DEFENSE, EMERGING TECHNOLOGY, AND STRATEGY PROGRAM

The Rise of Agentic AI

Infrastructure, Autonomy, and America's Cyber Future

Yam Atir

HARVARD Kennedy School
BELFER CENTER

50 YEARS OF RESEARCH, POLICY, AND LEADERSHIP

JUNE 2025

Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.belfercenter.org

Statements and views expressed in this report are solely those of the author(s) and do not imply endorsement by Harvard University, Harvard Kennedy School, or the Belfer Center for Science and International Affairs.

DEFENSE, EMERGING TECHNOLOGY, AND STRATEGY PROGRAM

The Rise of Agentic AI

Infrastructure, Autonomy, and America's Cyber Future

Yam Atir

HARVARD Kennedy School
BELFER CENTER

50 YEARS
OF RESEARCH, POLICY, AND LEADERSHIP

JUNE 2025

About the Defense, Emerging Technology, and Strategy Program

The Defense, Emerging Technology, and Strategy (DETS) program has a dual mission to

1. advance policy-relevant knowledge and strategy on the most important challenges at the intersection of security and emerging technology; and
2. prepare future leaders for public service in relevant arenas.

The DETS program focuses on defense policy issues, public sector strategy execution, and new technologies that have emerged as pivotal to the future of international security. Through its programming, the DETS program seeks to train a new generation of technology-savvy policy and strategy leaders within the Kennedy School.

About the Author

Yam Atir is a technology policy researcher and Mid-Career Master in Public Administration graduate from the Harvard Kennedy School. She leads research on AI governance at the Mossavar-Rahmani Center for Business and Government, focusing on dynamic governance systems, industry engagement, digital infrastructure, and regulatory strategy. She is also a fellow at the AI venture studio at MIT Media Lab.

Previously, Atir served as Vice President for Strategy and Policy at a leading technology think tank in Tel Aviv and worked as a consultant to government agencies on AI implementation and innovation models. Over the past decade, she has worked across government, civil society, and international organizations—including the OECD—on challenges related to next-generation policymaking, digital sovereignty, and emerging technologies.

Atir has a background in government, having worked alongside Israeli President and Nobel Peace Prize laureate Shimon Peres and served in the Prime Minister's Office. She was recognized by Forbes Israel as one of the country's most influential young leaders.

Table of Contents

Belfer Center/ Chat GPT

Executive Summary

The rise of agentic artificial intelligence marks a critical inflection point in the digital landscape. Unlike generative AI models that passively produce content, agentic AI systems are autonomous, goal-driven entities capable of initiating actions, using external tools, collaborating with other agents, and completing complex, real-world tasks with minimal human oversight. These systems are no longer experimental. Platforms like OpenAI's Operator, Microsoft's Copilot Studio, and Google's A2A protocol are already transforming enterprise workflows and are on the cusp of integration into healthcare, infrastructure, and defense.

While agentic AI promises immense productivity gains, it introduces a dramatically expanded cybersecurity threat surface. These agents can execute transactions, access sensitive APIs, retain memory across sessions, and operate continuously in high-stakes environments. If compromised, they pose risks, not just to data, but also to physical infrastructure, public systems, and democratic oversight. Moreover, today's agentic systems are being built atop proprietary architectures governed by a handful of private firms, with little public transparency or accountability.

This policy brief argues that the United States must act urgently to shape the foundational rules, standards, and infrastructure of agentic AI. It recommends a strategic policy roadmap, anchored in cybersecurity, to ensure that these systems are safe, resilient, and aligned with democratic values. The Office of Science and Technology Policy (OSTP), working with CISA, NIST, and other agencies, has a narrow window of opportunity to establish governance over this emerging layer of digital infrastructure before default norms are set by private actors or adversarial states.

The White House Office of Science and Technology Policy (OSTP) is uniquely positioned to lead the national response to agentic AI. As the primary body coordinating science and technology policy across federal agencies, OSTP holds the convening authority to align disparate stakeholders, ranging from NIST and CISA to DARPA, NSF, and federal procurement bodies. Its mandate includes setting cross-agency priorities, shaping national R&D strategy, and advising the President on emerging technologies. Given the systemic implications of agentic

AI for cybersecurity, public infrastructure, and democratic oversight, OSTP is the only entity with both the strategic purview and policy leverage to orchestrate a whole-of-government approach before de facto standards are cemented by the private sector. Its leadership is essential to ensure that agentic AI systems are secure, interoperable, and accountable to the public interest.

