R Street
Free markets. Real solutions.

R Street Policy Study
No. 325
May 2025

The Rise of AI Agents: Anticipating Cybersecurity Opportunities, Risks, and the Next Frontier

By Haiman Wong and Tiffany Saade

To prepare for the next frontier of AI, this study explores the cybersecurity implications of AI agents and presents a three-pronged framework to guide secure design and responsible deployment.

Table of Contents

Executive Summary

The rise of AI agents represents a significant evolution in artificial intelligence (AI)—shifting from passive, prompt-based tools to increasingly autonomous systems capable of reasoning, memory, learning, and complex task execution. As industries adopt these agents, their impact on business operations, human-machine collaboration, and national security is accelerating.

In cybersecurity, AI agents are already proving to be valuable copilots to human analysts—enhancing threat detection, expediting incident response, and supporting overstretched cyber teams. Yet with greater power across the agentic infrastructure stack—spanning perception, reasoning, action, and memory—comes greater responsibility to ensure that agents are secure, explainable, and reliable.

To prepare for this next frontier of AI, this study explores the cybersecurity implications of AI agents and presents a three-pronged framework to guide secure design and responsible deployment: (1) identifying policy priorities, including voluntary, sector-specific guidelines for navigating human–agent collaboration; (2) anticipating emerging technological solutions to strengthen agentic oversight and cyber resilience; and (3) outlining best practices for

developers, organizations, and end-users to ensure that agents augment—rather than replace—human talent and decision-making.

Ultimately, the goal is not merely to keep pace with the advancement of agentic systems, but to shape their trajectory—harnessing their benefits, minimizing their risks, and safeguarding America's technological leadership, national security, and values in the process.

Introduction

While 2023 was dubbed the year of "generative artificial intelligence (Gen AI)" and 2024 was marked by a steady march toward "AI practicality," 2025 opened with high expectations that it would become the year of "AI agents."[1] At its core, an AI agent is an "autonomous intelligent system powered by artificial intelligence and designed to perform specific tasks independently without the need for human intervention."[2] However, as seen with previous AI advancements—whether Gen AI, open-source AI, or large language models (LLMs)—a single, universally adopted definition remains elusive.[3] Some experts describe AI agents as "applications that attempt to achieve a goal by observing the world and acting upon it using the tools that [they have] at [their] disposal."[4] Others characterize agents as "layers on top of the language models that observe and collect information, provide input to the model and together generate an action plan and communicate that to the user—or even act on their own, if permitted."[5] Though these definitions vary in their precise phrasing and perspective, each consistently emphasizes the agent's ability to pursue and complete goals autonomously using a suite of capabilities that includes learning, memory, planning, reasoning, decision-making, and adaptation.[6]

Notably, not all AI agents are created equal. Non-agentic and agentic AI systems differ in how they operate, particularly in autonomy and goal-setting. Non-agentic systems, such as earlier versions of ChatGPT or Alexa, respond to user prompts without retaining memory, setting goals, or initiating actions on their own.[7] In contrast, agentic AI systems pursue objectives over time, using contextual awareness, autonomous planning, and adaptive reasoning to carry out multi-step tasks with minimal human input or oversight.[8] For example, an agentic system might autonomously research a topic across multiple websites, generate a tailored report, and distribute it through email.[9] This added layer of sophistication gives AI agents greater potential to navigate across different domains and produce more complex,

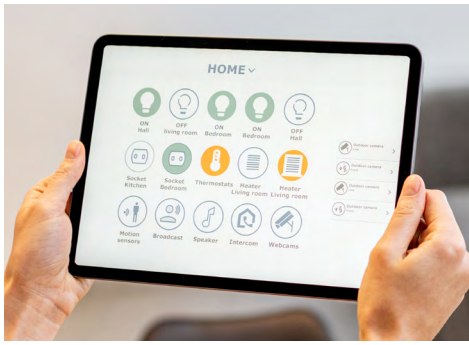
NON-AGENTIC AI SYSTEMS
such as earlier versions of ChatGPT or Alexa, respond to user prompts without retaining memory, setting goals, or initiating actions on their own.

