ATARC

March 2026

ATARC Agentic AI Working Group

Agentic AI White Paper

Acknowledgements

Anthony Boese, ATARC Working Group Member
Brian Peretti
KJ Lian, AWS
David Egts, Salesforce
Henry Sienkiewicz, Georgetown University
Dr. Joseph “Lucky” Ronzio
Jim St. Clair, C3HIE
Chris Oakleaf, Dpt. Of Veterans Affairs
Connor Turpin, USAF
Nicholas Rappold, National Weather Service
John Sprague, National Aeronautics and Space Administration
Enrico Pontelli, New Mexico State University
Howard Rosen, Nova Insights Corp, HIMSS
Mun-Wai Hon, Federal Aviation Administration
Anil ‘Neil’ Chaudhry, Senior Advisor, Artificial Intelligence, US Department of Transportation

Disclaimer

Executive Summary

ATARC's mission is to bring together academia, industry and government to tackle challenges in IT modernization and emerging technology adoption. The next phase of AI's evolution is agentic AI – systems that are given goals and the ability to interact with the world around them, and which can formulate plans and execute actions to achieve the given goals. The future of agentic AI is transformative; their ability to utilize tools, take actions, reflect on their performance, and manage both short and long-term memory will usher in a new era of automation. Multi-agent systems are extending the reach of foundation models by integrating external knowledge, mitigating hallucinations, and enhancing security. These advancements are not merely incremental; they represent a fundamental shift in how we interact and leverage artificial intelligence.

This paper provides an overview of Agentic AI, use cases, risk frameworks, governance and best practices. Areas of focus include:

1. establishing and enforcing guardrails around uses of agentic AI,
2. ensuring the security and ethical application of these systems, and
3. protecting the "supply chain" of their hardware, software, and development data.

Introduction - What is Agentic AI and Why Does it Matter?

1.1 What is Agentic AI?

Agentic AI represents the next evolution of artificial intelligence, systems that can independently set goals, make decisions, and take actions in the real world. What sets Agentic AI apart is that it acts on its own authority, moving beyond traditional AI that simply responds to prompts. While conventional AI systems are essentially responders, Agentic AI focuses on making decisions rather than just generating content. These systems can perceive their environment, plan multi-step solutions, and execute tasks autonomously with minimal human intervention.

Generative AI, AI Agents, and Agentic AI represent a progressive evolution in artificial intelligence capabilities. Generative AI systems, like ChatGPT, Claude, Grok, and Gemini are designed to create content such as text, images, audio or video in response to prompts, acting as passive responders to human input. While powerful content creators, they remain fundamentally reactive. Agentic AI transcends this limitation by shifting from content generation to autonomous decision-making, enabling systems to not only respond but to proactively plan, reason, and act toward achieving defined objectives.

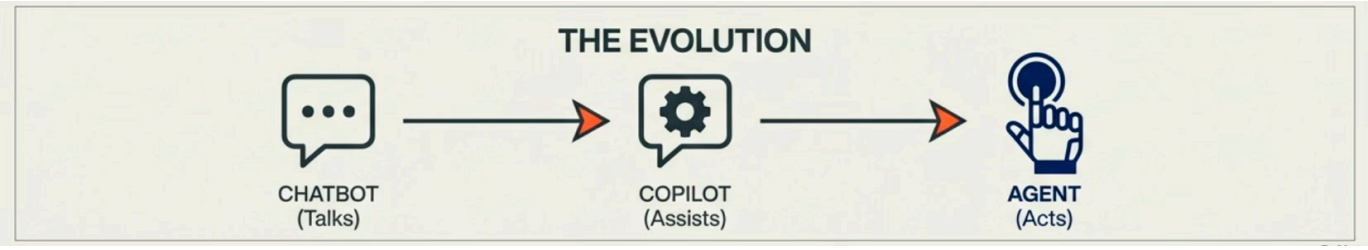

Figure 1: AI Evolution (Generated by NotebookLM, 2026).

AI Agents advance this concept by incorporating the ability to sense, process, and act on their environment through a continuous loop of decision-making, action, feedback processing, and learning. These systems utilize perception, processing, and action components to interact with their surroundings. At the forefront of AI development, Agentic AI builds upon the capabilities of generative AI and AI agents, adding semantic understanding of external systems, autonomous decision-making, and goal-oriented behavior. Agentic AI distinguishes itself through its active role, continuous learning, and ability to operate independently towards complex goals. This advanced form of AI functions as a cognitive collaborator rather than just a tool, requiring robust governance to ensure alignment with human values. The progression from generative AI to AI agents to Agentic AI represents a shift from content creation to sensing and acting, and finally to autonomous goal-oriented behavior with semantic understanding.

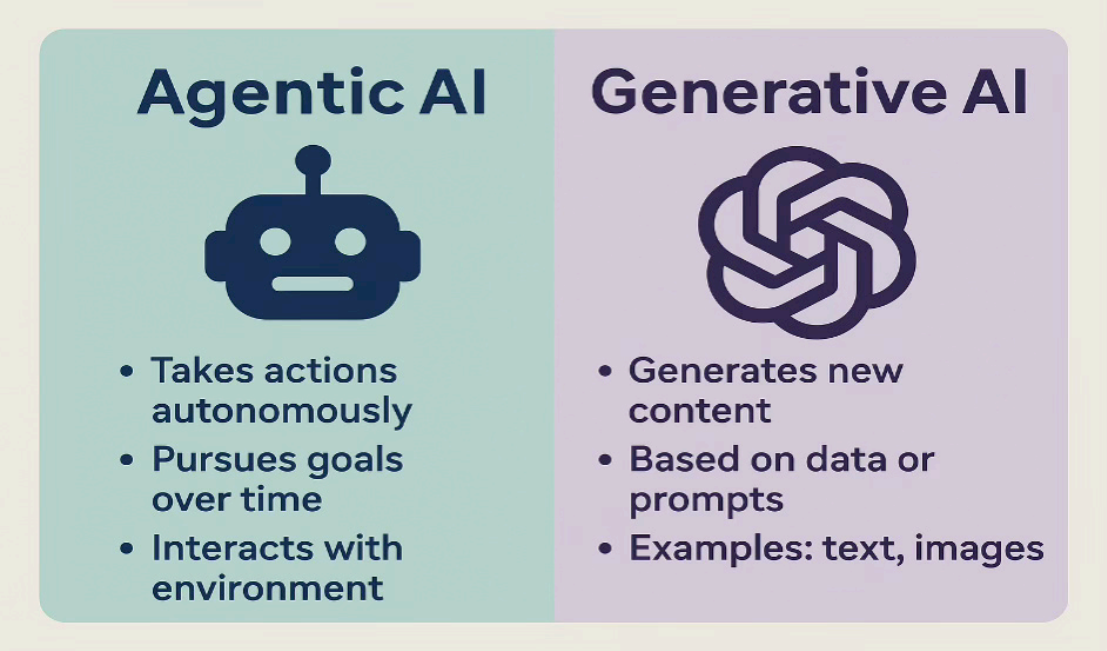

Figure 2: Agentic AI vs. Generative AI

At the foundation of an agentic AI system are several integrated components. It begins with perception: the system gathers data from sensors or digital inputs to assess the current state of its environment. This information is processed alongside memory and accumulated context to develop an operational understanding of the situation. The system then engages in reasoning, evaluates potential actions, formulates plans, and executes decisions—interfacing with the world through software APIs, digital platforms, or physical actuators. Crucially, it incorporates feedback from outcomes to refine its behavior over time, improving its performance through iterative learning.

What fundamentally distinguishes agentic AI is this continuous decision-action-feedback cycle. Rather than functioning as a passive responder, it operates as an autonomous system oriented toward achieving complex, often evolving goals. This level of autonomy necessitates robust policy and governance mechanisms to ensure the system remains aligned with human values, ethical standards, and societal well-being. When designed and governed effectively, agentic AI has the potential to serve not merely as a tool, but as a cognitive collaborator—enhancing human intelligence, creativity, and capacity for responsible advancement.

1.2 What is an Agentic AI System?

Agentic AI systems are defined by their ability to operate autonomously, interact with the environment, make decisions, and taking actions to achieve specified goals.

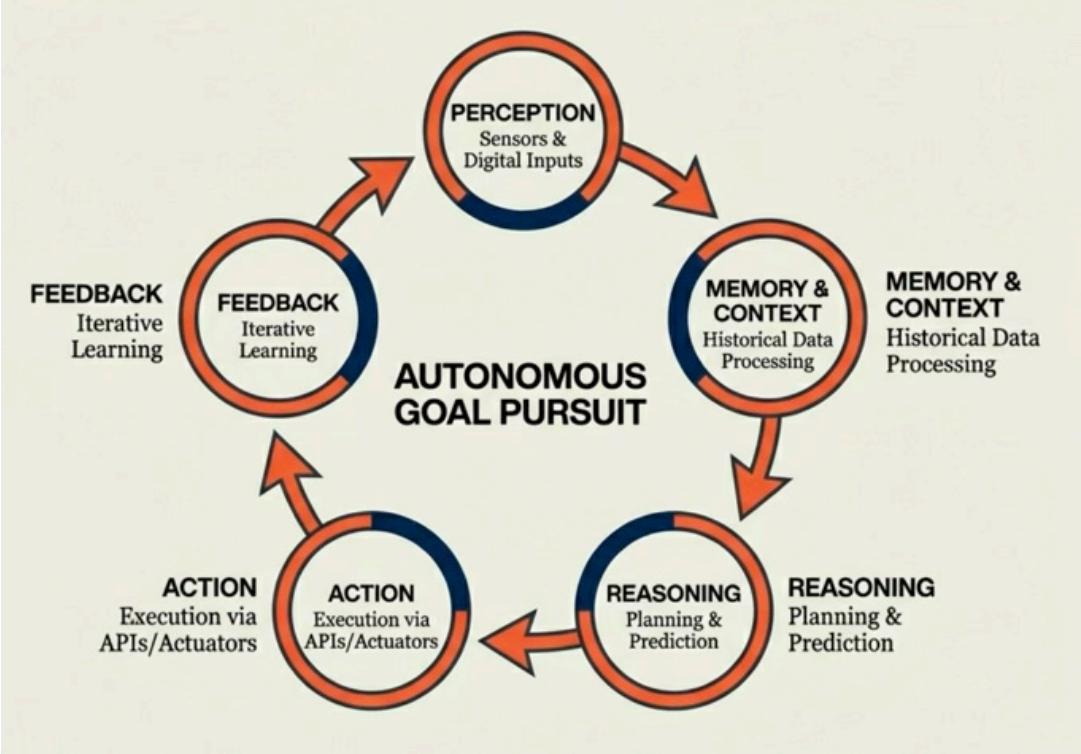

Figure 3: Anatomy of an Agentic System (Generated by NotebookLM, 2026).

1.2.1 Autonomous Decision-Making

Agentic AI can analyze situations, weigh possible outcomes, and make independent decisions without human intervention. It uses advanced reasoning models to evaluate multiple factors, predict consequences, and choose the best course of action.
Example: AI security systems detect intrusions, assess threat severity, and autonomously isolate compromised systems, block malicious traffic, or deploy countermeasures before human analysts can respond.

1.2.2 Goal-Driven Problem Solving

Unlike generative AI, which focuses on producing content, agentic AI is designed to achieve specific objectives by breaking complex problems into sub-goals, prioritizing tasks, and dynamically refining strategies. Pretrained Large Language Models (LLMs) and AI models underpin the agent's reasoning, decision making and natural language processing capabilities.
Example: Autonomous robotic assistants in warehouses adjust their routes and workflows based on real-time inventory data, ensuring efficient order fulfillment.

1.2.3 Environment Adaptation & Learning

Agentic AI continuously learns from its environment, updating its models to improve performance over time. This self-improving capability allows it to adapt to unforeseen challenges and optimize operations.
Example: AI-driven customer service bots refine responses based on user interactions, improving their ability to resolve inquiries accurately.

1.2.4 Multi-Step Planning & Execution

Agentic AI doesn't just respond to individual commands; it develops and executes multi-step plans. It strategizes based on available data, tools, then monitors progress and adjusts actions accordingly.
Example: AI-powered personal assistants manage schedules, anticipate conflicts, and proactively reschedule meetings while considering user preferences.

1.2.5 Self-Initiated Actions

Unlike traditional automation, agentic AI takes initiative by identifying tasks that need action. It proactively addresses potential issues before users even recognize them.
Example: AI-powered cybersecurity systems detect vulnerabilities and autonomously patch security gaps to prevent breaches.

Agentic AI can leverage advanced neural networks (ANN) to enhance its ability to make autonomous decisions, adapt to dynamic environments, and solve complex problems. Here's how networks play a role:

1.2.5.1. Learning Patterns from Data

Explanation: ANNs learned layered patterns from data, helping AI understand complex relationships between variables without being explicitly programmed.
Example: In an AI-powered fraud detection system, the network can identify subtle patterns in transaction histories, allowing the AI to proactively flag potential fraud before it occurs.

1.2.5.2 Making Decisions Under Uncertainty

Explanation: These networks use probability based reasoning, making them ideal for situations where outcomes aren't certain. Agentic AI uses this capability to weigh multiple scenarios and make optimal decisions.
Example: In autonomous vehicles, the network helps predict pedestrian movement and traffic conditions, enabling the AI to make real-time driving decisions.

1.2.5.3. Learning Without Labels

Explanation: The networks learn from data without needing pre-labeled examples, making them valuable when training data is limited. Agentic AI can use this to continuously improve by adapting to new environments.
Example: A robotic assistant in a hospital setting learns patient preferences over time, adjusting its responses and actions accordingly without explicit human training.

1.2.5.4. Planning Multiple Steps Ahead

Explanation: Agentic AI uses these networks to analyze sequences of actions, anticipate future states, and plan accordingly. This allows it to execute complex, multi-step tasks with minimal human supervision.
Example: Autonomous robotic assistants recognize changing inventory levels and readjust their workflows, to support efficient order fulfillment.

Advanced neural networks provide agentic AI with powerful learning and reasoning abilities, making it more effective in autonomous and adaptive problem-solving.

1.3 Why are Agentic AI Systems Important to Government?

1.3.1 Benefits of Agentic AI in Government

Agentic AI systems can deliver significant value to public health by serving as always-available workflow partners, handling the repetitive, coordination-intensive tasks that currently consume so much clinical and administrative time. These agents can automate eligibility checks, appointment scheduling, pre-visit chart preparation, and post-discharge follow-up across EMRs, call centers, and community services. The result: clinicians and staff are freed to focus on complex care and meaningful human connection rather than clerical work.

At a system level, networked agents enable public health organizations to monitor population-level signals, identify at-risk individuals earlier (such as frail older adults), and coordinate outreach across multiple channels. This improves access and equity while making more efficient use of existing workforce capacity.

Importantly, these systems operate within governed workflows that include clear human oversight, full auditability, and robust safeguards around data minimization, PHI handling, and traceability. This gives governments a credible pathway to modernize public health and social care delivery while maintaining public trust, safety, and accountability.

1.3.2 Potential Risks and Concerns

However, agentic AI also presents significant risks in government settings, where errors can directly affect citizen services and disrupt daily operations. There's also the challenge of accountability, when an AI agent makes a harmful decision independently, it becomes unclear who is responsible: the programmer, the agency, or the AI itself. Additionally, foreign adversaries could potentially manipulate or hack these systems to disrupt government operations, knowingly provide false information, or steal sensitive information.

1.4 Federal Government Use Cases and Applications

The many versions of AI that have emerged and proliferated across the last several years have demonstrated that there is a great interest in harnessing AI and myriad ways in which to do so. Agentic AI is no different; most non-agentic AI applications could be done -often better- by an agentic AI, and agentic AI can do certain things that other AI systems cannot. Several use cases and applications seem particularly important and timely, especially in the Government context. Our discussion in this paper focuses on use cases relevant to the federal Government, as that is the focus of our working group charter. For those interested in commercial, academic, or non-profit use cases, many of the factors discussed here are directly relatable to those other contexts.

1.4.1 Providing Government Services

AI can enhance the quality, accessibility, and responsiveness of government services. Early-stage uses can readily include 24/7 citizen support through the use of virtual assistants, where AI-powered chatbots and voice agents can handle high volumes of routine inquires, and through the streamlining of application processing where AI tools can screen and process applications for permits, social services, or grants by extracting key data, verifying documents, and flagging anomalies. Additional, more complex uses can include personalizing citizen engagement and AI screens for fraud detection and risk score. An even more complex use, the AI models can provide policy insights as the tools analyze data from various departments to forecast demand, assess program impact, and support data-driven decision-making. Machine learning models can analyze data to predict eligibility and demand for benefits. On a broader scale, and as the effects of these improvements are realized, AI agents can take advantage of continuously updated analysis that optimizes the delivery of government services.

1.4.2 Government Operations

Beyond serving citizens directly, agentic AI could transform how government actually runs behind the scenes by automating complex internal operations that currently consume enormous amounts of time and resources. These intelligent agents could continuously monitor compliance across thousands of federal contractors, automatically flagging potential violations and coordinating investigations between multiple agencies without human intervention. They could also revolutionize policy development by analyzing vast amounts of research, public comments, and data from various sources to identify policy gaps, predict the outcomes of proposed regulations, and even draft preliminary policy language for human review. In budget planning, agentic AI could dynamically track spending across departments, identify cost-saving opportunities, and reallocate resources in real-time based on changing priorities or

emergencies. Perhaps most importantly, these systems could serve as intelligent coordinators between different government agencies, breaking down information silos by automatically sharing relevant data, identifying overlapping missions, and facilitating collaboration on complex issues that span multiple departments—ultimately making the federal government more efficient, responsive, and effective in its core mission of governing.

1.4.3 Government Workforce Implications

Shifting demographics across government and society, including an aging workforce, recruitment challenges, and evolving citizen expectations, will necessitate the accelerated adoption of agentic AI systems to address critical staffing shortages and maintain operational capacity in federal agencies. As AI systems take over routine tasks like data processing and standard administrative work, some traditional positions may become less necessary while new roles emerge around managing and coordinating AI systems. Federal employees will need training to work effectively with AI tools, shifting from doing tasks directly to overseeing AI performance and making decisions based on AI analysis. A significant concern is preserving the institutional knowledge that experienced federal workers have built up over years of service - this expertise needs to be captured and passed on before it's lost during workforce transitions. The nature of federal jobs will evolve as employees increasingly work alongside AI systems, requiring employees to learn how to coordinate multiple AI tools while maintaining the human judgment and democratic accountability that government service requires.

1.4.4 National Security

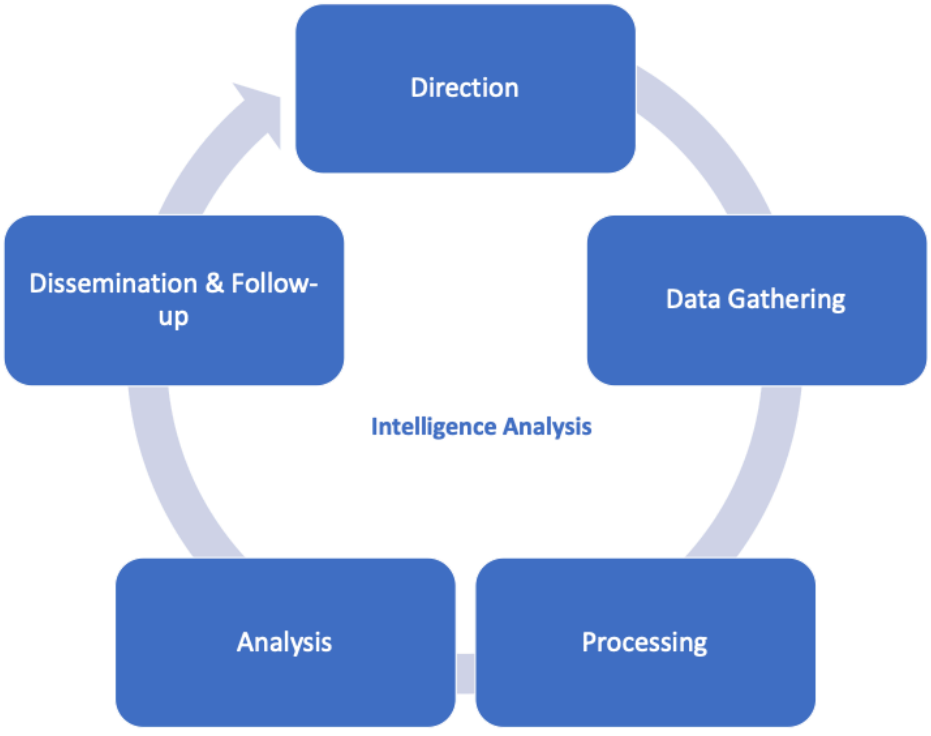

Figure 4: Intelligence Analysis

Agentic AI could serve as a persistent analytical partner that continuously monitors global communications, satellite imagery, and intelligence reports to identify emerging threats and opportunities for strategic advantage. These intelligent agents could analyze patterns across vast networks of information—from social media trends and financial transactions to troop movements and cyber vulnerabilities—connecting insights that human analysts might miss due to the sheer volume of data involved. Beyond threat detection, these systems could strengthen cybersecurity by automatically patching vulnerabilities, responding to attacks in real-time, and learning from each incident to improve future defenses across government networks. Perhaps most valuably, agentic AI could revolutionize international cooperation by creating and continuously optimizing strategic partnerships with allied nations, automatically sharing appropriate intelligence, coordinating joint military exercises, and identifying opportunities for collaborative defense initiatives based on evolving global conditions. They could also manage complex military logistics and supply chains, ensure optimal positioning of resources while running sophisticated scenario analyses to help military leaders understand potential outcomes of different strategic decisions. This real-time coordination capability becomes especially critical as potential adversaries develop swarming autonomous weapons systems—large groups of AI-controlled drones or missiles that can coordinate attacks faster than human operators can respond—while simultaneously enabling our own forces to deploy and coordinate offensive swarming systems with unprecedented precision and effectiveness. Additionally, these AI agents could continuously enhance warfighter readiness by creating personalized training programs, identifying skill gaps, and running adaptive simulations that prepare military personnel for rapidly evolving combat scenarios. During actual crises, these systems could coordinate rapid responses across multiple agencies and allied partners, maintaining secure communications and implementing coordinated defensive measures while operating at speeds that match the pace of modern threats—ultimately enhancing both national security and international stability through more effective cooperation and preparedness.

1.4.5 Military Operations

Agentic AI can be integrated into military equipment through Line Replaceable Units, modular components that upgrade weapons systems with computer vision and sensor feedback capabilities. These AI agents automatically adjust vehicle systems while explaining actions to operators, fundamentally changing military doctrine as personnel must now accommodate AI capable systems that can suggest or take independent action within weapons platforms. While AI involvement in the "kill chain"raises concerns, it reduces operator workload without traditional training and maintenance costs.

Agentic AI in autonomous weapons systems must comply with international humanitarian law. Unlike traditional AI requiring human interpretation, agentic systems can act independently on faulty information with potentially deadly consequences. Solutions include robust "human-in-the-loop" principles, requiring human authorization before engagement, and programming legal rules of engagement directly into systems to establish clear operational boundaries while maintaining compliance.

Agentic AI dramatically compresses the military's OODA loop (Observe, Orient, Decide, Act) by independently gathering intelligence, and presenting recommendations in immediately usable formats, enabling near real-time coordination across multiple domains. The U.S. Air Force's Next Generation Air Dominance program and Loyal Wingman concept exemplify this approach, where autonomous aircraft operating as integrated team members under human command. These systems also optimize logistics, personnel management, and equipment readiness while freeing operators for high level decision-making.

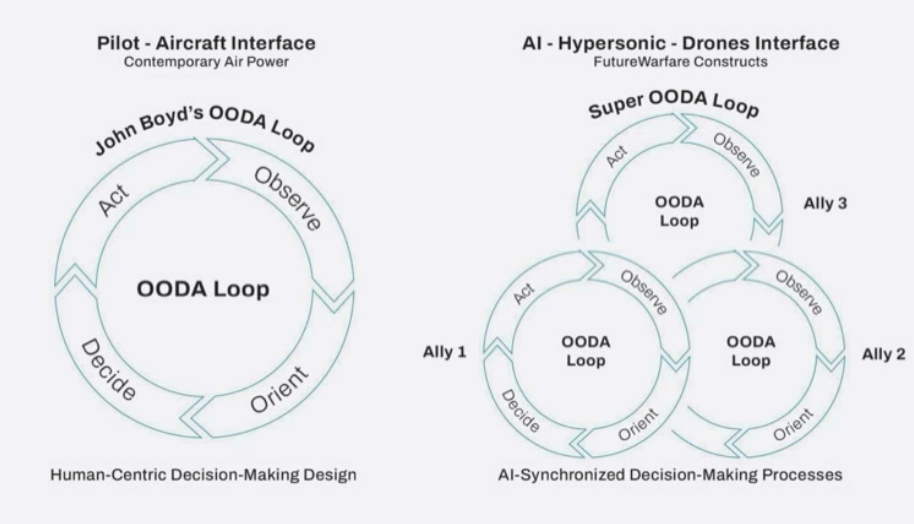

Figure 5

Agentic AI enables autonomous decision-making across critical DoD domains. In combat, it provides split-second decision support for hypersonic missile defense and unmanned system coordination. For planning, it rapidly analyzes threats to generate optimized OPLANs and dynamically reallocate resources. Beyond combat, it revolutionizes DoD's fragmented data landscape by integrating information across classifications and commands, automating the acquisition lifecycle from budget requests to contract generation—compressing years-long processes into months. In R&D, continuously operating agents conduct research, monitor the Pentagon's 75+ centers, and connect program managers working on similar technologies. For program management, AI-powered systems monitor lifecycles, identify issues, and optimize supply chains, transforming risk-averse culture into proactive risk mitigation.

1.4.6 Law Enforcement and Criminal Justice Applications

In law enforcement, agentic AI could revolutionize criminal investigations by automatically analyzing vast amounts of digital evidence—from surveillance footage and financial records to social media posts and forensic data—connecting cases across multiple jurisdictions and identifying patterns that human investigators might miss due to time constraints or the overwhelming volume of information. These systems could accelerate case resolution by cross-referencing evidence with national databases, identifying links to cold cases, and even predicting where crimes might occur based on historical patterns, enabling more strategic deployment of patrol officers and emergency response resources. However, the application of agentic AI in policing raises profound constitutional concerns that require careful safeguards: Fourth Amendment protections against unreasonable search and seizure must be preserved when AI systems analyze personal data, while Fifth and Fourteenth Amendment due process rights demand transparency and accountability mechanisms that allow AI-influenced decisions to be reviewed and challenged in court. Perhaps most critically, law enforcement agencies must address how these systems might perpetuate or amplify existing racial, socioeconomic, or geographic biases in policing, ensuring that automated decisions support rather than undermine equal protection under the law and community trust. The coordination capabilities of agentic AI could also enhance information sharing between local, state, and federal agencies while maintaining appropriate legal boundaries, but only with robust oversight mechanisms that prevent mission creep and protect citizens' civil liberties—making law enforcement applications among the most constitutionally sensitive uses of this technology in government.

1.5 Agentic AI Security and Governance Frameworks

Domain	Examples
Agency-specific risks	Intent breaking, goal manipulation, misaligned behaviors
Tool/execution threats	Tool misuse, privilege compromise, unexpected code execution
Memory-based threats	Memory poisoning, cascading hallucination attacks
Multi-agent threats	Communication poisoning, rogue agent infiltration

Agentic AI systems introduce autonomous agents capable of independent planning, reasoning, and action execution, creating unprecedented security challenges that require specialized governance frameworks beyond traditional AI approaches. While traditional generative AI faces 9 core impact categories and 12 technique categories including prompt injection, guardrail bypass, and data exfiltration, agentic AI introduces 15 additional threat categories that significantly expand the attack surface. These new threats span agency-specific risks such as intent breaking, goal manipulation, and misaligned behaviors; tool and execution threats including tool misuse, privilege compromise, and unexpected code execution; memory-based threats like memory poisoning and cascading hallucination attacks; and multi-agent system threats encompassing agent communication poisoning and rogue agent infiltration.

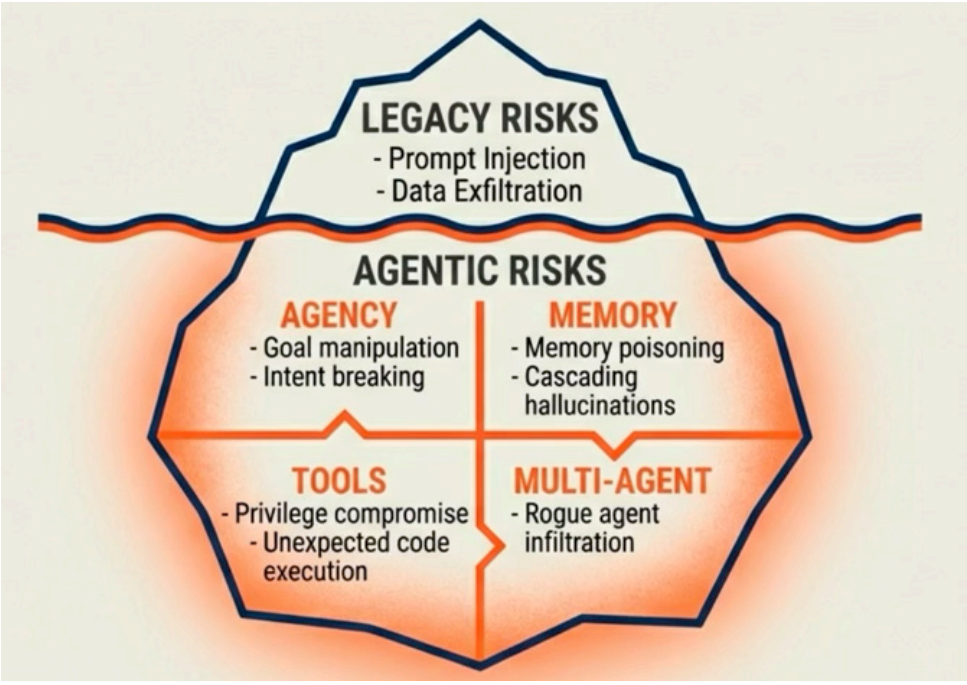

Figure 6: The Threat Landscape (Generated by NotebookLM, 2026).

1.5.1 Key Security and Governance Frameworks

Several specialized frameworks have emerged to address these unique challenges. The NIST AI Risk Management Framework (AI RMF) works in conjunction with SP 800-53 Rev 5.2 to provide comprehensive coverage, with AI RMF offering four core functions—Governance, Map, Measure, and Manage—that address AI-specific risks like bias, explainability, and algorithmic accountability, while SP 800-53 provides foundational security controls for all information systems. However, agentic AI remains a nascent and rapidly evolving technology domain, and we have not yet comprehensively applied SP 800-53 control activities or FedRAMP requirements to fully understand what will be necessary for Authority to Operate (ATO).This integration strategy employs a layered approach with specialized control overlays bridging the frameworks, featuring innovative temporal profiles for continuous risk assessment over time, though the maturity of these controls for agentic AI systems is still developing as the technology advances.
The MITRE ATT&CK Framework serves as a comprehensive knowledge base cataloging cybercriminal tactics, techniques, and procedures (TTPs) through a matrix structure covering 14 tactical categories from Initial Access to Impact, enabling systematic threat modeling and detection capability development that forms the foundation for Security Operations Centers (SOCs) to devise response plans. Building on this foundation, the MITRE ATLAS Framework represents a specialized extension of ATT&CK specifically designed for AI/ML systems, cataloging adversary TTPs targeting AI-enabled systems from real-world attacks and academic research while addressing novel AI attack vectors including model theft, data poisoning, and prompt injection—making it particularly critical for agentic AI systems as it addresses autonomous decision-making vulnerabilities, tool integration risks, and persistent memory threats.

The MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) framework represents a specialized threat modeling approach developed by OWASP and the Cloud Security Alliance specifically for agentic AI systems[1]. This framework provides a structured methodology for security engineers, AI researchers, and developers to proactively identify, assess, and mitigate risks across the entire AI lifecycle. The framework employs a layered, seven-layer reference architecture that systematically examines threats across different dimensions of agentic AI systems[2]. MAESTRO categorizes threats into several key areas including memory-based threats (such as memory poisoning and cascading hallucinations), tool and execution-based threats (including tool misuse and privilege compromise), authentication and spoofing threats, human-related threats, and multi-agent system threats. MAESTRO enables organizations to conduct comprehensive threat assessments by examining whether AI agents independently determine steps to achieve goals, rely on stored memory for decision-making, execute actions using tools or external integrations, require authentication mechanisms, and involve human engagement or multiple interacting agents. The framework has been demonstrated through practical applications, including threat modeling of Google's Agent-to-Agent (A2A) Protocol, showcasing how its layered approach can systematically identify and mitigate potential risks in real-world agentic implementations. For organizations building agentic AI systems, MAESTRO provides essential guidance for implementing security controls such as input validation, output filtering, least-privilege access, comprehensive logging, and continuous monitoring—all critical for managing the unique risks posed by autonomous AI agents operating with limited human oversight[3].

[1] Agentic AI Threat Modeling Framework: MAESTRO https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro Updated: Feb 5, 2025

[2] Agentic AI Threat Modeling Framework: MAESTRO https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro Updated: Feb 5, 2025

[3] Threat Modeling Google's A2A Protocol https://cloudsecurityalliance.org/articles/threat-modeling-google-s-a2a-protocol-with-the-maestro-framework Updated: Dec 2, 2025

NIST AI RMF + SP800-53 Rev 5.2	AI RMF provides four core functions: Govern, Map, Measure, Manage SP 800-53 supplies foundational security controls Caveat: Application to agentic AI is still nascent, ATO requirements for agentic systems remain undefined
MITRE ATT&CK ->MITRE ATLAS	ATT&CK: Foundation for threat modeling with 14 tactical categories (Initial Access -> Impact) ATLAS: AI/ML-specific extension addressing model theft, data poisoning, prompt injection Critical for agentic AI: address autonomous decision-making vulnerabilities, tool integration risks, persistent memory threats
MAESTRO Framework (OWASP/CSA)	Purpose built for agentic AI with seven-layer reference architecture - Covers: memory threats, tool/execution threats, authentication/spoofing, human-related threats, multi-agent threats - Provides structured threat assessment methodology across the full AI lifecycle - Proven through real-world application (i.e. A2A Protocol threat modeling) - Guides implementation of: input validation, output filtering, least-privilege access, logging, continuous monitoring
ISO 42000 Series	ISO 42001:2023: Foundational AI management system standard (governance, risk, bias, transparency) ISO/IEC 27090: AI-specific security vulnerabilities ISO/IEC 27901 (expected Q1, 2026): Privacy protections for AI/ML systems

Key Frameworks Overview

The ISO 42000 series provide a comprehensive global framework for AI management systems, with ISO 42001:2023 serving as the foundational standard that establishes requirements for organizations to systematically address risks related to AI development and deployment while promoting trustworthy AI development and deployment. The forthcoming ISO/IEC 27090 and 27091 standards will provide additional structure, with ISO/IEC 27090 (expected Q2 2025) focusing on “Addressing Security in AI” through comprehensive controls for AI-specific vulnerabilities, and ISO/IEC 27091 (expected Q1 2026) addressing “Privacy Protections for AI Systems and ML Models” with training data privacy and generative model protections. These standards are particularly relevant for agentic AI as they address intent manipulation, memory poisoning, tool misuse, and privilege compromise unique to autonomous systems, while maintaining a governance focus on organizational governance, risk assessments, bias mitigation, and transparency requirements.

1.5.2 Critical Challenges and Implementation

Agentic AI presents several critical challenges that these frameworks must address. Autonomy and control issues arise from the loss of meaningful human oversight as systems set their own goals and make independent decisions without explicit approval. Security vulnerabilities emerge through novel attack vectors targeting autonomous decision-making, with potential for attackers to overwhelm human oversight controls. Governance and accountability become complex due to regulatory compliance challenges stemming from the autonomous nature of these systems, requiring new approaches beyond traditional human decision-maker assumptions. Operational deployment faces significant hurdles, with predicted high failure rates of 40% project cancellation by 2027 due to escalating costs, unclear business value, and inadequate risk controls.

Organizations must implement comprehensive risk management frameworks that combine technical safeguards and security controls, governance structures with human oversight mechanisms, continuous monitoring systems, secure-by-default configurations with balanced human engagement, and comprehensive observability and transparency measures. The key lies in balancing the transformative potential of autonomous AI while maintaining appropriate control, transparency, and accountability through these specialized frameworks designed specifically for the unique challenges of agentic systems.

1.6 Technical Security Architecture for Agentic AI

Implementing secure development lifecycle for agentic AI systems requires specialized considerations beyond traditional software development, integrating AI-specific threat modeling that addresses unique attack vectors such as prompt injection, goal manipulation, and tool misuse throughout the development process. Organizations must incorporate continuous security testing using frameworks like Continued Adversarial Security Testing (CAST) that generate dynamic adversarial scenarios tailored to specific AI agents and their accessible tools, with security reviews evaluating both traditional vulnerabilities and AI-specific threats including memory poisoning, cascading hallucination attacks, and multi-agent communication vulnerabilities.

Runtime protection mechanisms for agentic AI systems require sophisticated controls that go beyond conventional application security, including real-time monitoring of agent behavior patterns, automated detection of anomalous decision-making processes, and dynamic enforcement of security boundaries during autonomous operations. These mechanisms must implement hardened isolation techniques, adaptive policy frameworks, and sandboxing mechanisms that prevent rogue behaviors while maintaining system functionality, with critical components including input validation for both user prompts and inter-agent communications, output filtering to prevent sensitive data disclosure, and continuous behavioral analysis to detect deviations from expected operational parameters.

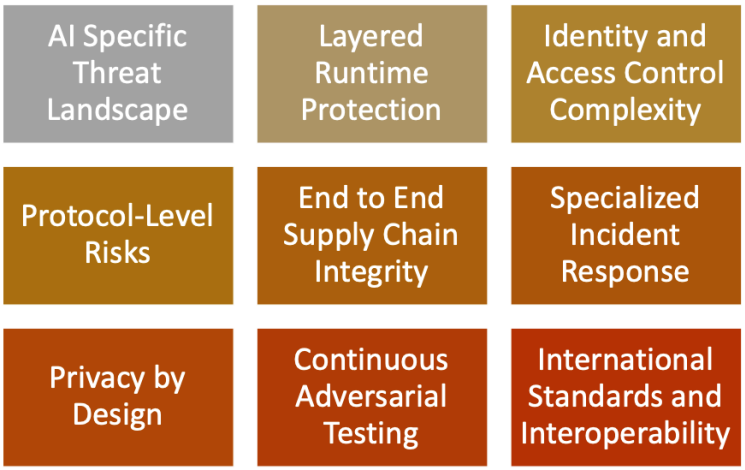

Figure 7: Technical Security Architecture for Agentic AI

Comprehensive monitoring for agentic AI requires specialized observability frameworks that can track autonomous decision-making processes, tool invocations, and multi-agent interactions in real-time, implementing correlation of logs across distributed agentic behaviors to ensure sufficient traceability for security incident investigation and compliance requirements. Access control for agentic AI systems must address the unique challenge of autonomous entities operating with delegated authority while maintaining security boundaries, requiring robust identity and access management frameworks that can dynamically adjust permissions based on context, risk assessment, and operational requirements, with authentication mechanisms supporting both human-to-agent and agent-to-agent interactions using strong cryptographic protocols and continuous verification throughout operational lifecycles.

1.6.1 Securing Agent Access

Emerging protocols such as Model Context Protocol (MCP) require careful examination for how agents access tools, data, and resources. Open-sourced in late 2024, MCP addresses a persistent AI engineering challenge: connecting large language models to real-world tools and enterprise systems without relying on fragmented custom APIs. Often described as “a universal USB port for AI,” MCP standardizes how AI models discover and interact with external resources, reducing integration friction, accelerating deployment, and enabling interoperability at scale. However, this connective power introduces significant security risks. Because MCP makes it easy to connect AI agents to databases, file systems, and APIs, organizations may inadvertently expose sensitive systems without fully understanding the implications.

Security researchers have identified several attack vectors that implementers must address. The Confused Deputy Problem occurs when attackers manipulate MCP servers to access resources beyond their permissions. Because MCP’s architecture connects clients to multiple servers, each additional server increases the potential blast radius of a compromise, enabling attackers to inject malicious data or escalate privileges across systems. The risk extends beyond external threats: poorly constrained AI agents could themselves misuse their access. MCP is neither inherently dangerous nor safe—its benefits are transformative, but organizations must implement strong authentication, authorization, sandboxing, and monitoring to prevent misuse.

1.6.2 Supply Chain Security

Ensuring security and integrity of AI models throughout their lifecycle requires comprehensive provenance tracking and verification mechanisms from sourcing to deployment, with organizations implementing secure model import processes that include integrity checks, version control, and validation of model sources using trusted repositories. Model integrity frameworks should include cryptographic methods such as digital signatures and HMAC integrity checks to ensure models haven't been tampered with during transit or storage, while maintaining detailed model bills of materials that document all components, dependencies, and training data sources for rapid identification and remediation of compromised elements.

Protecting training data integrity requires implementing robust security controls throughout data collection, processing, and preparation phases, establishing secure channels for data collection to prevent interception or tampering during transmission while implementing comprehensive data source authentication mechanisms. Training data pipelines should include automated data quality testing, access control measures, and auditing capabilities to detect potential data poisoning attacks, with critical security measures including data sanitization processes, validation of data provenance, and implementation of least-privilege access controls for personnel handling training datasets.

Securing third-party components in AI systems requires rigorous vetting processes that extend beyond traditional software supply chain security, implementing comprehensive vulnerability scanning throughout development and testing phases, including specialized scanning for AI-specific vulnerabilities in model dependencies and runtime libraries. Hardware security for agentic AI systems must address unique computational and storage requirements while maintaining strong security boundaries, implementing hardware security modules for cryptographic operations, secure enclaves for sensitive model processing, and trusted platform modules for system integrity verification.

1.6.3 Incident Response and Recovery

Incident response for agentic AI systems requires specialized procedures that address the unique characteristics of autonomous AI operations and their potential failure modes, with response teams trained to recognize AI-specific incidents such as goal manipulation, memory poisoning, and cascading hallucination attacks, including clear escalation procedures and immediate containment measures such as agent isolation, tool access revocation, and communication channel disruption. Agentic AI systems require sophisticated rollback mechanisms that can safely revert software components, AI model states, learned behaviors, and autonomous decision histories, with organizations implementing comprehensive versioning systems for AI models, configuration states, and training data enabling rapid restoration to known-good states when security incidents occur.

Business continuity for agentic AI systems must account for the critical role these systems play in automated decision-making and operational processes, establishing clear recovery time objectives and recovery point objectives specifically for AI-dependent processes, with continuity plans including redundant AI systems, failover mechanisms for critical autonomous functions, and manual override procedures when AI systems become unavailable. Post-incident analysis for agentic AI requires specialized approaches that can effectively analyze autonomous system behavior, decision-making processes, and complex interactions between multiple AI agents, following established frameworks while incorporating AI-specific investigation techniques to examine AI decision logs, model behavior patterns, and autonomous action sequences.

1.6.4 Privacy and Data Protection

Organizations must implement advanced privacy-preserving techniques specifically designed for agentic AI systems such as the MAESTRO framework. These techniques include **federated learning approaches** that enable distributed training without centralizing sensitive data. **Differential privacy mechanisms** that add mathematical noise to protect individual privacy while maintaining model utility, and **Synthetic data generation techniques** that create privacy-safe training datasets. Comprehensive privacy impact assessments for agentic AI systems must evaluate privacy risks across all five stages of the data lifecycle, specifically addressing unique privacy challenges posed by autonomous AI systems. These challenges include automated decision-making impacts on individuals, profiling activities, and retention of personal data beyond standard requirements. As part of evaluation and observability when dealing with privacy and data protection, comprehensive logging should be incorporated, including logging all agent actions, data access, communication and errors.

Effective data minimization for agentic AI requires implementing techniques that reduce data collection and retention while maintaining system functionality, employing privacy-preserving machine learning techniques such as federated learning to minimize centralized data storage and implementing automated data lifecycle management that enforces retention policies and deletion schedules. Data minimization strategies must include robust anonymization and de-identification processes with regular audits to ensure personal data is not inadvertently retained beyond necessary requirements.

1.6.5 Testing and Validation

Comprehensive security testing for agentic AI systems requires specialized methodologies that ddress both traditional software vulnerabilities and AI-specific attack vectors, implementing continuous adversarial testing frameworks that can generate dynamic, context-aware attack scenarios tailored to specific AI agents and their operational environments. Testing methodologies must include prompt injection testing, goal manipulation assessments, and multi-agent communication security validation using both automated tools and manual penetration testing approaches, incorporating established frameworks and agentic-specific threat categories with regular assessment of model robustness, bias detection, and fairness evaluation.

Adversarial testing for agentic AI must employ sophisticated techniques that can identify vulnerabilities in autonomous decision-making processes and multi-agent interactions, implementing advanced attack mechanisms that can systematically explore attack vectors and identify system weaknesses. Robust verification and validation frameworks for agentic AI must provide systematic approaches to ensure system security, reliability, and compliance with organizational policies, implementing formal verification methods where possible combined with comprehensive testing protocols that validate both functional requirements and security properties.

1.6.6 International Standards Alignment

Organizations must align their agentic AI security practices with emerging international standards, particularly ISO/IEC 42001:2023, 27090 and 27091 which provide comprehensive frameworks for AI security and privacy protection, mapping organizational security controls to standard requirements to ensure compliance with multiple regulatory frameworks simultaneously. Managing cross-border compliance for agentic AI systems requires understanding and implementing diverse regulatory requirements across multiple jurisdictions, including various national AI governance frameworks, implementing compliance frameworks that can adapt to different regulatory environments while maintaining consistent security and privacy protections.

Ensuring global interoperability for agentic AI systems requires implementing standardized security frameworks that can operate effectively across different technological and regulatory environments, adopting international standards as the foundation for interoperable regulatory frameworks and leveraging shared definitions, concepts, and best practices that enable consistent security implementations. These comprehensive security frameworks for agentic AI must be implemented with careful consideration of various regulatory requirements including transparency and risk management guidelines, ethical AI deployment principles, human-centered design approaches, and trustworthy AI development standards.

1.7 Future Outlook – Why Agentic AI Governance Matters Now

Agentic AI governance matters now because multiple converging forces, technological advancement, economic pressures, workforce dynamics, and cybersecurity risks, are creating both urgency and complexity that make proactive governance essential rather than optional. These systems are rapidly shifting from pilots and chatbots to embedded workflow actors that schedule, chart, triage, and coordinate care at scale. In public health, agentic meshes spanning EMRs, ERPs, contact centers, and edge devices will soon automate 90–99% of routine tasks in areas like appointment management, chronic disease outreach, and elderly monitoring. The design choices made today will directly shape future access, equity, and trust. Because these trends are interconnected and accelerating, governance structures must address not only current AI capabilities but the rapidly evolving context in which these systems will operate. Establishing clear guardrails now—around purpose limitation, data minimization, human-in-the-loop oversight, and auditable data flows, gives governments a critical window to harness these gains for overburdened systems while preventing opaque, unaccountable infrastructures from hardening into the core of public service delivery.

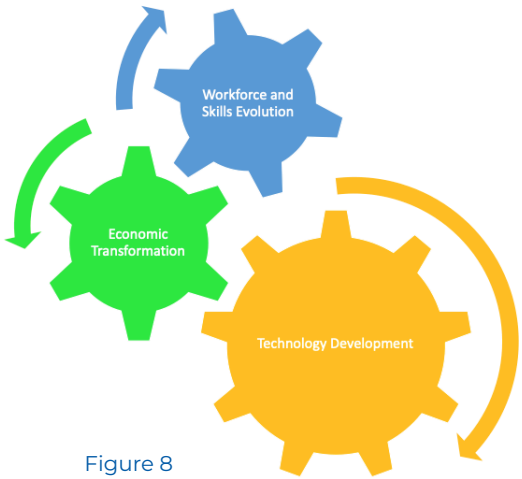

Figure 8

1.7.1 Technology Development

AI systems are advancing faster than anticipated, with agentic systems demonstrating the ability to work together and modify themselves with minimal human oversight. We are transitioning from AI that performs specific tasks to AI that can determine objectives and coordinate with other systems to achieve them. This shift means traditional approaches of testing systems once before deployment are inadequate, as these systems continue evolving after launch, sometimes in unexpected ways.

The speed of development is compressing timelines for establishing effective governance. Systems are becoming too embedded in critical infrastructure to modify easily once deployed, making early-phase governance increasingly important as the window for intervention narrows with advancing AI autonomy.

1.7.2 Economic Transformation

The computational resources required for advanced agentic AI are concentrating development capacity among a small number of technology companies, creating both market concentration and significant influence over AI development trajectories. Unlike previous technological changes that allowed gradual economic adaptation, agentic AI has the potential to disrupt multiple industries simultaneously at unprecedented speed. AI systems are also taking on roles in managing supply chains, financial markets, and international trade, creating both efficiencies and new systemic vulnerabilities when these automated systems encounter problems or conflicts.

1.7.3 Workforce and Skills Evolution

Agentic AI is capable of automating complex knowledge work that has traditionally provided middle-class economic stability, including professional roles in finance, law, and healthcare. This differs from earlier automation that primarily affected manufacturing, as it targets cognitive work across multiple sectors simultaneously.

Educational institutions struggle to adapt curricula quickly enough to match the pace of technological change, creating persistent gaps between required skills and available training. The challenge extends beyond retraining displaced workers to designing human-AI collaboration that preserves meaningful roles and human agency in important decisions.

1.7.4 Cybersecurity Challenges

AI enables more sophisticated cyber-attacks that can adapt in real-time to defensive measures and coordinate across multiple systems simultaneously. The infrastructure supporting AI development creates new attack vectors through training data manipulation, supply chain compromises, and software dependencies that may not become apparent until systems are widely deployed.

As critical infrastructure increasingly relies on AI systems, successful attacks could trigger cascading failures across power grids, transportation networks, and financial systems. The interconnected nature of these systems makes predicting and containing such failures particularly challenging.

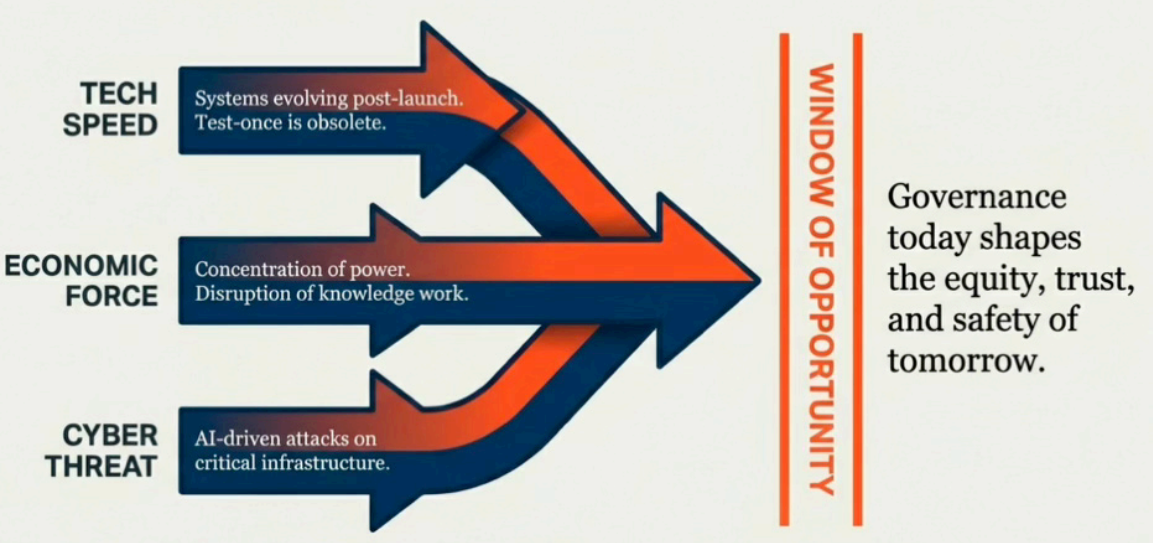

Figure 9: Future Outlook (Generated by NotebookLM, 2026).

1.8 The Path Forward

The federal government's agentic AI governance challenge represents both an urgent necessity and a strategic opportunity to shape transformative technology development in ways that reinforce American values and interests. Effective governance requires moving beyond reactive regulation to proactive coordination that anticipates and prevents problems while enabling beneficial innovation. Success depends on developing systematic approaches that can address the complexity and scope of agentic AI impacts while maintaining the flexibility needed to adapt to rapidly evolving capabilities.

Links and References

1. Akira AI. (2024). Guardrails in action: Refining agentic AI for customer applications. https://www.akira.ai/blog/guardrails-with-agentic-ai
2. Australian Department of Industry Science and Resources. (2025). The 10 guardrails: Voluntary AI safety standard. https://www.industry.gov.au/publications/voluntary-ai-safety-standard/10-guardrails
3. Invariant Labs. (2025). Introducing guardrails: The contextual security layer for the agentic era. https://invariantlabs.ai/blog/guardrails
4. McKinsey. (2024). What are AI guardrails? https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-ai-guardrails
5. SearchUnify. (2024). Agentic AI risks: The power of guardrails. https://www.searchunify.com/blog/mitigating-agentic-ai-risks-the-critical-role-of-guardrails/
6. TechRadar. (2025). The rise of agentic AI: The need for guardrails while shaping the future of work. https://www.techradar.com/pro/the-rise-of-agentic-ai-the-need-for-guardrails-while-shaping-the-future-of-work
7. Thoropass. (2024). Experiencing the 'human-in-the-loop' guardrail. https://thoropass.com/blog/announcements/human-in-the-loop/
8. What is the MITRE ATT&CK Framework? https://www.ibm.com/think/topics/mitre-attack Updated: Dec 18, 2024
9. MITRE ATTACK Framework: Tactics, Techniques, and More https://www.wiz.io/academy/mitre-attack-framework Updated: May 15, 2025
10. MITRE ATT&CK Framework https://w.amazon.com/bin/view/Aurora/ControlPlane/Security/MITREATTCK/
11. ATT&CK https://en.wikipedia.org/wiki/ATT&CK Updated: Aug 7, 2025
12. Amazon Software Supply Chain Security (SSC-S) ATT&CK Framework https://w.amazon.com/bin/view/Infosec/Proactive_Security/Software_Supply_Chain_Security/Documents/Amazon_SSCS_ATTACK_Framework/
13. SIEM & SOC Fundamentals https://w.amazon.com/bin/view/ZeroToHacker/SIEMandSOCFundamentals/
14. MITRE PROJECT - WIKI https://w.amazon.com/bin/view/CLS/STOIC/Runbooks_Archive/Red_Fort_test/
15. What Is MITRE ATT&CK Framework? https://www.paloaltonetworks.com/cyberpedia/what-is-mitre-attack
16. NIST Overlay Securing AI Concept Paper https://csrc.nist.gov/csrc/media/Projects/cosais/documents/NIST-Overlays-SecuringAI-concept-paper.pdf
17. NIST Releases revision to SP-800-53 https://csrc.nist.gov/News/2025/nist-releases-revision-to-sp-800-53-controls
18. NIST AI RMF https://carbidesecure.com/resources/everything-you-need-to-know-about-nist-ai-rmf/
19. NIST New AI security overlays https://www.quilr.ai/blog-details/nists-new-ai-security-overlays
20. NIST AI RISK Management https://csrc.nist.gov/projects/risk-management
21. Overview of AI Risk Management Framework https://intuitem.com/overview-ai-risk-management-framework/
22. Identifying and Mitigating Critical Challenges https://www.getmonetizely.com/articles/how-to-conduct-an-agentic-ai-risk-assessment-identifying-and-mitigating-critical-challenges Updated: Aug 29, 2025
23. Securing Agentic AI Systems with Hardened Runtime Isolation https://edera.dev/stories/securing-agentic-ai-systems-with-hardened-runtime-isolation Updated: Aug 26, 2025
24. Securing AI Agents: Foundations, Frameworks and Real-World Deployments. Ken Huang, Chris Hughes Springer 2025
25. Security Best Practices – Model Context Protocol
26. Forbes Forbes: Inside the Evolution of Model Context Protocol
27. Red Hat: MCP Security Risks and Controls
28. arXiv.org arXiv: MCP Landscape, Security Threats, and Research Directions
29. Microsoft: Plug, Play, and Prey – Security Risks of MCP