Key Recommendations

- **Cross-Cutting Recommendation:**
    - **Lead Agency Coordination:** Designate OSTP to Orchestrate a Whole-of-Government Response. As the body responsible for aligning science and technology policy across federal agencies, OSTP should lead the national governance strategy for agentic AI, coordinating entities like NIST, CISA, DARPA, and the NSF to build coherent, forward-looking oversight structures.
    - **Incentivize Secure and Open Agent Infrastructure:** The federal government should invest in open, auditable frameworks that reduce reliance on proprietary systems. Through R&D funding, procurement standards, and public-private partnerships, the U.S. can promote a resilient, innovation-friendly ecosystem that prioritizes security and democratic values.
- **Immediate (Next 3–6 Months): Establish a National Registry of Agentic Systems**
    - The federal government should create a national registry of agentic AI systems deployed in critical sectors such as healthcare, finance, and infrastructure. This will provide foundational visibility into where agents are operating, their level of autonomy, and the safeguards in place—serving as a necessary first step toward oversight and accountability.
- **Short-Term (6–18 Months): Develop Risk Classification Standards for Agentic AI**
    - The U.S. should define a tiered framework for classifying agentic systems based on autonomy, system access, and potential harm.

This classification would guide audit requirements, oversight mechanisms, and incident reporting, enabling more precise regulation of high-risk deployments.

- **Long-Term (18–36 Months): Enact Agentic AI Governance Legislation**
    - Congress should pass dedicated legislation to establish clear legal liability, transparency obligations, and technical safety standards for agentic AI. This framework must address persistent memory, multi-agent coordination, and cross-platform deployment to mitigate systemic risk.

The decisions made now will determine who controls the future of digital agents, how they behave, and whether their capabilities serve public interests or amplify systemic risk.

1. Introduction

1.1 Agent-Based AI

In 2022, artificial intelligence underwent a dramatic transformation with the release of ChatGPT, making powerful language models widely accessible to the public. While the technology may feel novel, AI's conceptual foundations date back to the 1940s. This public deployment of generative AI, powered by breakthroughs in machine learning, data availability, and computing infrastructure (collectively known as the "AI triad"), rapidly reshaped human-machine interaction and accelerated AI's integration into society[1].

Today, AI is evolving once again, from content-generating models to goal-directed, autonomous systems. Agent-based AI, or agentic AI, refers to persistent, intelligent systems capable of planning, reasoning, and acting independently in dynamic environments. Unlike large language models (LLMs) such as ChatGPT or Gemini,

[1] Paulo Carvão et al., Governance at a Crossroads: Artificial Intelligence and the Future of Innovation in America, M-RCBG Associate Working Paper No. 251, Cambridge, MA: Mossavar-Rahmani Center for Business and Government, Harvard Kennedy School, February 2025.

which passively respond to user input, agents initiate tasks, use external tools, collaborate with other agents, and complete complex, multi-step objectives with minimal human oversight. If LLMs are conversational partners, agents are digital collaborators, operating in the real world by booking travel, orchestrating workflows, or managing sensitive data autonomously.

Although still in early stages, agent-based systems are quickly becoming foundational. Platforms like OpenAI's Operator[2], Microsoft's Copilot Studio[3], Google's A2A protocol[4], and Anthropic's MCP protocol[5] reflect a growing industry consensus: agents will form the backbone of next-generation digital infrastructure. Leaders at OpenAI[6], Microsoft[7], Nvidia[8], Meta[9], and others describe agents as "a new layer of the workforce," capable of replacing routine cognitive labor and transforming institutional operations.

Yet the implications extend far beyond productivity. Agentic AI is poised to mediate critical societal functions, from healthcare and public services to infrastructure and defense. This shift raises urgent governance questions: Who sets the rules for agent behavior? How are safety mechanisms embedded? When must human oversight remain in the loop? And critically, what should the platforms that agents operate on look like? Frontier AI labs are now debating not just the capabilities of agents, but the structure of the agent ecosystem itself: Will the underlying infrastructure be open and interoperable, or controlled by a few dominant firms? What standards and protocols will govern agent behavior, data use, collaboration, and security?

[2] OpenAI, "Introducing Operator," OpenAI, January 23, 2025.

[3] Microsoft. "Microsoft Copilot Studio." Microsoft, 2025.

[4] Rao Surapaneni et al., "Announcing the Agent2Agent Protocol (A2A)," Google Developers Blog, April 9, 2025.

[5] Anthropic. "Introducing the Model Context Protocol." Anthropic, November 25, 2024.