AGENTIC AI SYSTEMS
pursue objectives over time, using contextual awareness, autonomous planning, and adaptive reasoning to carry out multi-step tasks with minimal human input or oversight.

1. Haiman Wong et al., "The Transformative Role of AI in Cybersecurity: Anticipating and Preparing for Future Applications and Benefits," R Street Institute, Jan. 24, 2024. https://www.rstreet.org/commentary/the-transformative-role-of-ai-in-cybersecurity-anticipating-and-preparing-for-future-applications-and-benefits; Eric Johnson, "2023 Was The Year Of AI Hype—2024 Is The Year Of AI Practicality," *Forbes*, April 2, 2024. https://www.forbes.com/councils/forbestechcouncil/2024/04/02/2023-was-the-year-of-ai-hype-2024-is-the-year-of-ai-practicality; Tae Kim, "Nvidia CEO Says 2025 Is the Year of AI Agents," Barron's, Jan. 7, 2025. https://www.barrons.com/articles/nvidia-stock-ceo-ai-agents-8c20ddfb.
2. Kinza Yasar, "What are AI agents?," TechTarget, December 2024. https://www.techtarget.com/searchenterpriseai/definition/AI-agents.
3. Maxwell Zeff and Kyle Wiggers, "No one knows what the hell an AI agent is," TechCrunch, March 14, 2025. https://techcrunch.com/2025/03/14/no-one-knows-what-the-hell-an-ai-agent-is.
4. Michael Nuñez, "Google maps the future of AI agents: Five lessons for businesses," VentureBeat, Jan. 6, 2025. https://venturebeat.com/ai/google-maps-the-future-of-ai-agents-five-lessons-for-businesses; Julia Wiesinger et al., "Agents," Google, September 2024. https://www.kaggle.com/whitepaper-agents.
5. Susanna Ray, "AI agents—what they are, and how they'll change the way we work," Microsoft, Nov. 19, 2024. https://news.microsoft.com/source/features/ai/ai-agents-what-they-are-and-how-theyll-change-the-way-we-work.
6. "What is an AI Agent?," GoogleCloud, last accessed March 29, 2025. https://cloud.google.com/discover/what-are-ai-agents.
7. Anna Gutowska, "What are AI agents?," IBM, July 3, 2024. https://www.ibm.com/think/topics/ai-agents.
8. Ibid.
9. Ibid.

real-world outcomes. These capabilities not only unlock new opportunities but also raise novel challenges for oversight, governance, and cybersecurity that differ in both scope and scale from earlier AI systems.[10]

AI agents have already begun transforming workflows across various sectors, especially in software engineering, where they are increasingly embedded into routine development tasks.[11] For example, Claude 3.7 and Cursor AI are automating software development tasks such as code generation, refactoring, and debugging.[12] In cybersecurity, Microsoft's Security Copilot can autonomously triage phishing alerts, dynamically update its detection capabilities based on analyst feedback, and flag security policy configuration issues.[13] Other similar and emerging cybersecurity-focused AI agents include Exabeam's Copilot, Cymulate AI Copilot, and Oleria Copilot, all of which streamline cyber incident investigations and simulations.[14] Beyond these programming- and cybersecurity-centric agents, general-purpose agents like OpenAI's Operator, Anthropic's Computer Use, and Google's Project Astra also stand out for their potential to coordinate tasks such as multi-step web navigation, form completion, and cross-application integration.[15] Looking ahead, many experts anticipate that agentic AI will continue advancing in three key areas: (1) enhancing reasoning and contextual understanding to solve problems more effectively, (2) delivering greater autonomy for complex task execution, and (3) augmenting the human workforce.[16]



Some experts have even suggested that—in the next five years—workers will face a reality in which AI is doing 80 percent of their day-to-day tasks.

