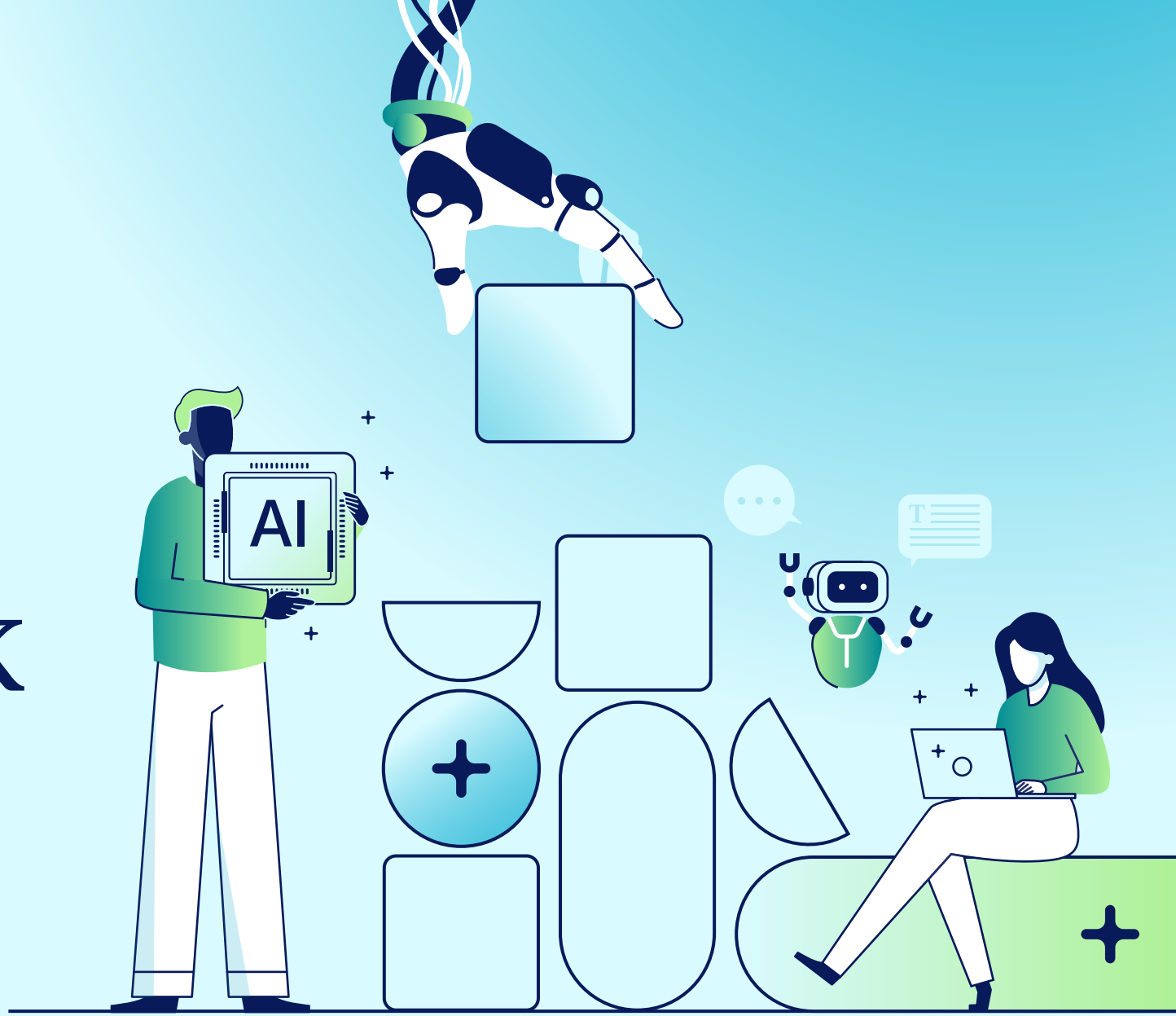
netskope

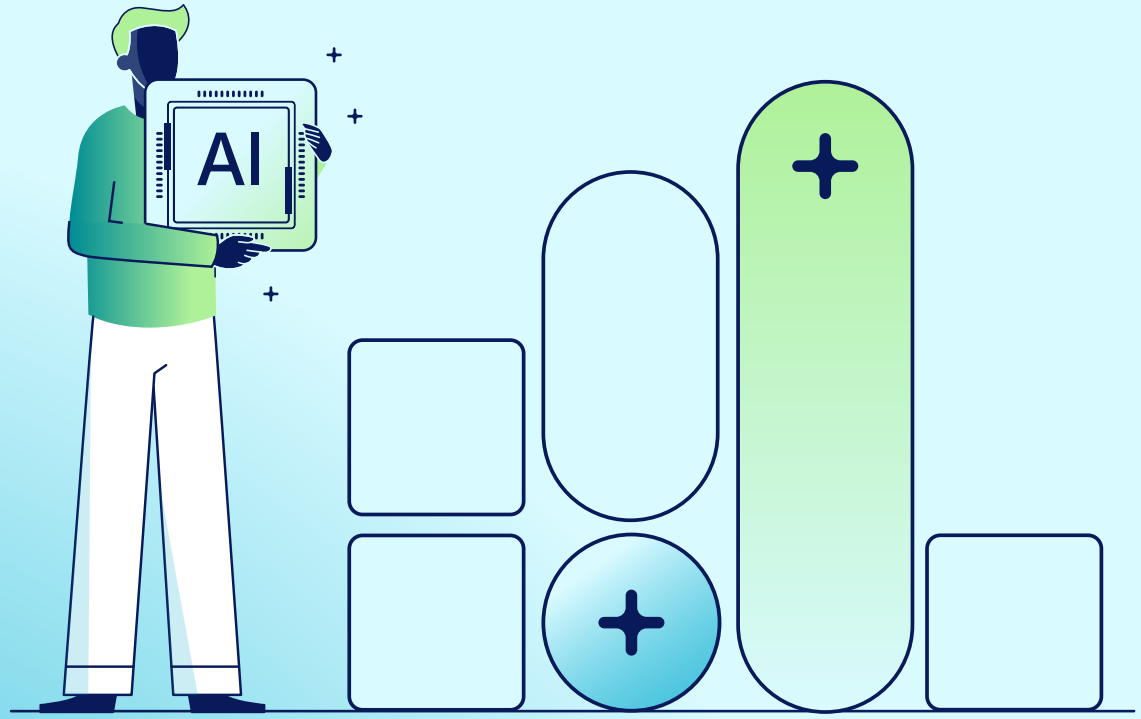
The Federal AI Security Playbook

Closing the gap between policy and practice



The Federal

AI Security Playbook



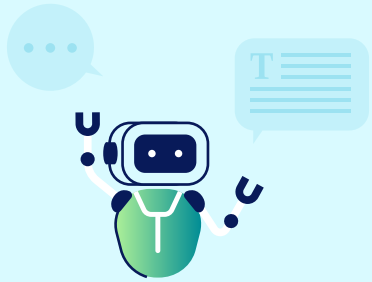


Table of Contents

Introduction	3
Chapter 1: Security Challenges in AI	4
Chapter 2: From Mandate to Implementation	5
Chapter 3: Navigating AI Security	6
Chapter 4: The Future of AI Security	12
Conclusion	13
About Netskope	14

Introduction

Federal agencies are deploying AI faster than they can secure it. Across the government, agencies disclosed more than 3,600 AI use cases in 2025, a 105% jump in a single year[1], even as GAO found most lack adequate secure safeguards for those deployments. The policy environment is shifting faster than most security teams can respond.