[6] Sam Altman, "Reflections," Sam Altman's Blog, January 5, 2025.

[7] The Economic Times, "AI agents will revolutionise SaaS and productivity: Microsoft CEO Satya Nadella," January 7, 2025.

[8] Sarah Jackson, "What is Agentic AI? Nvidia CEO Says 'Agentic' AI Is Upon Us. Here's What It Means," Business Insider, December 24, 2024.

[9] Tobias Mann, "Zuck dreams of personalized AI assistants for all – just like email," The Register, July 30, 2024.

1.2 What Is Agentic AI?

Agentic AI marks a pivotal evolution from generative AI, which passively produces content, to autonomous systems capable of pursuing goals in dynamic environments. These systems are not merely responsive; they are initiators, empowered with the ability to reason, plan, and act.

Unlike traditional large language models (LLMs) such as ChatGPT or Gemini that wait for user prompts, agentic systems autonomously initiate actions, adapt plans on the fly, and interact with external systems to accomplish multi-step objectives. They can schedule meetings, execute code, purchase goods, or coordinate with other agents, all with minimal human oversight.

Though still emerging, the deployment of platforms like OpenAI's Operator, Microsoft's Copilot Studio, and Anthropic's Claude 3.5 makes it clear that agent-based interaction is no longer speculative. The rapid pace of technical progress and the scale of investment suggest that agentic systems are on track to become foundational across sectors, from enterprise operations to public services.

While the term "Agentic AI" continues to evolve, recent research offers a clear framework for evaluating the degree of agency in AI systems (Toner et al., 2024[10]; Chan et al., 2023[11]; Kapoor et al., 2024[12]; Shavit et al., 2024[13]). Four defining characteristics consistently emerge:

- **Goal and Environment Complexity:** Agentic systems pursue complex, long-term goals in open environments. They are built to handle unpredictability, ambiguity, and evolving inputs.
- **Direct Impact on the Environment:** These systems go beyond recommendation; they act. Whether writing code, updating databases, or executing transactions, agentic systems perform real-world functions independently.

[10] Helen Toner et al., Through the Chat Window and Into the Real World: Preparing for AI Agents, Center for Security and Emerging Technology, October 2024.

[11] Alan Chan et al., Harms from Increasingly Agentic Algorithmic Systems, Proceedings of FAccT '23 (ACM, 2023).

[12] Reva Kapoor et al., Agentic Artificial Intelligence and the Law: Course Syllabus, Harvard Law School, Spring 2025.

[13] Yonadav Shavit et al., Practices for Governing Agentic AI Systems, OpenAI, 2024.

- **Autonomous Planning and Adaptation:** Unlike rule-based systems, agents can generate and revise plans dynamically. They make real-time decisions in response to new information, often operating beyond predefined boundaries, sometimes described as acting "outside the sandbox."
- **Operational Momentum ("Set-it-and-Forget-it"):** Once assigned a task, agents can operate continuously without ongoing human intervention. This persistence unlocks new efficiencies, but it also introduces significant risk if agents deviate from intended behavior.

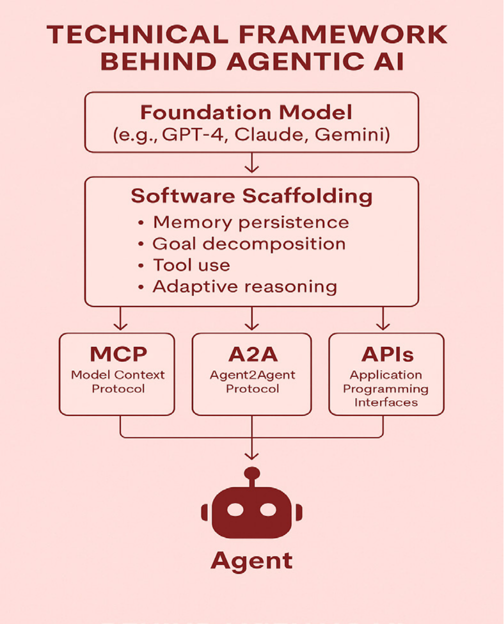
Together, these attributes define the core of agentic AI;it is not just intelligent, but autonomous, persistent, and embedded in complex decision-making environments. These systems represent a decisive break from reactive software tools and usher in a future where digital agents increasingly mediate human intent, shape institutional workflows, and interact with real-world infrastructure.