Some experts have even suggested that—in the next five years—workers will face a reality in which AI is doing 80 percent of their day-to-day tasks.[17] As AI agents begin shouldering more cognitive and operational tasks, their influence over business operations, workforce dynamics, and digital infrastructure will become increasingly consequential. Accordingly, experts now view technological leadership in agentic AI as a matter of strategic national importance. Although many of today's leading agents continue to be developed by American AI and technology companies, the global competition to lead in agentic AI persists.[18]

10. Shomit Ghose, "The Next 'Next Big Thing': Agentic AI's Opportunities and Risks," UC Berkeley Sutardja Center for Entrepreneurship & Technology, Dec. 19, 2024. https://scet.berkeley.edu/the-next-next-big-thing-agentic-ais-opportunities-and-risks.
11. Janakiram MSV, "GitHub Copilot Agent And The Rise Of AI Coding Assistants," *Forbes*, Feb. 8, 2025. https://www.forbes.com/sites/janakirammsv/2025/02/08/github-copilot-agent-and-the-rise-of-ai-coding-assistants; Matt Marshall, "Anthropic's stealth enterprise coup: How Claude 3.7 is becoming the coding agent of choice," VentureBeat, March 11, 2025. https://venturebeat.com/ai/anthropics-stealth-enterprise-coup-how-claude-3-7-is-becoming-the-coding-agent-of-choice; Julie Bort, "AI coding assistant Cursor reportedly tells a 'vibe coder' to write his own damn code," TechCrunch, March 14, 2025. https://techcrunch.com/2025/03/14/ai-coding-assistant-cursor-reportedly-tells-a-vibe-coder-to-write-his-own-damn-code.
12. Ibid.
13. Vasu Jakkal, "Microsoft unveils Microsoft Security Copilot agents and new protections for AI," Microsoft, March 24, 2025. https://www.microsoft.com/en-us/security/blog/2025/03/24/microsoft-unveils-microsoft-security-copilot-agents-and-new-protections-for-ai; Jeffrey Schwartz, "Microsoft Gives Security Copilot Some Autonomy," DarkReading, March 24, 2025. https://www.darkreading.com/cybersecurity-operations/microsoft-gives-security-copilot-autonomy.
14. Louis Columbus, "From alerts to autonomy: How leading SOCs use AI copilots to fight signal overload and staffing shortfalls," VentureBeat, March 24, 2025. https://venturebeat.com/security/ai-copilots-cut-false-positives-and-burnout-in-overworked-socs
15. Will Douglas Heaven, "OpenAI launches Operator—an agent that can use a computer for you," MIT Technology Review, Jan. 23, 2025. https://www.technologyreview.com/2025/01/23/1110484/openai-launches-operator-an-agent-that-can-use-a-computer-for-you; Samuel Axon, "Anthropic publicly releases AI tool that can take over the user's mouse cursor," Ars Technica, Oct. 22, 2024. https://arstechnica.com/ai/2024/10/anthropic-publicly-releases-ai-tool-that-can-take-over-the-users-mouse-cursor; Will Douglas Heaven, "Google's new Project Astra could be generative AI's killer app," MIT Technology Review, Dec. 11, 2024. https://www.technologyreview.com/2024/12/11/1108493/googles-new-project-astra-could-be-generative-ais-killer-app.
16. Melissa Heikkilä and Will Douglas Heaven, "Anthropic's chief scientist on 4 ways agents will be even better in 2025," MIT Technology Review, Jan. 11, 2025. https://www.technologyreview.com/2025/01/11/1109909/anthropics-chief-scientist-on-5-ways-agents-will-be-even-better-in-2025; Kate Whiting, "The rise of 'AI agents': What they are and how to manage the risks," IBM, Dec. 16, 2024. https://www.weforum.org/stories/2024/12/ai-agents-risks-artificial-intelligence.
17. Kate Whiting, "What is an AI agent and what will they do? Experts explain," World Economic Forum, July 24, 2024. https://www.weforum.org/stories/2024/07/what-is-an-ai-agent-experts-explain.
18. Alexander Puutio, "The Agentic AI Race Is On, And The Blue Chips Are All In," *Forbes*, Nov. 15, 2024. https://www.forbes.com/sites/alexanderpuutio/2024/11/15/the-agentic-ai-race-is-on-and-the-blue-chips-are-all-in; Sean Oesch et al., "Agentic AI and the Cyber Arms Race," arXiv, Feb. 10, 2025. https://arxiv.org/html/2503.04760v1.