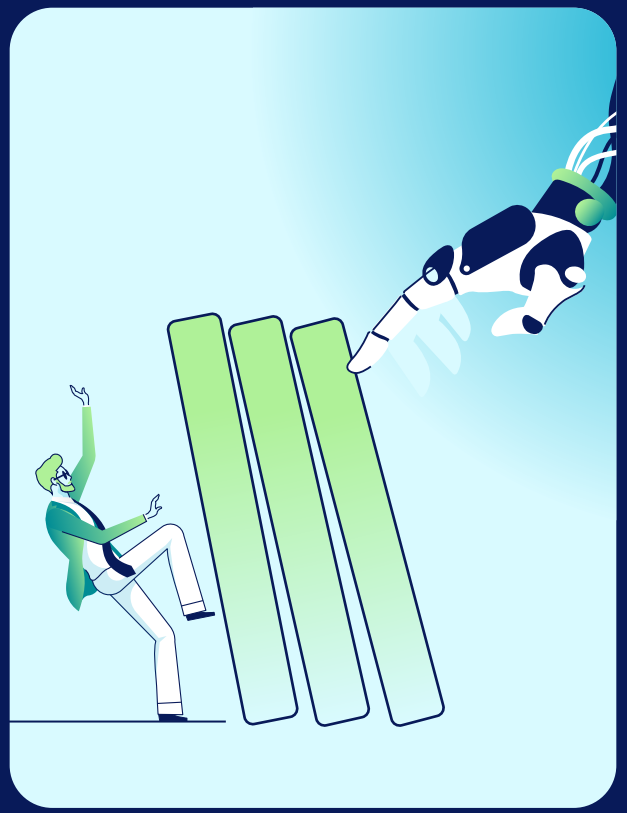
Only 11% of federal agencies have AI fully integrated into their Zero Trust architecture[2], even as AI deployment accelerates across mission-critical operations.

The most immediate risk is data. Sensitive government data, Controlled Unclassified Information (CUI), personally identifiable information (PII), and mission-critical operational data, is being entered into third-party AI applications without adequate visibility or controls. Model Context Protocol (MCP) makes data sharing with AI applications even easier, and exposure harder to control.

The challenge is intensifying. Agentic AI systems can operate autonomously, executing taste, making decisions, and accessing agency data without constant human intervention. Gartner forecasts that by 2028, 25% of all breaches will be tied to AI agent abuse[3].

This playbook maps the six AI security challenges federal agencies face today to the controls that address each one, from data protection and threat defense to AI governance under an accelerating mandate environment.

In June 2025, a zero-click vulnerability in Microsoft 365 Copilot, deployed across federal agencies, was found capable of silently exfiltrating documents from users' SharePoint and OneDrive without a single interaction[4].



[1] Fedscoop, Disclosed government AI use increased by 70% in 2025 per OMB, April 2026

[2] Market Connections / Netskope, May 2026, From Policy to Practice: Operationalizing Zero Trust for the Federal Government in the Age of AI

[3] Gartner's Top Predictions for 2025

[4] The Hacker News, Zero-Click AI Vulnerability Exposes Microsoft 365 Copilot Data Without User Interaction, June 2025

Security Challenges in AI

Top three issues facing security teams today

1 Expanding risk surface

Every path AI takes is a path adversaries can take out. That surface now spans public chat tools, embedded SaaS, privately hosted models, and connections to mission systems.

- Public genAI tools and AI features embedded in existing SaaS apps expose CUI and mission-critical data outside federal data security controls.
- Privately hosted LLMs and custom AI apps introduce new vectors, such as misconfigured access controls or vulnerabilities in data pipelines.
- MCP connections bridge modern AI tools to legacy federal systems, opening pathways outside traditional security controls.
- Classification violations occur when personnel use AI tools at the wrong sensitivity level. The AI doesn't flag the mismatch. The data leaves.

2 Sensitive data exposure and exfiltration

The most immediate risk in AI adoption is data loss, whether it's accidental or malicious.

- Inadvertent exposure happens when agency personnel input CUI, PII, or other sensitive data into public AI tools, often unaware that federal handling rules fully apply, per ISOO Notice 2026-01.
- Malicious insiders or nation-state adversaries actively exploit AI tools to exfiltrate data or abuse model outputs.
- Custom AI training on government records creates persistent risk. Sensitive data embedded in model parameters can surface in outputs long after training.

3 Responsible AI governance

Federal AI governance isn't aspiration. The mandates are active and agency leadership is accountable for fulfilling them.