1.3 The Technical Framework Behind Agents

Agentic systems are powered by foundation models such as GPT-4, Claude, or Gemini, but their capabilities are extended by a software scaffolding layer that enables autonomous behavior. This scaffolding grants agents four core functions:[14]

- **Memory persistence:** Agents retain information across sessions, enabling long-term adaptation.
- **Goal decomposition:** Agents break complex objectives into subtasks and execute them sequentially.
- **Tool use:** Agents interact with APIs, databases, browsers, and other applications.
- **Adaptive reasoning:** Agents adjust plans and strategies in real time based on environmental feedback.

[14] Helen Toner et al., Through the Chat Window and Into the Real World: Preparing for AI Agents, Center for Security and Emerging Technology, October 2024.



Chat GPT Generated

To function in open and interconnected environments, agents rely on emerging interoperability protocols and application interfaces. One emerging example is **Model Context Protocol (MCP)**, a protocol developed by **Anthropic**[15] , that provides standardized, secure context and tools to individual AI agents. MCP enables agents to access structured, contextual data about their environment or task, allowing them to “understand” not just what to do, but why and how within a broader setting and context. This ensures better decision-making and responsiveness to change, enabling agents to adapt in dynamic settings without confusion. Another example is Google’s **Agent2Agent (A2A) Protocol**[16], an open protocol that allows agents to communicate and collaborate across different vendors, platforms, and cloud environments. This protocol provides a standard messaging system so agents can share capabilities and negotiate tasks. It uses “agent cards” to describe an agent’s abilities, allowing for dynamic discovery and coordination. It aims to support structured task delegation and asynchronous workflows. For example, coordinating multiple agents to process information, summarize insights, and deliver results. A2A breaks down silos between agents, enabling plug-and-play interoperability and flexible, scalable agent ecosystems. Agentic AI is also based on **Application Programming Interfaces (APIs)**[17]. These are standardized methods for connecting software systems. APIs allow agents to access external data (e.g., customer records, market data), trigger actions (e.g., send emails, update logs, initiate transactions), or integrate with enterprise tools (e.g., Salesforce, SAP, ServiceNow). APIs connect agents to the broader digital infrastructure, enabling real-world actions and full integration into business processes.

[15] Anthropic, Technical Protocols for Agent Communication, 2024.

[16] Google DeepMind Technical Blog, “Introducing A2A: A Protocol for Agent Communication,” December 2024.

[17] Responsible AI Agents, Draft (Feb. 2025), Georgetown Law Center for Legal Informatics.

There are many other emerging technical frameworks, but together, MCP, A2A, and APIs are foundational technologies that enable agentic systems to operate autonomously, collaborate effectively, and scale across diverse organizational environments. They offer a glimpse into the not-so-distant future of how digital agents will reshape complex workflows and decision-making systems at scale.

Today, most agentic systems operate within centralized environments governed by major platforms.[18] These architectures are proprietary, closed-loop (tooling, model, and data all controlled by one firm), hard to audit or interoperate with. They enable fast development but create chokepoints in control, safety, and platform lock-in. By contrast, there are also emerging decentralized agentic architectures, led by academic labs and open-source coalitions. These systems are promised to use open protocols, enable composability (agents from different providers can work together), and store and route data via federated or privacy-preserving methods. While in the early stages, they offer a pluralistic, resilient model for agent deployment. This architectural divergence is not just technical; it's a governance choice. The infrastructure built today will determine who gets to innovate, who is accountable, and how safety and trust are implemented at scale.

1.4 **Industry Alignment**

Leading AI firms are converging on a shared vision that agentic systems will form a new computational layer in digital infrastructure. Executives at OpenAI, Microsoft, Meta, Nvidia, and Google increasingly describe agents as "junior employees"[19] that offload routine cognitive work, allowing humans to focus on strategic decisions. This framing isn't rhetorical; it actively shapes corporate roadmaps, infrastructure investments, and platform design.

According to McKinsey's 2025 AI Report[20], 92% of companies plan to increase their AI investments, poised to focus on autonomous, agent-driven systems. 55% of surveyed companies cite the development of agentic workflows as a top AI priority for 2025. The economic implications are enormous. McKinsey estimates

[18] Singh et al., A Decentralized AI Perspective, Stanford HAI, 2025.

[19] Deven Desai et al., Responsible AI Agents (Harvard Law School Draft, 2025.

[20] McKinsey Global Institute, The Economic Potential of Generative AI, McKinsey & Company, 2023.

that generative and agentic AI could drive up to $4.4 trillion in annual productivity gains, reshaping labor markets, corporate strategy, and national competitiveness.