Outside of the United States, Manus, developed by Wuhan-based startup Butterfly Effect, garnered global attention in March 2025 when it claimed to be "the world's first general AI agent, using multiple AI models (such as Anthropic's Claude 3.5 Sonnet and fine-tuned versions of Alibaba's open-source Qwen) and various independently operating agents to act autonomously on a wide range of tasks."[19] Manus is designed to execute diverse, goal-oriented tasks independently, including language translation, online purchasing, research synthesis, and 3D game development from a single prompt.[20] The launch of Manus underscores the strategic significance of AI agents in global innovation ecosystems and mirrors trends already observed in the open-source AI movement and the broader competitive landscape around AI innovation and deployment.[21]

In March 2025, for example, the Pentagon contracted with ScaleAI to give AI what one reporter termed "its most prominent role in the Western defense sector to date."

As with earlier waves of AI development, establishing technological leadership in agentic AI may carry both economic and substantial geopolitical implications, especially if agents become embedded in critical workflows across sensitive sectors, such as finance, healthcare, and defense.[22] In March 2025, for example, the Pentagon contracted with ScaleAI to give AI what one reporter termed "its most prominent role in the Western defense sector to date."[23] Framed as the Department of Defense's "first foray" into deploying agentic systems across military workflows, the initiative aims to accelerate strategic assessments, simulate war-gaming scenarios, and modernize campaign development.[24] This deal not only underscores the rising stakes of securing America's technological leadership but also ushers in the age of "agentic warfare" at a moment when many are only beginning to grapple with the full scope of this technology's capabilities, limitations, and risks.[25]

The rise of AI agents presents a critical window of opportunity to take a closer look—not only at how AI agents are being developed—but also at how they can be secured and governed.[26] As agentic systems begin streamlining business operations, problem-solving, and human–machine collaboration, the implications extend far beyond technological innovation.[27] In light of these rapidly evolving shifts, policymakers must craft governance strategies that are balanced, forward-looking, and flexible—strategies that support agentic innovation and use while proactively mitigating risks and ensuring long-term national security resilience.

19. Caiwei Chen, "Everyone in AI is talking about Manus. We put it to the test," MIT Technology Review, March 11, 2025. https://www.technologyreview.com/2025/03/11/1113133/manus-ai-review.
20. Rhiannon Williams, "The Download: testing new AI agent Manus, and Waabi's virtual robotruck ambitions," MIT Technology Review, March 12, 2025. https://www.technologyreview.com/2025/03/12/1113172/the-download-testing-new-ai-agent-manus-and-waabis-virtual-robotruck-ambitions; Coco Feng, "Manus draws upbeat reviews of nascent system," MSN, March 3, 2025. https://www.msn.com/en-sg/news/other/less-structure-more-intelligence-ai-agent-manus-draws-upbeat-reviews-of-nascent-system/ar-AA1AJyOE.
21. Ibid.
22. Kieran Garvey, "How Agentic AI will transform financial services with autonomy, efficiency, and inclusion," World Economic Forum, Dec. 2, 2024. https://www.weforum.org/stories/2024/12/agentic-ai-financial-services-autonomy-efficiency-and-inclusion; Sara Heath, "Epic's take on agentic AI designed to boost patient experience," TechTarget, March 5, 2025. https://www.techtarget.com/patientengagement/feature/Epics-take-on-agentic-AI-designed-to-boost-patient-experience; Brandon Vigliarolo, "It begins: Pentagon to give AI agents a role in decision making, ops planning," The Register, March 5, 2025. https://www.theregister.com/2025/03/05/dod_taps_scale_to_bring.
23. Ibid.
24. Ibid.
25. Julia Hornstein, "AI agents are coming to the military. VCs love it, but researchers are a bit wary," Business Insider, March 8, 2025. https://www.businessinsider.com/ai-agents-coming-military-new-scaleai-contract-2025-3.
26. 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. https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents.
27. Ibid.

This study examines the benefits, risks, cybersecurity considerations, and policy needs likely to define the emerging frontier of AI advancement. It begins by outlining the architecture of agentic systems and explaining how they differ from earlier generations of AI tools. It then explores how AI agents are already being deployed in cybersecurity use cases and identifies the new categories of risk they introduce across four distinct infrastructure layers: perception, reasoning, action, and memory. Finally, the paper presents a framework for secure and responsible deployment, emphasizing the roles of policy, technical safeguards, and organizational practices in strengthening long-term resilience and demonstrating that—if guided with care and foresight—agentic systems could mark not just the next phase of AI, but a turning point in how digital security is built and sustained.

Background

Although AI agents dominated news headlines in late 2024 and early 2025, their conceptual foundations trace back to the 1970s and 1980s, when research explored how capable systems were of sensing and acting intelligently within an environment.[28] These early systems, often referred to as "intelligent agents," powered linguistic analysis, biomedical applications, and robotics, relying on rule-based logic and limited autonomy due to constraints in hardware, computing power, and algorithmic sophistication.[29] At the time, these agents were described as "a new type of AI system capable of adapting, learning from data, and making complex decisions in changing environments."[30]

The renewed surge of interest in agents today reflects a convergence of technological advancements: scalable cloud infrastructure, advanced foundation models such as GPT-4 and Claude 3.5, and modular architectures that support planning, reasoning, and action with minimal human oversight.[31] Tools like AutoGPT, an "experimental, open-source Python application that uses GPT-4 to act autonomously," also helped popularize this shift from reactive, prompt-based tools to proactive, goal-driven systems capable of coordinating complex tasks.[32] As a result, AI agents are now positioned to be practical tools with significant operational and economic utility—from automating software development to automating customer service and even augmenting real-time cybersecurity defense.[33]

AI agents are now positioned to be practical tools with significant operational and economic utility—from automating software development to automating customer service and even augmenting real-time cybersecurity defense.

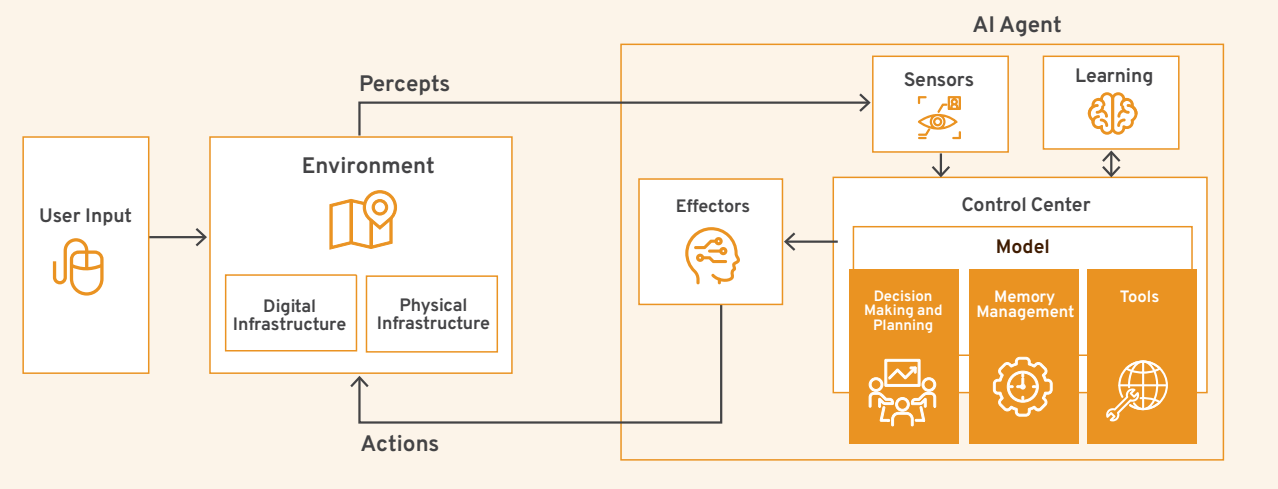
Architecturally, AI agents typically operate as a layer above LLMs and include four foundational components: perception, reasoning, action, and memory.[34] The perception module is responsible for ingesting data from external sources, such

28. Matvii Diadkov, "AI Agents: From Inception to Today," *Forbes*, March 18, 2025. https://www.forbes.com/councils/forbesbusinesscouncil/2025/03/18/ai-agents-from-inception-to-today.
29. Joseph Reagle, "The Etymology of 'Agent' and 'Proxy' in Computer Networking Discourse," Harvard University, Sept. 18, 1998. https://cyber.harvard.edu/archived_content/people/reagle/etymology-agency-proxy-19981217.html.
30. Diadkov. https://www.forbes.com/councils/forbesbusinesscouncil/2025/03/18/ai-agents-from-inception-to-today.
31. Timothy R. McIntosh et al., "From Google Gemini to OpenAI Q* (Q-Star): A Survey on Reshaping the Generative Artificial Intelligence (AI) Research Landscape," MDPI, Jan. 30, 2025. https://www.mdpi.com/2227-7080/13/2/51.
32. Sabrina Ortiz, "What is Auto-GPT? Everything to know about the next powerful AI tool," ZDNet, April 14, 2023. https://www.zdnet.com/article/what-is-auto-gpt-everything-to-know-about-the-next-powerful-ai-tool.
33. Diadkov. https://www.forbes.com/councils/forbesbusinesscouncil/2025/03/18/ai-agents-from-inception-to-today.
34. Whiting, "The rise of 'AI agents': What they are and how to manage the risks." https://www.weforum.org/stories/2024/12/ai-agents-risks-artificial-intelligence.

as user inputs or application programming interfaces (APIs).[35] After the data is gathered, the reasoning module leverages the LLM's capabilities to plan or infer the best course of action.[36] The action module can then execute tasks through tools, APIs, or integrations with third-party systems.[37] Finally, the memory module stores contextual information, often using vector databases or session-based memory managers.[38] This modular stack enables agents to operate across real-world applications and adapt while completing tasks in ways that static prompt chains or retrieval-augmented generation (RAG) pipelines cannot.[39]

Figure 1 illustrates how advanced AI agents may sense and interact with their environment, process information, and coordinate actions.

Figure 1: Key Components of Advanced AI Agents



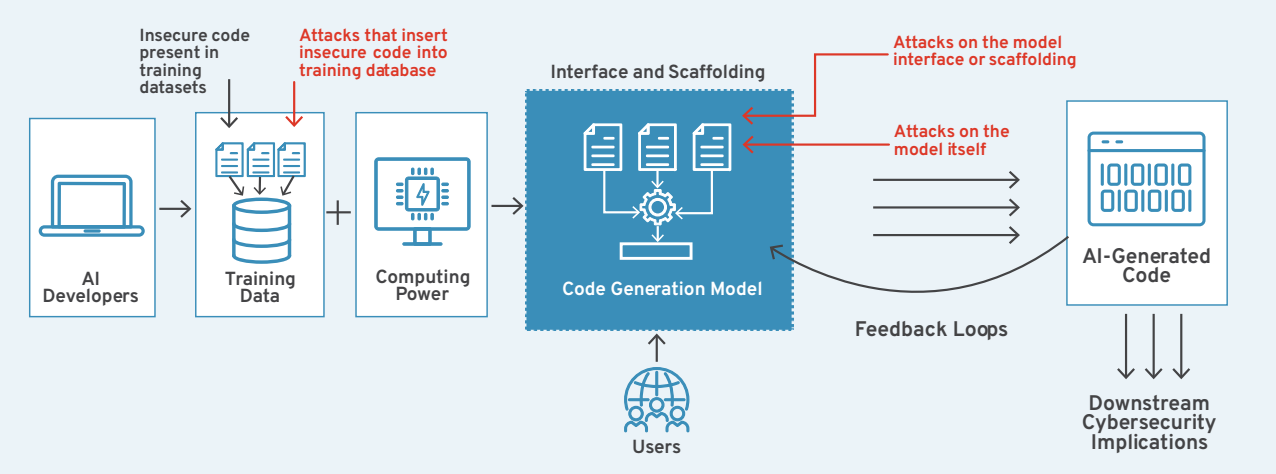
Source: Kate Whiting, "The rise of 'AI agents': What they are and how to manage the risks," World Economic Forum, Dec. 16. 2024. https://www.weforum.org/stories/2024/12/ai-agents-risks-artificial-intelligence.

Behind this architecture lies a supporting infrastructure stack: model APIs for LLM access, memory stores for quick retrieval, session managers for coordinating task state, external tool integrations for operational output, and even open-source frameworks and libraries that enable modular development.[40] Multi-agent systems add another layer of sophistication, allowing agents to collaborate or delegate tasks to other agents within a shared environment.[41] While this growing interconnectedness can enhance agentic capabilities, it can also introduce new challenges around explainability, privacy, system security, and reliability.[42]

35. Yasar. https://www.techtarget.com/searchenterpriseai/definition/AI-agents.
36. Rine Diane Caballar and Cole Stryker, "What is agentic reasoning?," IBM, March 20, 2025. https://www.ibm.com/think/topics/agentic-reasoning.
37. Yasar. https://www.techtarget.com/searchenterpriseai/definition/AI-agents.
38. Cole Stryker, "What is AI agent memory?," IBM, March 18, 2025. https://www.ibm.com/think/topics/ai-agent-memory.
39. Ibid.
40. Yifeng He et al., "Security of AI Agents," arXiv, June 20, 2024. https://arxiv.org/html/2406.08689v2; Cole Stryker, "What are the components of AI agents?," IBM, March 10, 2025. https://www.ibm.com/think/topics/components-of-ai-agents.
41. Talha Zeeshan et al., "Large Language Model Based Multi-Agent System Augmented Complex Event Processing Pipeline for Internet of Multimedia Things," arXiv, Jan. 3, 2025. https://arxiv.org/html/2501.00906v2.
42. Cole Stryker, "Agentic AI: 4 reasons why it's the next big thing in AI research," IBM, Oct. 11, 2024. https://www.ibm.com/think/insights/agentic-ai; Adrian Bridgwater, "Okay AI, Solo Kagent Is An Agentic AI Framework For Kubernetes," *Forbes*, March 17, 2025. https://www.forbes.com/sites/adrianbridgwater/2025/03/17/okay-ai-solo-kagent-is-an-agentic-ai-framework-for-kubernetes; Elizabeth Wallace, "How Agentic AI is Changing Decision-Making," CD Insights, March 8, 2025. https://www.clouddatainsights.com/how-agentic-ai-is-changing-decision-making.

As visualized in Figure 2, vulnerabilities can emerge across multiple layers of the agent lifecycle—from insecure training data and compromised model interfaces to attacks on the agents themselves. These risks, if left unaddressed, can embed insecure behaviors into models or outputs and trigger downstream cybersecurity consequences that expose users, systems, and software supply chains.

Figure 2: Cybersecurity Risks of AI-Generated Code

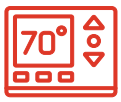


Source: Jessica Ji et al., "Cybersecurity Risks of AI-Generated Code," Center for Security and Emerging Technology, Nov. 2024. https://cset.georgetown.edu/publication/cybersecurity-risks-of-ai-generated-code. qtd. in Christopher Jablonski, "AI software supply chain risks prompt new corporate diligence," Zscaler, Jan. 13, 2025. https://www.zscaler.com/cxorevolutionaries/insights/ai-software-supply-chain-risks-prompt-new-corporate-diligence.

To further contextualize how agents work, it is important to be familiar with the seven main types of agents that have emerged in AI research and development.[43] Each category represents a varying level of autonomy, complexity, and adaptability:

1. Simple Reflex Agents. These agents represent the most basic form of agent, as they operate based on predefined condition-action rules only.[44] These types of agents are typically used in systems like keyword-based spam filters, where emails are assigned a label (spam or not spam) based on a predetermined rule or list of keywords.[45]

2. Model-Based Reflex Agents. Building on the simple reflex agent foundation, model-based reflex agents maintain an internal state, allowing them to adapt actions based on historical context or data, which is similar to how smart thermostats adjust temperature based on past patterns.[46]



3. Goal-Based Agents. These agents introduce a layer of intentionality, selecting actions based on whether they can help fulfill a defined objective.[47] A travel booking agent that books flights and coordinates lodging accommodations with minimal human input would fall into this third category of agents.[48]

43. "Agents in Artificial Intelligence," GeeksforGeeks, June 5, 2023. https://www.geeksforgeeks.org/agents-artificial-intelligence; Douglas B. Laney, "Understanding And Preparing For The 7 Levels Of AI Agents," *Forbes*, Jan. 3, 2025. https://www.forbes.com/sites/douglaslaney/2025/01/03/understanding-and-preparing-for-the-seven-levels-of-ai-agents.
44. Gutowska. https://www.ibm.com/think/topics/ai-agents.
45. Ibid.
46. Ibid.
47. Ibid.
48. Ibid.

4. Utility-Based Agents. Utility-based agents take this further by weighing possible outcomes by determining which course of action is likely most beneficial, such as optimizing delivery routes to save time, value, or fuel.[49]

5. Learning Agents. These agents extend beyond the fixed strategies in the aforementioned four categories by continuously updating their approach based on feedback and new data.[50] These systems are capable of refining their performance over time, such as an AI system that personalizes lesson plans based on a given student's behavior and progress.[51]

6. Multi-Agent Systems. In more complex environments, multi-agent systems bring together different agents that work cooperatively or competitively to complete shared tasks, such as coordinating supply chain logistics.[52]

7. Hierarchical Agents. Hierarchical agents structure decision-making across various levels, delegating sub-tasks and managing dependencies in a way that mirrors organizational workflows.[53]

Many of today's leading agents, including Google's Project Astra, OpenAI's Operator, and CrewAI, reflect a growing trend: the emergence of general-purpose systems designed for flexible use across diverse environments and industries.[54]

As agentic AI matures, efforts to establish cybersecurity, interoperability, and governance standards are already underway.[55] These include initiatives like the novel Multi-Agent Environment, Security, Threat, Risk and Outcome (MAESTRO) threat modeling framework; sandboxing and permissioning strategies; and increased attention toward memory constraints and data boundaries.[56] Understanding the historical roots, agentic infrastructure stack layers, and practical distinctions between existing types of agents is essential for making sense of today's evolving developments and anticipating the cybersecurity and policy considerations that lie ahead.

Cybersecurity Benefits of AI Agents

Thanks to their enhanced autonomy, advanced reasoning, and capacity for continuous self-improvement, AI agents are already being deployed to streamline customer service workflows, assist with preliminary legal research, and automate data entry.[57] Though each of these business applications is significant, AI agents

49. Ibid.
50. Ibid.
51. Ibid.
52. "Agents in Artificial Intelligence." https://www.geeksforgeeks.org/agents-artificial-intelligence.
53. Ibid.
54. Mark Purdy, "What is Agentic AI, and How Will It Change Work?," Harvard Business Review, Dec. 12, 2024. https://hbr.org/2024/12/what-is-agentic-ai-and-how-will-it-change-work; Heaven. https://www.technologyreview.com/2024/05/14/1092407/googles-astra-is-its-first-ai-for-everything-agent; Bhavishya Pandit, "CrewAI: A Guide With Examples of Multi AI Agent Systems," DataCamp, Sept. 12, 2024. https://www.datacamp.com/tutorial/crew-ai.
55. Toner. https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents.
56. Ken Huang, "Agentic AI Threat Modeling Framework: MAESTRO," Cloud Security Alliance, Feb. 6, 2025. https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro; Sam Adler, "Interoperable Agentic AI: Unlocking the Full Potential of AI Specialization," Tech Policy Press, Dec. 3, 2024. https://www.techpolicy.press/interoperable-agentic-ai-unlocking-the-full-potential-of-ai-specialization; Stephen Weigand, "Cybersecurity in 2025: Agentic AI to change enterprise security and business operations in year ahead," SC Media, Jan. 9, 2025. https://www.scworld.com/feature/ai-to-change-enterprise-security-and-business-operations-in-2025; Shubham Sharma, "The new paradigm: Architecting the data stack for AI agents," VentureBeat, