting AI-driven systems.

5. **Implement Continuous Monitoring, Testing, and Model Maintenance:** AI models are not static; they are living systems that can drift or be evaded over time. Agencies must implement continuous performance monitoring, conduct regular red-teaming and adversarial testing, and maintain a disciplined model retraining and maintenance lifecycle.
6. **Employ Hybrid Defense Strategies:** The most robust defense posture combines the strengths of deterministic, rule-based systems with the adaptive, pattern-finding capabilities of AI. Agencies should layer AI and traditional controls to create a defense-in-depth architecture with no single point of failure.
7. **Ensure Compliance, Security, and Privacy in AI Systems:** All AI deployments must adhere to federal security and privacy mandates, including FISMA and the Privacy Act. This involves securing the AI development lifecycle, managing the software supply chain, and ensuring all decisions are auditable.
8. **Collaborate and Share Best Practices Across Agencies:** Cyber threats are a whole-of-government problem, and so too must be the defense. Agencies should actively share successes, failures, threat intelligence, and best practices through interagency forums to accelerate collective learning and improvement.

Ultimately, federal leadership must foster a culture that views AI/ML as a powerful but evolving capability—one that demands constant learning, caution, and agility. The path forward requires diligence and strategic investment. Still, the reward is a safer, more resilient government digital infrastructure capable of defending the nation's critical assets against the cyber adversaries of today and tomorrow.

Conclusion: Synthesizing the Framework: A Call to Action for Federal Leadership

The eight strategic recommendations outlined in this framework—spanning governance, data, integration, workforce, monitoring, hybrid strategies, compliance, and collaboration—are not a menu of discrete options from which to choose. They represent an integrated, mutually dependent system. Robust governance is meaningless without high-quality data; a skilled workforce cannot succeed without well-integrated tools; and the most advanced tools will fail without continuous monitoring and maintenance. Success in leveraging AI for federal cybersecurity requires a holistic approach that integrates technology, people, policy, and process into a unified, resilient framework.

This paper serves as a call to action for federal leaders. The role of leadership is to champion this holistic vision and provide the sustained, top-down support required to bring it to fruition. This includes allocating dedicated budgets for not only the initial procurement of AI tools but also for the ongoing costs of data curation, model maintenance, and workforce training. It requires establishing clear policies that empower and guide the responsible use of AI. Most importantly, it demands the organizational will to break down silos and foster the culture of collaboration and continuous improvement necessary to stay ahead of adversaries.

Fostering the Culture of Agile and Responsible Innovation

Implementing this framework requires a significant cultural shift within many agencies. This cultural transition suggestion is rooted in federal guidance (DoD's AI strategy, GAO's AI audit commentary, NIST's AI RMF, etc.). The traditional government posture, often characterized by a deep aversion to risk, must evolve into one of managed risk-taking. The development and deployment of AI/ML systems is an iterative, experimental process. Not every model will succeed, and some will underperform. A culture of innovation treats these "failures" not as career-ending mistakes but as invaluable learning opportunities—data points that provide the feedback necessary to build better, more resilient systems.

Leadership must create an environment that encourages experimentation within safe boundaries, rewards learning, and accepts that the path to a powerful AI capability involves continuous adaptation. This means fostering psychological safety, where analysts and data scientists feel empowered to challenge assumptions, report model weaknesses, and test the limits of their systems without fear of reprisal.

The Path Forward: Sustaining Momentum in the AI Arms Race

The strategic imperative for adopting AI in cybersecurity is clear and urgent. The cyber threat landscape will not stand still; adversaries will continue to innovate, refining their use of AI for attack and actively targeting our AI-based defenses. Our defense strategies must therefore be equally dynamic. As the NSA's AI Security Center asserts, protecting our AI systems has become an integral part of protecting our national security.

The path forward requires continuous vigilance and adaptation. By embracing AI proactively but never uncritically, by fostering cross-disciplinary expertise, by prioritizing security and ethics by design, and by committing to a cycle of continuous learning, federal agencies can effectively meet this challenge. With the strategic foresight of leadership and the rigorous implementation of best practices outlined in this framework, the federal government can confidently leverage AI and machine learning as a decisive force for strengthening national cybersecurity, ensuring a safer and more resilient digital infrastructure for the nation in the era of AI.

Sources:

- DHS S&T Feature Article – *"Leveraging AI to Enhance the Nation's Cybersecurity,"* Oct 17, 2024.
- GAO Report – *"Artificial Intelligence: DHS Needs to Ensure Responsible Use for Cybersecurity,"* GAO-24-106246, Nov 2023.
- CISA AI Use Cases – *"CISA Artificial Intelligence Use Cases,"* Dec 2024; including descriptions of ML for PII detection, threat indicator scoring, malware analysis, network anomaly detection.
- NIST News – *"NIST Identifies Types of Cyberattacks That Manipulate AI Systems,"* Jan 4, 2024.
- PurpleSec Blog – *"AI in Cybersecurity: Defending Against Latest Threats,"* 2023 (discussing benefits like speed, and risks like data poisoning, adversarial attacks, model drift).
- NSA/CISA/FBI Joint Cybersecurity Info Sheet – *"AI Data Security: Best Practices for Securing Data Used in AI Systems,"* May 2025 (discussing data supply chain, poisoned data, and drift).
- LinkedIn Article – *"Hybrid Approaches: Combining ML Techniques to Tackle Cyber Threats,"* Apr 2025.
- FedTech Magazine – *"Machine Learning Models Expedite Federal Tech Efforts,"* Jun 2025 (notes on agencies using cloud ML platforms like AWS SageMaker).
- Splunk Product Info – *"Splunk User Behavior Analytics (UBA),"* accessed 2025 (ML baselining for insider threat detection).
- NIST AI Risk Management Framework – *Overview of AI RMF 1.0*, 2023.