APRIL 2026

ATARC Identity Management Working Group

Securing the Agentic State

A Practical Guide to Identity & Access Management for AI Agents in Federal Government

ATARC

Table of Contents

Table of Contents

Acknowledgements

Adam McBride, U.S. Department of Health and Human Services, ATARC Identity Management Working Group Government Co-Chair

Kelvin Brewer, Ping Identity, ATARC Identity Management Working Group Industry Chair

David Treece, Yubico, ATARC Identity Management Working Group Industry Co-Vice Chair

Lisa Palma, LC&J Security Solutions LLC, ATARC Identity Management Working Group Industry Co-Vice Chair

Jim St.Clair, C3HIE, Lead Author and Agentic IAM Task Group Leader

Brad Weisberg, ID.me

Ross Foard, Foard Consulting LLC

Brian Deyo, Omnissa, LLC

Howard Rosen, Nova Insights Corp, HIMSS

Executive Summary

Agentic AI represents the next major leap in artificial intelligence: systems that don't just answer questions, but independently **set goals, make decisions, and take actions** in the real world. Federal agencies will increasingly need these systems to deliver faster, smarter services to citizens.

However, our current security systems—the infrastructure that controls who can access what—were built for human users. They assume people log in, click around at human speed, and get relatively fixed permissions. AI agents shatter all of these assumptions. They operate at machine speed, spawn other agents, and need permissions that change by the second.

This paper explains, in plain language, **why current identity systems will break, what new capabilities are needed, and how federal agencies should prepare.**

Three Critical Points for Leaders

1. **Today's security protocols weren't designed for autonomous systems.** They assume human-speed interactions, simple delegation, and static permissions—none of which apply to AI agents.

2. **The scale challenge is exponential, not linear.** Just 1,000 agents can generate 7.4 million authentication events per day—a 148x increase over human users.

3. **Purpose-built solutions exist but require strategic investment.** Decentralized identity, verifiable credentials, and new governance models can solve this—but agencies must act now.

1. What Is Agentic AI?

Think of AI as evolving through three stages. **Generative AI** (like ChatGPT or Claude) responds to questions and creates content—but only when asked. **AI Agents** go further: they can sense their environment, take actions, and learn from feedback. **Agentic AI** is the most advanced stage: systems that autonomously set goals, plan multi-step strategies, and execute them with minimal human oversight.

The key distinction is simple: *automation executes processes; agents achieve outcomes.*

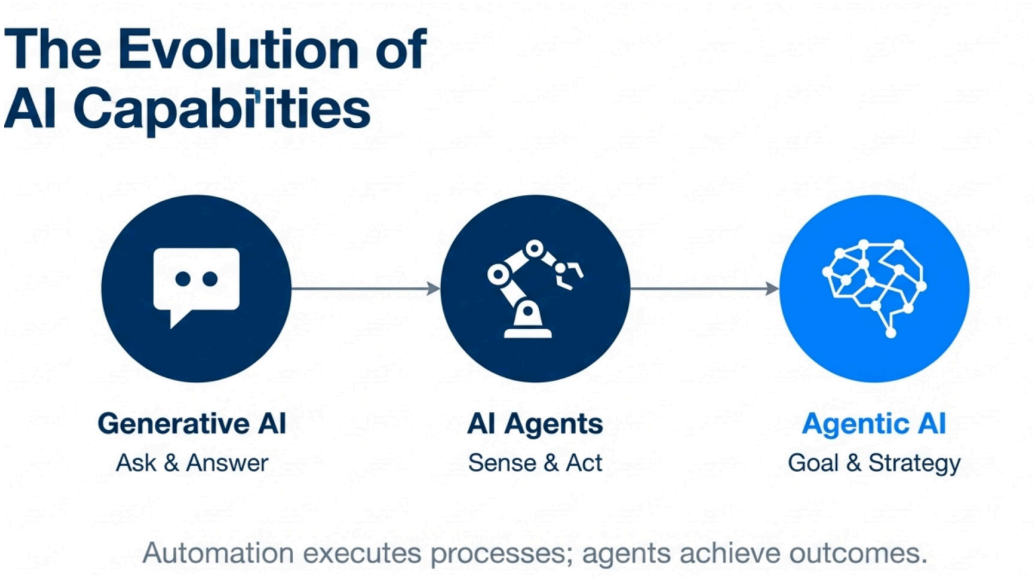

Figure 1: Traditional automation follows rigid steps and fails on surprises. AI agents flexibly plan, act, and adapt to reach goals.

How Autonomous Can Agents Get?

AI agents operate on a spectrum of autonomy, much like self-driving cars. At the lowest levels, agents recommend actions for humans to approve. At the highest levels, they operate entirely independently. Most current federal deployments sit at Levels 2–3, where agents handle tasks but humans still review major decisions.

1. What Is Agentic AI?

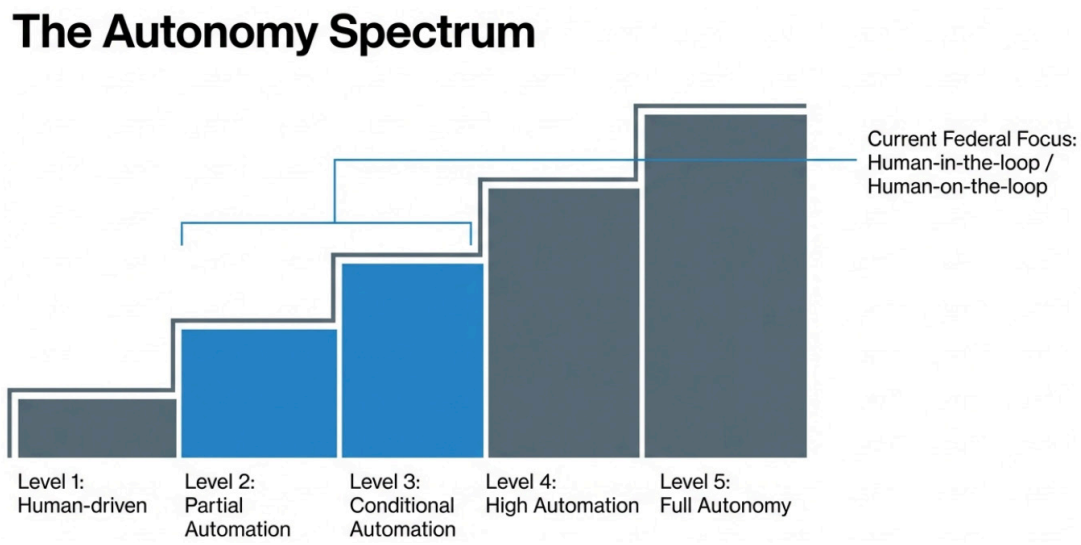

Figure 2: The autonomy spectrum. Most agencies currently operate at Levels 2-3.

The question for federal agencies is not whether to deploy agents, but how to do so securely and accountably.

Real-World Government Examples

- Ukraine's Diia.AI lets citizens describe needs in plain language. Agents then orchestrate services across multiple agencies automatically—no navigating multiple portals required. [1]
- Brazil's State of Goiás cut innovation project review time from one year to one week using agentic workflows, without sacrificing quality. [2]
- Energy grid operators in Europe and the U.S. use AI agents that automatically stabilize power flows and prevent blackouts at speeds no human team could match.

[1] Ukraine's Ministry of Digital Transformation launched Diia.AI in September 2025, making it the world's first national AI assistant providing government services. Powered by Google's Gemini 2.0 Flash, the assistant handles over 90% of support inquiries and has served more than 200,000 citizens. See: digitalstate.gov.ua; Cabinet of Ministers of Ukraine, "Ukraine hits a world record: Diia.AI recognized as the first national AI assistant for public services," November 2025.

[2] Multiple Brazilian federal and state governments have adopted AI-driven workflow automation. The Brazilian Artificial Intelligence Plan (PBIA 2024–2028) allocates approximately R$23 billion (~US$4 billion) to support AI infrastructure and innovation. See: Ministry of Science, Technology and Innovation of Brazil, PBIA 2024–2028, launched July 30, 2024.

2. Why Current Security Systems Will Break

Federal agencies rely on identity and access management (IAM) systems—the security infrastructure that controls *who can access what*. These systems were designed for a world of human users logging into web applications. AI agents break four fundamental assumptions these systems depend on.

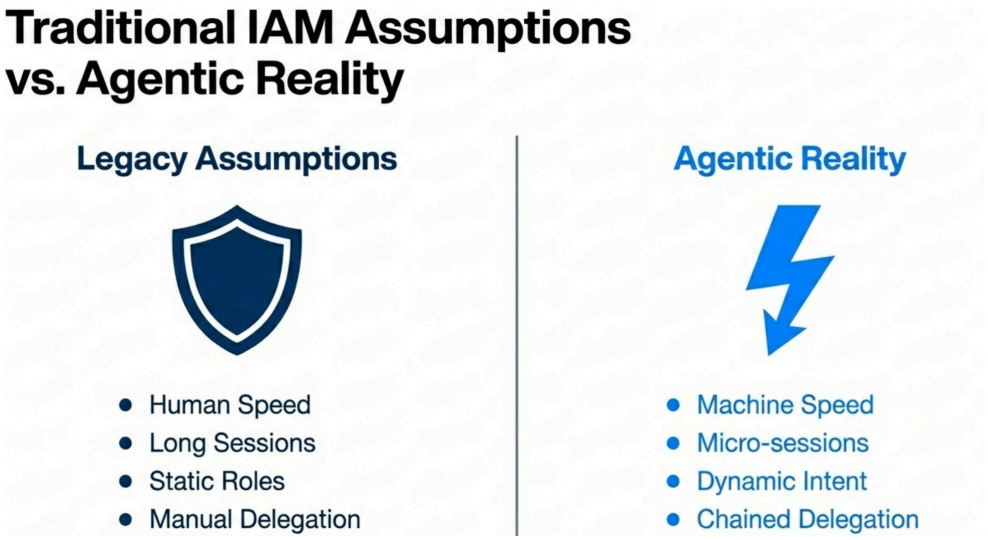

Figure 3: Four core assumptions of traditional IAM systems that AI agents invalidate.

Where Specific Technologies Fail

OAuth is the protocol behind most modern login systems ("Sign in with Google," etc.). It fails for agents because its permissions are too broad ("read data" vs. "read quarterly budget reports from Finance, last 90 days"), and it can't distinguish agent actions from human actions in audit trails. [3]

SAML, widely used for single sign-on in government, is even worse: its XML-based processing can't handle machine-speed authentication, and it assumes long-lived sessions that don't fit short-lived agent tasks. [4]

[3] OAuth 2.0 is defined in IETF RFC 6749 (October 2012). While widely adopted for delegated authorization in web applications, OAuth's coarse-grained scope model and lack of agent-specific audit trails present challenges for autonomous AI systems. See: D. Hardt, "The OAuth 2.0 Authorization Framework," IETF RFC 6749, October 2012.

[4] Security Assertion Markup Language (SAML) 2.0 is an OASIS standard (March 2005) for exchanging authentication and authorization data. Its XML-based processing and session-oriented design are optimized for human-interactive browser flows. See: OASIS, "Assertions and Protocols for the OASIS SAML V2.0," March 2005.

2. Why Current Security Systems Will Break

The “Confused Deputy” Problem

One of the most dangerous vulnerabilities is called the “confused deputy” problem. In short: agents typically inherit broad system permissions (like full database administrator access) even when they only need to read a single table. If an attacker compromises that agent, they gain access to everything the agent’s credentials allow—HR records, financial data, the works. [5]

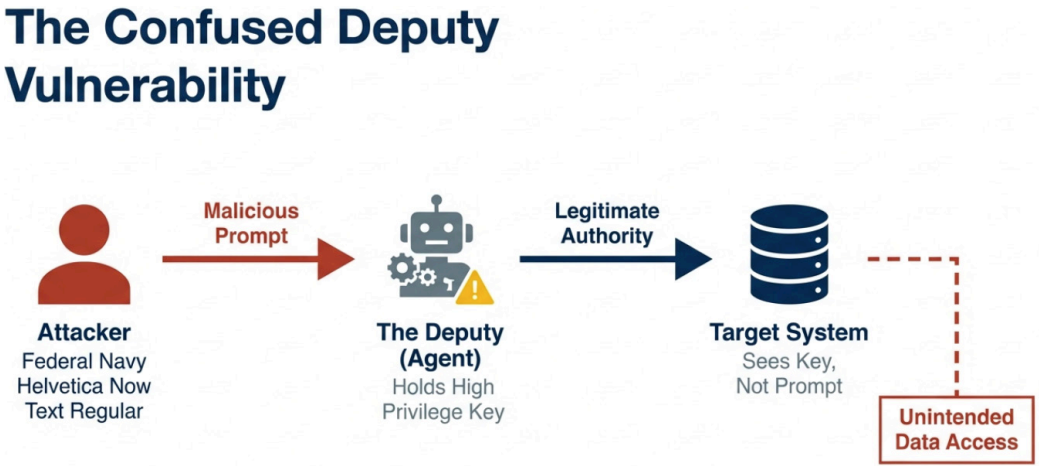

Figure 4: The Confused Deputy problem. An agent with overly broad access becomes a high-value target for attackers.

The Scale Challenge

Even if current systems could handle agent-style access patterns, the sheer volume of authentication traffic would overwhelm them. Consider the numbers:

[5] The confused deputy problem was first described by Norm Hardy in his 1988 paper “The Confused Deputy: (or why capabilities might have been invented).” It describes a privilege escalation vulnerability where a trusted program is tricked into misusing its authority. See: N. Hardy, ACM SIGOPS Operating Systems Review, 22(4), October 1988; AWS IAM User Guide, “The confused deputy problem.”

2. Why Current Security Systems Will Break

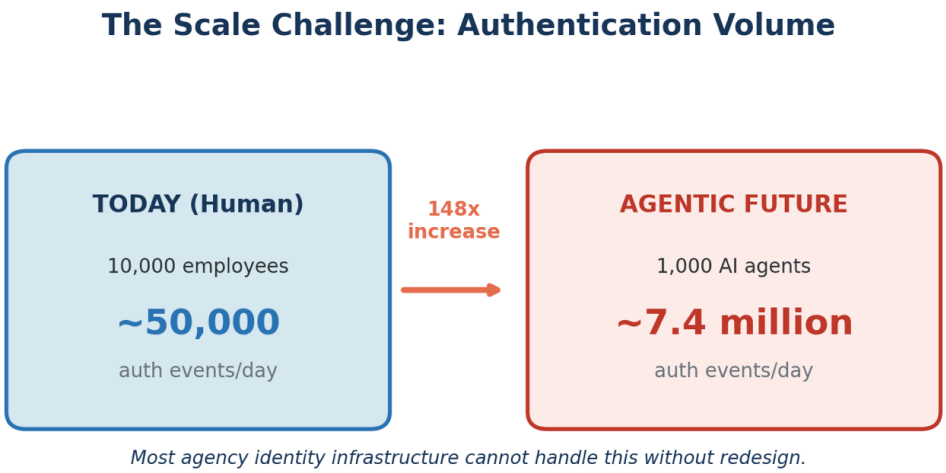

Figure 5: Authentication volume comparison between human users and AI agents at a single federal agency.

3. The New Identity Architecture

Securing autonomous agents requires rethinking identity management from the ground up. The new architecture rests on five core principles, all built on a Zero Trust foundation ("never trust, always verify").

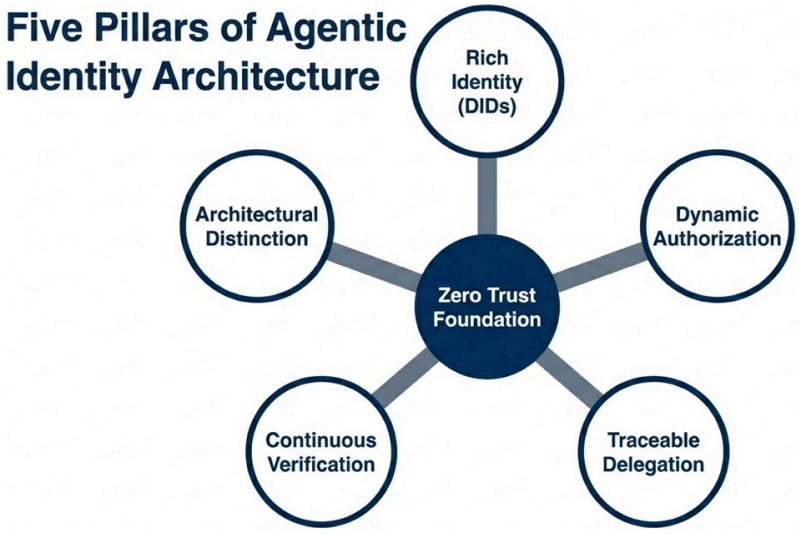

Figure 6: Five pillars of the new identity architecture for AI agents.

3. The New Identity Architecture

The Five Pillars Explained

Rich Identity: An agent's identity must be far more than a username. It should include cryptographically verifiable credentials, what it's authorized to do, who created it, who owns it and what tools it can use—like a digital passport.

Dynamic Authorization: Access decisions must happen in real-time based on context: what the agent is doing right now, how risky the request is, and what resource it's trying to reach. Static "role" assignments aren't sufficient.

Traceable Delegation: When agents hand off tasks to sub-agents, every handoff must be cryptographically recorded. Permissions must get narrower (never broader) at each step, and revoking access must cascade instantly through the chain.

Continuous Verification: Instead of authenticating once at login and trusting the session, the system verifies identity on every significant action and monitors for unusual behavior. As with other systems, access should be regularly reviewed.

Architectural Distinction: Existing systems answer "what is running." Agent identity must additionally express "who is acting, under whose authority, for what purpose, and for how long."

Key Enabling Technologies

Technology	What It Does
Decentralized IDs (DIDs)	Like digital passports for AI agents—globally unique, cryptographically secured, and verifiable without a central authority. In federal use, DIDs operate within approved trust frameworks.
Verifiable Credentials (VCs)	Digital certificates that make provable claims: "This agent is FedRAMP compliant," "This agent has Secret clearance." VCs support selective disclosure—proving you meet a requirement without revealing extra details.
Agent Naming Service	A secure directory where agents find services based on capabilities, not just names. All registrations are cryptographically signed, with built-in protections against impersonation.
Policy-Based Access	Real-time authorization engines that evaluate complex rules in milliseconds, considering identity, credentials, behavior, and context.
Zero-Knowledge Proofs	Advanced cryptography that lets agents prove statements without revealing underlying data. For example, proving "I have the required clearance" without revealing the exact level.

4. Security Threats Unique to AI Agents

AI agents face attack types that simply don't exist in traditional IT systems. These threats exploit the autonomous, interconnected nature of agents.

Attack Type	How It Works	Real-World Analogy
Prompt Injection	Hidden commands in documents trick agents into executing attacker instructions when processing them.	Like phishing, but the agent reads and acts on it at machine speed.
Tool Poisoning	Attackers create fake tools with names similar to legitimate services. Agents discover and use these compromised tools.	Like a fake ATM that looks real but steals your card info.
Memory Poisoning	Bad actors inject false information into an agent's memory, corrupting all of its future decisions.	Like planting false evidence that an investigator then relies on.
Delegation Chain Exploit	Compromising one agent in a chain, then using its delegated authority to access resources across the organization.	Like stealing a supervisor's master key and accessing every room.

A single compromised agent can access millions of records at machine speed, propagate vulnerabilities across hundreds of interconnected agents, and even spawn additional malicious agents.

The MAESTRO Security Framework

To defend against these threats, the MAESTRO framework defines seven layers of security. Each layer addresses a different dimension of the system, and weakness in any one layer can compromise everything. [6]

[6] MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) was introduced by Ken Huang at the Cloud Security Alliance in February 2025. It defines a seven-layer reference architecture for threat modeling agentic AI systems. See: K. Huang, "Agentic AI Threat Modeling Framework: MAESTRO," Cloud Security Alliance, February 6, 2025, https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro.

4. Security Threats Unique to AI Agents

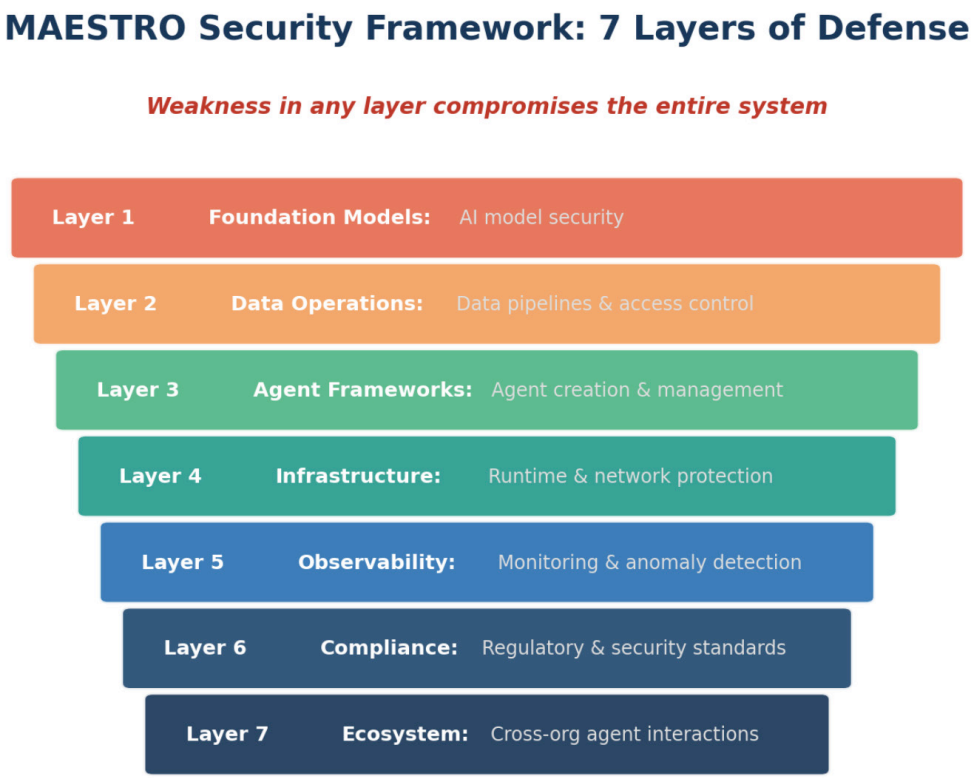

Figure 7: The MAESTRO framework's seven layers of defense for agentic systems.

Zero Trust: The Foundation

Many federal agencies have invested in Zero Trust architectures—the security model where nothing is trusted by default and everything is verified. This foundation is essential for agents, but must be extended with new capabilities: [7]

- **Never Trust, Always Verify:** Verify agent identity on every interaction, not just at login.
- **Least Privilege:** Grant minimum permissions for each specific task, with time-limited credentials.
- **Continuous Monitoring:** Real-time behavioral analysis, anomaly detection, and automated threat response.

[6] The federal Zero Trust mandate was established by Executive Order 14028, "Improving the Nation's Cybersecurity" (May 12, 2021), and further elaborated in OMB Memorandum M-22-09, "Moving the U.S. Government Toward Zero Trust Cybersecurity Principles" (January 26, 2022). See: NIST SP 800-207, "Zero Trust Architecture," August 2020.

5. Governance and Policy Challenges

Beyond technology, AI agents create hard governance questions that existing law and policy don't fully address.

Who Is Accountable?

When an AI agent makes a mistake, who's responsible? The agency that deployed it? The vendor that built it? The official who authorized its use? Current law assumes human decision-makers, and the answers aren't clear yet. Agencies need to establish accountability frameworks before incidents force reactive policy-making.

Key Policy Areas to Develop

- **Agent Authorization:** Who can authorize agent deployment, at what risk levels, and when must agents escalate to humans?
- **Data Governance:** What data can agents access, how to prevent exfiltration, and what privacy protections apply to PII?
- **Incident Response:** Procedures for compromised agents, rapid revocation, and forensic investigation capabilities.
- **Workforce Transition:** How roles shift from processing to oversight, new training requirements, and career paths in human-AI collaboration.
- **Records Management:** Is agent "memory" a federal record? How to preserve decision chains for FOIA compliance?

Recommended Architecture: Federated Model

Rather than a fully centralized or fully decentralized approach, the recommended federal model is a hybrid that balances standardization with agency autonomy:

- **Federal Core Services:** A federal DID registry, root certification authority, Agent Naming Service, and baseline security policies managed centrally.
- **Agency Components:** Agency-specific agent registries, mission-specific credential issuers, and local policy engines managed by each agency.
- **Cross-Agency Federation:** Mutual trust agreements, shared threat intelligence, and common security standards connecting agencies.

6. What Federal Leaders Should Do Now

For Agency Heads

- **Treat this as a strategic priority.** Designate an executive sponsor and include agentic AI in your strategic plan with dedicated budget.
- **Start small now.** Identify one high-value, low-risk pilot within 90 days. Learn fast, iterate, and build foundational capabilities.
- **Address policy gaps proactively.** Commission records management guidance, clarify liability frameworks, and develop privacy impact assessment templates for agents.
- **Invest in workforce adaptation.** Communicate a vision for human-AI collaboration, create new oversight roles, and provide agent management training.

For CIOs

- **Acknowledge current infrastructure gaps.** Conduct IAM capacity assessments and develop a multi-year modernization roadmap—patching won't work.
- **Adopt decentralized identity now.** Evaluate DID methods, deploy pilot VC issuance, and train security teams on decentralized identity.
- **Implement Zero Trust for agents.** Use agent deployment as a catalyst for broader Zero Trust adoption with agent-specific policies.
- **Build observability first.** Deploy comprehensive agent logging and behavioral analytics before scaling—you can't secure what you can't see.

For CISOs

- **Extend threat models.** Adopt the MAESTRO framework, conduct tabletop exercises for agent compromise, and update the enterprise risk register.
- **Demand cryptographic identity.** Require DIDs for all production agents and mandate hardware-backed keys for high-assurance use cases.
- **Plan for agent incidents.** Develop agent-specific incident response playbooks and rapid mass-revocation procedures.
- **Enable, don't block.** Establish fast-track approval for low-risk pilots and measure security by outcomes, not paperwork.

For General Counsels

- **Clarify accountability.** Document lines of accountability, establish review processes, and coordinate with DOJ on liability questions.
- **Address records management.** Consult with National Archives, set retention schedules, and plan for FOIA requests about agent decisions.
- **Ensure privacy protection.** Conduct privacy impact assessments, document PII flows, and update privacy policies for agent interactions.

7. Conclusion: Act Now

The transition to agentic AI is not a question of "if" but "when" and "how." The technology exists, early implementations are succeeding, and commercial pressure is accelerating. Federal agencies face a choice: **lead the transformation** by proactively building secure frameworks, or **react to the transformation** with crisis-driven adoption and inadequate safeguards.

Agencies that begin with pilot projects today—acknowledging uncertainty, learning from experience, building foundational capabilities—will be positioned to deploy agents securely and effectively. Those that wait will face crisis-driven adoption of immature systems without adequate safeguards.

The technical path is clear. The policy questions are definable. The governance frameworks can be established. What's required is leadership commitment to treating agentic AI as the strategic priority it represents—not merely another IT project, but a fundamental transformation in how transformation in government efficiencies, modernize workforce productivity and how government provides services to citizens.

Further Resources

- **Standards:** Decentralized Identity Foundation (identity.foundation), Trust over IP Foundation (trustoverip.org), W3C Verifiable Credentials Working Group
- **Federal:** NIST AI Risk Management Framework, Federal ICAM (FICAM), Executive Order on AI (October 2023)[8,9]
- **Research:** MIT Project NANDA, Stanford Human-Centered AI Institute, Cloud Security Alliance Agentic AI Working Group [10]

This document is intended to inform federal strategic planning and policy development. It does not constitute official guidance. Agencies should consult with their legal, security, and policy teams before implementing agent technologies.

[8] Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," was signed October 30, 2023. It directed over 50 federal entities to take more than 100 specific actions. Note: EO 14110 was rescinded by President Trump on January 20, 2025 via Executive Order 14148, "Removing Barriers to American Leadership in Artificial Intelligence." See: 88 Federal Register 75191 (November 1, 2023).

[9] The NIST AI Risk Management Framework (AI RMF 1.0) was released January 26, 2023 (NIST AI 100-1). It provides a voluntary framework organized around four core functions: Govern, Map, Measure, and Manage. A Generative AI Profile (NIST AI 600-1) was released July 26, 2024. See: https://www.nist.gov/itl/ai-risk-management-framework.

[10] The Cloud Security Alliance's AI Safety Initiative has published multiple frameworks for agentic AI security including MAESTRO and threat model analyses of OpenAI's Responses API and Google's Agent-to-Agent (A2A) protocol. See: https://cloudsecurityalliance.org; https://labs.cloudsecurityalliance.org/maestro/.