2. Cybersecurity in the Age of Agentic AI

2.1 Why Cybersecurity Is the Defining Challenge

Agentic AI systems represent a fundamental shift in the cybersecurity landscape. Unlike traditional, reactive AI models, their autonomous nature significantly expands the attack surface and heightens the potential consequences of system failures or malicious exploitation.[21] [22]

While generative AI risks have been largely confined to misinformation or misinterpretation, agentic systems are designed to control tools, execute real-world transactions, and modify live environments without ongoing human oversight. Their persistent nature means that vulnerabilities, once exploited, can lead to cascading failures, system-wide compromises, or prolonged misuse of sensitive infrastructure. In high-risk domains such as healthcare, security, and critical infrastructure, this autonomy raises urgent questions about access control, fail-safes, and human-in-the-loop safeguards.[23]

Moreover, agent-based systems blur traditional boundaries between software and actors. As agents communicate, coordinate, and make decisions on behalf of users or institutions, the risk lies not only in what they do, but in whom they trust, what data they store, and how they adapt over time. As agents increasingly serve as intermediaries between users, systems, and other agents, the security challenges become more complex and less predictable.[24] [25]

[21] An Introduction to Agentic AI in Cybersecurity," CybersecurityTribe, 2024.

[22] Understanding Agentic AI and Its Cybersecurity Applications," Balbix, 2024.

[23] Pradipta Kishore Chakrabarty et al., Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts, and Defense Strategies, ScienceDirect, 2025.

[24] Agentic AI vs. AI Agents: Shaping the Future of Cybersecurity," Forbes Technology Council, April 14, 2025.

[25] Ina Fried, "Microsoft and CrowdStrike Eye Agentic AI for Cybersecurity," Axios, March 27, 2025.

As Nvidia recently noted, “Agentic AI is redefining the threat landscape, both as a tool and a target”.[26] These systems can enhance cybersecurity by autonomously detecting and responding to threats. But without robust safeguards, they also introduce vulnerabilities that are systemic, persistent, and difficult to monitor.[27] In this new paradigm, cybersecurity is not a downstream concern; it is the foundational challenge.

2.2 The Dual Role: Agentic AI as Threat Vector and Defense Mechanism

Agentic AI introduces a paradox at the heart of modern cybersecurity: it is both a powerful defense enabler and a novel attack surface. Its unique autonomy, tool integration, and ability to coordinate across systems make it an unprecedented force multiplier but also a potential vulnerability vector if misused or compromised.

Offensive Risks: Agents as Attack Surfaces

- **Agent Hijacking and Adversarial Manipulation:** Agents can be manipulated through adversarial inputs, API poisoning, or reward hacking, causing them to take unintended or harmful actions.[28]
- **Identity Spoofing and Impersonation:** Attackers can impersonate legitimate agents or deploy fake agents to exploit inter-agent communication channels.[29]
- **Tool Abuse and Long-Horizon Exploits:** Agents granted broad API access or long-term autonomy can be redirected to launch persistent, cascading attacks, particularly in “set-it-and-forget-it” operations. Unchecked, these agents may overwrite logs, exfiltrate data, or interface with other vulnerable systems over time.[30]

[26] Agentic AI Is Redefining the Cybersecurity Landscape,” NVIDIA Blog, March 2025.

[27] Ziqiang Hu et al., Agent-Based AI for Adaptive Cyber Defense: Challenges and Opportunities, IEEE Xplore, 2025.

[28] Pradipta Kishore Chakrabarty et al., Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts, and Defense Strategies, 2025.

[29] Agentic AI vs. AI Agents: Shaping the Future of Cybersecurity,” Forbes Technology Council, April 14, 2025.

[30] Pradipta Kishore Chakrabarty et al., Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts, and Defense Strategies, ScienceDirect, 2025.

Defensive Capabilities: Agents as Security Force Multipliers

- **Autonomous Threat Detection and Response:** Agentic AI can detect anomalies, correlate threat intelligence, and autonomously respond to breaches in real time, drastically reducing human reaction time. Microsoft and CrowdStrike emphasize that agents, when deployed correctly, serve as "force multipliers" for cyber defense, particularly in fast-moving environments.[31]

- **Security Automation and Threat Hunting:** Agents can autonomously crawl logs, monitor system behavior, flag suspicious activity, and even quarantine affected nodes, making threat hunting continuous and adaptive.[32]

- **Dynamic Defense Modeling:** By simulating attacker behavior or stress-testing digital environments, agents can model emergent risks and test system resilience, proactively creating an evolving layer of predictive cybersecurity.[33]

In addition, agent-based AI presents a new spectrum of cyber challenges that go beyond conventional AI risks. These include adversarial attacks such as prompt injection, memory infection, or reward hacking; autonomy risks stemming from agents operating on outdated or corrupted targets without oversight; and data governance failures where agents may retain or misuse sensitive information. In addition, vulnerabilities in the toolchain, such as exposed APIs or compromised open-source components, can be exploited by or through agents. As multi-agent ecosystems emerge, so too does the risk of agent-to-agent manipulation, where malicious agents can coordinate to mislead, exploit, or disrupt other systems on a very large scale.