- AI models can unintentionally encode and propagate biases, leading to regulatory scrutiny and reputational damage.
- Improper handling of data in agency AI workflows may violate FISMA, CUI rules, FedRAMP boundaries, or HIPAA.
- The autonomous deployment of AI in place of human decision-making, especially in high-stakes areas creates gaps existing policy hasn't resolved.
- OMB M-25-21 requires every agency to designate a Chief AI Officer, but most agencies are still building the governance infrastructure to fulfill it.

From Mandate to Implementation

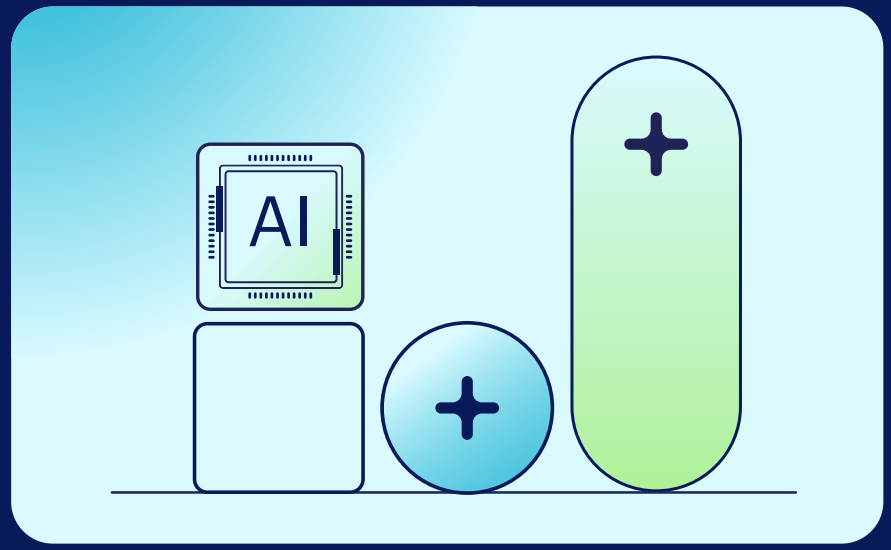
The mandate exists. The readiness gap does too.

Federal agencies understand the AI security challenge. The mandate stack is clear: EO 14028, OMB M-25-21, ISOO Notice 2026-01. The risks are documented. The gap is in implementation.

A May 2026 survey of 275 federal civilian and defense respondents shows how wide that gap is: only 31% have high or complete visibility into shadow AI tools; only 29% describe their security controls as fully aligned with federal guidance; just 13% apply policy consistently across AI agents, applications, and cloud environments[1].

Only 28% of federal respondents describe their security stack as prepared to support expanded AI adoption. Among DoW respondents, that number drops to 9%[2].
The mandate is not the gap.
The implementation is.

That gap matters because AI doesn't wait. It's embedded in agency workflows, SaaS applications, and mission systems whether or not controls are in place. The six challenges ahead map directly to where enforcement is breaking down, and where sensitive data, mission continuity, and mandate compliance are at risk.



Pro Tip

OMB M-25-21 requires agencies to maintain an AI use case inventory. That inventory is also your attack surface map. Agencies that treat compliance reporting as a security input close the implementation gap faster.

[1,2] Market Connections / Netskope, May 2026, From Policy to Practice: Operationalizing Zero Trust for the Federal Government in the Age of AI

Navigating AI Security

Top six challenges and solutions

Challenge #1: Lack of Visibility

As AI tools become embedded in daily workflows, federal agencies are grappling with a fundamental security challenge: they can't secure what they can't see.

Agency personnel and contractors access sanctioned and unsanctioned applications with both corporate and personal credentials, blurring the lines between approved and unapproved use. This uncontrolled sprawl increases the risk of data leakage, mission data loss, and compliance violations, especially when sensitive information is entered into unmanaged or shadow AI services. Personnel also spin up local LLMs on endpoint and server infrastructure without oversight and connect to public MCP servers both directly or as a user or via an Agent. AI agents, outside any visibility boundary.

Most agencies can't differentiate risky from legitimate AI use. Traditional tools fall short in identifying model interactions, distinguishing personal from corporate accounts, or surfacing real-time insights at the user, app, or activity level. Without deep visibility, security teams are left blind to potential exposure points.

How Netskope solves for this

Netskope gives federal agencies the visibility to see every AI interaction, across managed tools, shadow applications, API connections, and MCP traffic, before exposure becomes an incident.

Key capabilities include:

- **Advanced Instance Awareness:** Distinguish between personal and government instances of AI applications like ChatGPT, Gemini, and Copilot.
- **AI Command Center:** Gain deep insights into AI usage trends, top applications, frequency of access, and granular user actions such as logins, posts, uploads, and downloads.
- **User and Entity Behavior Analytics (UEBA):** Detect anomalies and risky behavior using machine learning to identify threats such as data exfiltration, insider risks, and policy violations.
- **Foundational Visibility:** Unified coverage across user-to-app, API, and MCP traffic. Netskope unifies visibility from usage, inventory, and data flows, alongside both endpoint and server AI discovery capabilities.

Agencies gain the granular visibility needed to enforce AI policy, meet M-25-21 governance requirements, and respond before exposure becomes an incident.

Challenge #2: Understanding AI Application Risk

What looks like a routine SaaS application may now include an AI copilot, smart reply engine, or AI-generated summary, added by the vendor without notifying agency security teams. The application was authorized. The AI capability was not. This growing trend makes it increasingly difficult to understand which applications are using AI, how they're using it, and what risks they introduce to the agency.

Security teams need the ability to dynamically assess risk based on how AI features are integrated, whether they retain or train on agency data, and how they align with federal requirements. Without that visibility, agencies face data exposure, loss of sensitive government information, regulatory violations, and even AI model manipulation, risks that grow with every feature a vendor quietly ships.

How Netskope solves for this

Netskope tackles the evolving complexity of AI application risk with its Cloud Confidence Index (CCI), which provides real-time, continuously updated insights into more than 85,000 cloud and SaaS applications.

Key capabilities include:

- **Real-Time, AI-Aware Risk Scoring:** Identify applications with embedded AI capabilities and understand the risks associated with these features, including FedRAMP authorization status and ATO scope.
- **Agency Data Handling Insights:** Evaluate how applications manage agency data, including retention, model training, and third-party sharing.
- **Compliance Tracking:** Stay aligned with regulatory requirements including FedRAMP, FISMA, NIST SP 800-53, and CMMC.
- **Secure LLMs and MCP:** Evaluate AI apps, embedded AI features, and public MCP servers for risky attributes, authentication types, and protocol versions.

With CCI, security teams can confidently navigate AI application risks and ensure their agency remains secure and compliant.

Challenge #3: AI Model Integrity

As agencies deploy generative AI tools (both custom-built models and applications like Microsoft Copilot), ensuring the integrity of the data used to train these models becomes a critical concern. AI systems trained on agency datasets can surface sensitive documents, case files, and CUI in model outputs. RAG systems compound this: data retrieval requires decryption, creating a vulnerability window in memory pipelines that security teams rarely see.

Microsoft Copilot can be trained on the content within a user's Office suite, ranging from Word documents to Excel spreadsheets. If CUI or sensitive government data is stored in these locations, without proper access controls, then there is a potential risk that Copilot could surface controlled program information, or mission-critical details in responses. ISOO Notice 2026-01 is clear: agencies remain accountable for CUI protection even when that data is processed by AI tools; Copilot does not create an exception.

How Netskope solves for this

Netskope One DSPM (Data Security Posture Management) gives agencies visibility and control over sensitive data across cloud environments, preventing CUI, PII, and controlled program information from unauthorized use in AI model training.

Key capabilities include:

- **Continuous Monitoring of Cloud Environments:** Detect and classify sensitive data in real time, ensuring unauthorized use in AI model training is prevented.
- **Visibility into Data Access and Sharing:** Gain real-time insights into how data is accessed and shared across the cloud, allowing for immediate corrective action when necessary.
- **Compliance and Data Leakage Prevention:** Safeguard sensitive data to ensure compliance, prevent data leakage, and maintain control over controlled government information.
- **Robust Security Posture Management:** Ensure proper data posture and discover, label, and classify your structured and unstructured data.

With Netskope One DSPM, agencies close the enforcement gaps that put training data integrity and CUI protection at risk.

Challenge #4: Threats Targeting AI Systems

For federal agencies, AI threats aren't generic, they're nation-state. China, Russia, Iran, and North Korea are using AI to accelerate reconnaissance, automate exploitation, and exfiltrate sensitive data at scale.

AI is also compressing adversary timelines. Attacks that once required weeks can now be executed in hours. Federal agencies built around slow-burn detection and response cycles are structurally exposed to this shift.

Prompt injection is a specific federal attack vector. An adversary who manipulates the prompt context of a federal AI system, through a malicious document, poisoned RAG source, or a crafted API call, can extract CUI, redirect AI-assisted decisions, or trigger unauthorized agent actions.

CISA and NSA addressed this in May 2026, identifying five AI-specific risk categories: privilege escalation, design and configuration failures, behavioral misalignment, structural brittleness, and accountability gaps. Agencies without controls in each area have documented security gaps.

How Netskope solves for this

Netskope delivers AI-specific defenses built for the federal threat environment, detecting and blocking attacks at the point of AI interaction.

Key capabilities include:

- **Unified AI Defense:** Netskope One AI Guardrails mitigates sophisticated attacks, including prompt injection and jailbreak attempts, through deep, real-time analysis of all traffic.
- **Advanced Threat Protection:** Machine learning, sandboxing, and heuristic analysis to detect and block both known and zero-day threats, including malware hidden in files submitted to AI tools.
- **Red Teaming and Vulnerability Assessments:** Automate adversarial simulations to uncover vulnerabilities and ensure agency AI models are secure and resilient against advanced threats.
- **Proactive AI Activity Monitoring:** Real-time monitoring of AI interactions detects emerging threats and policy violations before they escalate.

By combining these technologies, Netskope provides an integrated solution that helps agencies secure their AI systems from sophisticated threats and evolving attack vectors.

Challenge #5: Data Exposure

One of the most urgent and high-stakes challenges in AI security is the risk of data exposure. As agency personnel across functions adopt AI tools to boost productivity, they may unknowingly upload or share sensitive data in the process. Source code for federal systems, CUI, law enforcement sensitive information, and PII are all entering public AI models, often without visibility or controls. Once exposed, this data can be retained, used for model training, or even leaked.

The consequences extend beyond compliance findings. Exposed CUI can compromise ongoing operations. Exposed law enforcement data can endanger investigations and confidential sources. Exposed source code for FISMA-categorized systems can reveal exploitable vulnerabilities in national security infrastructure.

Unlike traditional data sharing channels, AI platforms can act as black boxes, offering little transparency into how data is stored, accessed, or used. Without guardrails in place, agencies face FISMA compliance findings, CUI handling violations, potential counterintelligence exposure, and erosion of public trust.

How Netskope solves for this

Netskope protects agency data from exfiltration through AI tools, at rest, in motion, and at the point of AI interaction. By combining real-time risk assessments, inline and API-based controls, and posture checks, Netskope's unified security policies enable precise governance of both user and data interactions across the organization.

Key capabilities include:

- **Advanced Data Loss Prevention (DLP):** Prevent CUI, PII, source code, and other sensitive data from exfiltration through AI tools, whether agency personnel are in the office, remote, or in the field.
- **Granular Control:** Block or limit high-risk actions, such as uploading source code CUI, or mission-critical program documentation.
- **Real-Time User Coaching:** Educate users on policy violations with visual prompts, helping to reduce repeat offenses.
- **Inspect Every Request and Response:** Identify and block the delivery of patented or copyrighted data in AI responses to defend against data integrity risks.
- **Secure API Traffic:** Authenticate and centralize traffic management and content inspection between private apps and LLMs.

With these capabilities, Netskope ensures comprehensive, adaptive data protection that scales across an agency's entire AI and cloud environment.

Challenge #6: Governance, Compliance, and Ethical Use

Federal agencies don't just face governance pressure, they face an active, expanding mandate stack with accountability consequences.

The federal mandate stack for AI governance:

Mandate	What it requires
OMB M-25-21	Agency Chief AI Officers, AI use case inventory, governance accountability
EO on AI Innovation & Security (June 2026)	Cybersecurity requirements for frontier AI; NSA classified benchmarking within 60 days
NIST AI RMF	Govern, Map, Measure, and Manage functions for responsible AI
ISOO Notice 2026-01	CUI and classified handling rules apply to AI without exception
CISA/NSA AI Data Security (May 2025)	Secure AI training data — addresses supply chain integrity, data poisoning, and drift
CISA Agentic AI Guidance (May 2026)	Five agentic AI risk categories; cryptographic agent identity; human override requirements
DoD AI Strategy	Procurement transparency, training data provenance, and human oversight requirements

Meeting these requirements demands visibility into AI usage, control over how data is handled, and the ability to demonstrate compliance across every applicable mandate.

How Netskope solves for this

Netskope gives agencies the visibility and policy controls to govern AI across the mandate stack.

Key capabilities include:

- **Granular Policy Enforcement:** Control how data is shared with AI tools, preventing sensitive or regulated data from unauthorized use in AI model training.
- **Real-Time Compliance Controls:** Block uploads of PHI to noncompliant apps; prevent CUI from entering AI tools without proper authorization.
- **Regulatory Framework Support:** Support compliance with OMB M-25-21, NIST AI RMF, CISA AI security guidance, and more.
- **Real-Time Content Moderation:** Automatically filter and control harmful or discriminatory content, including hate speech, crimes, weapons, and violence.
- **Master AI Data Governance:** Automate discovery, classification, and pre-deployment hardening to ensure controlled government information remains protected and compliant.

With Netskope, agencies can demonstrate compliance, enforce AI policy in real time, and close the governance gaps that put mission data and mandate alignment at risk.

The Future of AI Security

Emerging technologies and threats

As AI adoption grows and new use cases, from copilots to As AI capabilities advance, so does the threat landscape. Three emerging areas will define the next phase of federal AI security.

Agentic AI systems capable of autonomous action are already being weaponized. In September 2025, a Chinese state-sponsored group used an AI coding agent to autonomously attack approximately 30 global targets, including government agencies[1]. CISA and NSA responded with joint guidance in May 2026 recommending scoped permissions, cryptographically anchored agent identity, and human override capabilities for all high-stakes agentic actions.

AI supply chain risk grows as agencies build and procure AI systems from third-party components they don't fully control. In 2024, researchers found approximately 100 malicious models on HuggingFace[2] containing embedded code execution payloads. A compromised pre-built model can introduce exploitable vulnerabilities that persist through deployment and into production.

AI enabled social engineering is scaling attacks against federal personnel. In July 2025, threat actors used AI voice cloning to impersonate the Secretary of State[3], targeting foreign ministers, senators, and governors. AI-generated voice calls surged 1,600% in Q1 2025[4]. The FBI logged over 22,000 AI-related fraud complaints with losses exceeding $893 million[5].

To stay ahead, agencies should prioritize:

- **AI visibility:** Maintain central visibility into all AI tools across the agency, sanctioned and unsanctioned.
- **Autonomy boundaries:** Enforce scoped permissions for agentic systems with human override and verify identity, per CISA's May 2026 guidance.
- **Supply chain vetting:** Validate pre-built AI models and components before deployment. Treat AI supply chain risk like software supply chain risk.
- **Workforce readiness:** 54% of federal respondents cite limited staff expertise as the top AI security barrier[6]. Controls alone are not enough.

The future of AI security will be defined by how thoughtfully agencies prepare, not just how well they respond.

[1] Market Connections / Netskope, May 2026, From lschwartz@netskope.com Policy to Practice: Operationalizing Zero Trust for the Federal Government in the Age of AI

[2] Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign, November 2025

[3] Dark Reading, Hugging Face AI Platform Riddled with 100 Malicious Code-Execution Models, February 2024

[4] CNN, Someone using AI to impersonate Marco Rubio contacted at least five people including foreign ministers, cable says, July 2025

[5] Right-Hand Cybersecurity, The State of Deep Fake Vishing Attacks in 2025

[6] FBI, Internet Crime Report, 2025

Prediction

Analyst firm Gartner predicts that, by 2028, at least 15% of daily business decisions will be made autonomously through agentic AI, up from 0% in 2024.

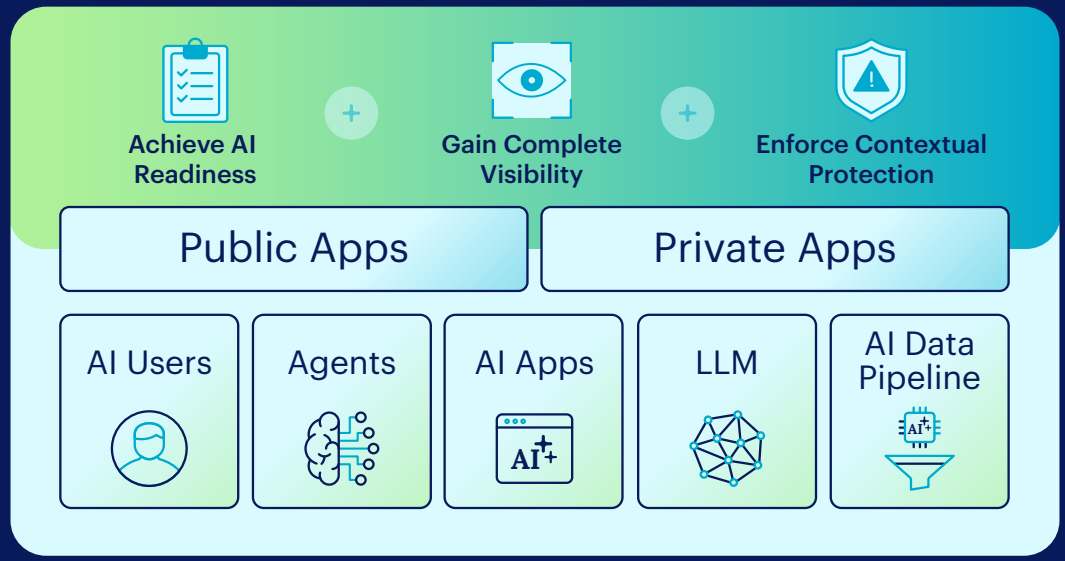
Conclusion

Secure AI end-to-end, everywhere with Netskope One

Federal agencies are deploying AI faster than they can secure it. The six challenges in this playbook represent the most urgent enforcement gaps between where agencies are and where federal mandates require them to be:

- Lack of Visibility
- Understanding AI Application Risk
- AI Model Integrity
- Threats Targeting AI Systems
- Data Exposure
- Governance, Compliance, and Ethical Use

Netskope One AI Security provides a single enforcement layer for agency personnel and automated agents across public SaaS, private AI tools, and agentic workflows. The gap between where most agencies are and where federal mandates require them to be is the mandate gap Netskope is built to close.



Research

Analyst firm Forrester found that Netskope delivers an 80% reduction in the risk of a severe breach caused by an external attack, equivalent to a $2m saving in annualized material breach costs.[1]

[1] Forrester Report: The Total Economic Impact™ of Netskope SSE
https://www.netskope.com/resources/analyst-reports/forrester-the-total-economic-impact-of-netskope-sse

About Netskope

Netskope (NASDAQ: NTSK), a leader in modern security and networking, addresses the needs of both security and networking teams by providing optimized access and real-time, context-based security for people, devices, and data anywhere they go. Thousands of customers, including more than 30 of the Fortune 100, trust the Netskope One platform, its Zero Trust Engine, and its powerful NewEdge Network to reduce risk and gain full visibility and control over cloud, AI, SaaS, web, and private applications—providing security and accelerating performance without trade-offs.

Interested in learning more?

Request a demo

Resources

Netskope One AI Security for Federal Agencies

From AI Adoption to AI Accountability in Federal Agencies Blog

Netskope Threat Labs: Generative AI Cloud Threat Report

From Policy to Practice: Operationalizing Zero Trust for the Federal Government in the Age of AI