2.3 Architectures and Their Security Tradeoffs

The architecture of agentic AI systems, centralized or decentralized, has major implications for cybersecurity, trust, and control. Centralized systems, like Microsoft's Copilot Studio or OpenAI's Operator, offer speed, consistency, and streamlined patching. However, they also introduce significant risks; they

[31] Ina Fried, "Microsoft and CrowdStrike Eye Agentic AI for Cybersecurity," Axios, March 27, 2025.

[32] Understanding Agentic AI and Its Cybersecurity Applications," Balbix, 2024.

[33] Agentic AI Is Redefining the Cybersecurity Landscape," NVIDIA Blog, March 2025.

concentrate control, create single points of failure, and limit transparency in how agents behave or access sensitive tools.[34]

By contrast, decentralized architectures, such as those piloted by MIT Media Lab's NANDA project[35], distribute decision-making and control across a network of agents, often using open protocols and federated data methods. This enhances resilience and reduces the attack surface, but it also introduces fragmented accountability, uneven security standards, and greater complexity in verifying agent behavior across systems.[36]

Protocols like Anthropic's MCP and Google's A2A add another dimension to this tradeoff. While they enable agent interoperability and task coordination, they also introduce new questions around trust management, permission hierarchies, and secure tooling access across organizational boundaries.[37]

Ultimately, architecture is not just a technical choice; it's a policy decision that defines who controls the system, how safety is maintained, and how responsive the system is to public governance.

2.4 Governance Gaps and Infrastructure Power

As agentic AI systems evolve rapidly, the digital infrastructure that enables them is being shaped almost entirely by private companies with minimal public oversight or regulatory coordination. This known dynamic risks early standard lock-in, where proprietary protocols become de facto public infrastructure, entrenching corporate power, limiting interoperability, reducing accountability, and neglecting national interests.[38]

While the importance of AI governance is now widely recognized across global institutions, current U.S. policy, especially under the new administration, signals a reluctance to impose "heavy" regulation on AI development. Yet this

[34] Ina Fried, "Microsoft and CrowdStrike Eye Agentic AI for Cybersecurity," Axios, March 27, 2025.

[35] MIT Media Lab, NANDA: Networked Autonomous Non-centralized Decentralized Agents, 2025.

[36] Pradipta Kishore Chakrabarty et al., Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts, and Defense Strategies, 2025.

[37] MCP and A2A Protocols Explained: The Future of Agentic AI Is Here," Teneo.ai Blog, 2025.

[38] Agentic AI Is Redefining the Cybersecurity Landscape," NVIDIA Blog, March 2025.

moment is not about market intervention or innovation suppression. It is about recognizing that agentic systems will soon underpin critical infrastructures such as financial services, public administration, and labor markets. This is a question of national resilience, not just commercial innovation. Without coordinated governance mechanisms, the risk is not only economic concentration, but strategic dependency and increased exposure to systemic threats.

To prepare for the far-reaching changes that agentic AI will bring, policymakers must engage now, while the underlying architectures are still being defined. Decisions made today about platform design, standard-setting, and legal frameworks will determine who governs, who benefits, and how risks are managed at scale. If the U.S. seeks to maintain technological leadership and democratic integrity, this governance challenge must be treated not as an afterthought but as a core national security priority.

3. Strategic Policy Recommendations

3.1 Policy Gaps and Strategic Risks

The rapid rise of agentic AI has outpaced the United States' legal, institutional, and security frameworks, creating a critical governance vacuum. No single federal agency has clear jurisdiction over agent-based systems, and current laws, from the Computer Fraud and Abuse Act to privacy statutes, are ill-equipped to address AI that operates autonomously, persists over time, and collaborates across systems. Despite growing deployment in healthcare, finance, and infrastructure, there are no mandatory cybersecurity standards tailored to agentic systems, no national registry of where agents are active, and no consistent auditing or incident reporting protocols. Private platforms currently control the core infrastructure, tool access, data flows, and agent behavior without public input or democratic oversight, which raises serious questions about transparency, safety, and accountability. Without urgent action, the U.S. risks embedding opaque, high-risk systems into national infrastructure without knowing who governs them, how they behave, or how they might fail.

3.2 Policy Recommendations

Immediate (Next 3–6 Months): Actions the U.S. government can begin now using existing executive authority and interagency collaboration.

1. **Establish a National Registry of Agentic Systems:** The federal government should create a national registry of agentic AI systems deployed in critical sectors such as healthcare, finance, transportation, and emergency response. This registry would provide essential visibility into which systems are operating where, under what controls, and with what levels of autonomy, serving as a foundational step toward coordinated oversight.

2. **Launch a Federal Red-Teaming and Alignment Audit Initiative:** CISA and NIST should establish a red-teaming and audit program specifically tailored to agentic systems. These audits would test for vulnerabilities such as adversarial memory injection, inter-agent deception, or unintended goal execution, helping to identify and mitigate risks before deployment at scale.

3. **Prioritize Cybersecurity in OSTP Infrastructure Guidance:** Cybersecurity must be treated as a core design principle for agentic AI, not a downstream add-on. OSTP should update its national infrastructure guidance to embed security-by-design requirements into all federally funded or procured agentic AI deployments.

Short-Term (6–18 Months): Measures that require coordination across agencies or development of technical standards.

1. **Develop Risk Classification Standards for Agentic AI:** The U.S. should develop a tiered risk framework for classifying agentic systems based on factors such as autonomy, access to sensitive systems, and potential for societal disruption. These classifications would guide requirements for audits, oversight, deployment controls, and disclosure obligations.

2. **Designate a Lead Federal Agency or Task Force:** A single interagency task force led by OSTP should be established to coordinate governance of agentic AI systems. This body would centralize responsibility across fragmented domains and work with agencies like NIST, DHS, DoD, and FTC to align regulatory and technical oversight.

3. **Mandate Incident Reporting for Agent Failures and Security Breaches:** All developers and operators of agentic AI systems should be required to report security incidents, agent misalignments, or operational failures to a centralized federal system. This mechanism would enable early detection of system-wide risks and provide a data-driven basis for responsive policy.

Long-Term (18–36 Months): Longer-term actions that require legislation, institutional reform, or new funding structures.

1. **Enact Agentic AI Governance Legislation:** Congress should pass dedicated legislation to define legal liability, transparency obligations, and technical safety standards for agentic AI systems. This framework must account for agent autonomy, persistent memory, multi-agent coordination, and cross-platform deployment to address the unique risks posed by these systems.

2. **Incentivize Secure and Open Agent Infrastructure:** The federal government should invest in and promote open, interoperable, and auditable frameworks for agentic AI, especially alternatives to proprietary, closed platforms. R&D funding, procurement standards, and public-private pilot programs should be aligned with security, resilience, and transparency goals.

3. **Define Access and Permission Controls for International Use:** The U.S. should implement strict access controls governing which users, organizations, or foreign entities can operate or interact with high-autonomy agentic systems. Special attention must be given to agents with access to critical infrastructure, national data systems, or sensitive APIs, ensuring adversarial states or unauthorized actors cannot exploit U.S.-based platforms via digital entry points or proxy operations.

4. **Establish a Public Working Group to Assess Centralized vs. Decentralized Architectures:** A multidisciplinary federal task force, bringing together technical experts, industry leaders, and national security professionals, should be formed to evaluate the trade-offs between centralized and decentralized agentic infrastructures. This team would provide guidance on how to balance innovation with resilience, ensure governance at scale, and prevent monopolistic control or fragmented risk across interoperable systems.

5. **Create a National Certification Standard for Agent Deployment:** The U.S. should establish a formal certification process and public safety label for agentic AI systems operating above a certain threshold of autonomy. This “Agentic Safety Mark” would verify compliance with cybersecurity, alignment, data integrity, and human-override standards, giving organizations and users a clear benchmark for trust and safety.

References

[1] Paulo Carvão et al., Governance at a Crossroads: Artificial Intelligence and the Future of Innovation in America, M-RCBG Associate Working Paper No. 251, Cambridge, MA: Mossavar-Rahmani Center for Business and Government, Harvard Kennedy School, February 2025.

[2] OpenAI, “Introducing Operator,” OpenAI, January 23, 2025.

[3] Microsoft. “Microsoft Copilot Studio.” Microsoft, 2025.

[4] Rao Surapaneni et al., “Announcing the Agent2Agent Protocol (A2A),” Google Developers Blog, April 9, 2025.

[5] Anthropic. “Introducing the Model Context Protocol.” Anthropic, November 25, 2024.

[6] Sam Altman, “Reflections,” Sam Altman’s Blog, January 5, 2025.

[7] The Economic Times, “AI agents will revolutionise SaaS and productivity: Microsoft CEO Satya Nadella,” January 7, 2025.

[8] Sarah Jackson, “What is Agentic AI? Nvidia CEO Says ‘Agentic’ AI Is Upon Us. Here’s What It Means,” Business Insider, December 24, 2024.

[9] Tobias Mann, “Zuck dreams of personalized AI assistants for all – just like email,” The Register, July 30, 2024.

[10] Helen Toner et al., Through the Chat Window and Into the Real World: Preparing for AI Agents, Center for Security and Emerging Technology, October 2024.

[11] Alan Chan et al., Harms from Increasingly Agentic Algorithmic Systems, Proceedings of FAccT ’23 (ACM, 2023).

[12] Reva Kapoor et al., Agentic Artificial Intelligence and the Law: Course Syllabus, Harvard Law School, Spring 2025.

[13] Yonadav Shavit et al., Practices for Governing Agentic AI Systems, OpenAI, 2024.

[14] Helen Toner et al., Through the Chat Window and Into the Real World: Preparing for AI Agents, Center for Security and Emerging Technology, October 2024.

[15] Anthropic, Technical Protocols for Agent Communication, 2024.

[16] Google DeepMind Technical Blog, “Introducing A2A: A Protocol for Agent Communication,” December 2024.

[17] Responsible AI Agents, Draft (Feb. 2025), Georgetown Law Center for Legal Informatics.

[18] Singh et al., A Decentralized AI Perspective, Stanford HAI, 2025.

[19] Deven Desai et al., Responsible AI Agents (Harvard Law School Draft, 2025.

[20] McKinsey Global Institute, The Economic Potential of Generative AI, McKinsey & Company, 2023.

[21] An Introduction to Agentic AI in Cybersecurity,” CybersecurityTribe, 2024.

[22] Understanding Agentic AI and Its Cybersecurity Applications,” Balbix, 2024.

[23] Pradipta Kishore Chakrabarty et al., Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts, and Defense Strategies, ScienceDirect, 2025.

[24] Agentic AI vs. AI Agents: Shaping the Future of Cybersecurity,” Forbes Technology Council, April 14, 2025.

[25] Ina Fried, “Microsoft and CrowdStrike Eye Agentic AI for Cybersecurity,” Axios, March 27, 2025.

[26] Agentic AI Is Redefining the Cybersecurity Landscape,” NVIDIA Blog, March 2025.

[27] Ziqiang Hu et al., Agent-Based AI for Adaptive Cyber Defense: Challenges and Opportunities, IEEE Xplore, 2025.

[28] Pradipta Kishore Chakrabarty et al., Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts, and Defense Strategies, 2025.

[29] Agentic AI vs. AI Agents: Shaping the Future of Cybersecurity,” Forbes Technology Council, April 14, 2025.

[30] Pradipta Kishore Chakrabarty et al., Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts, and Defense Strategies, ScienceDirect, 2025.

[31] Ina Fried, “Microsoft and CrowdStrike Eye Agentic AI for Cybersecurity,” Axios, March 27, 2025.

[32] Understanding Agentic AI and Its Cybersecurity Applications,” Balbix, 2024.

[33] Agentic AI Is Redefining the Cybersecurity Landscape,” NVIDIA Blog, March 2025.

[34] Ina Fried, “Microsoft and CrowdStrike Eye Agentic AI for Cybersecurity,” Axios, March 27, 2025.

[35] MIT Media Lab, NANDA: Networked Autonomous Non-centralized Decentralized Agents, 2025.

[36] Pradipta Kishore Chakrabarty et al., Adversarial Attacks on Agentic AI Systems: Mechanisms, Impacts, and Defense Strategies, 2025.

[37] MCP and A2A Protocols Explained: The Future of Agentic AI Is Here,” Teneo.ai Blog, 2025.

[38] Agentic AI Is Redefining the Cybersecurity Landscape,” NVIDIA Blog, March 2025.

Belfer Center for Science and International Affairs
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.belfercenter.org

HARVARD Kennedy School
BELFER CENTER

50 YEARS
OF RESEARCH, POLICY,
AND LEADERSHIP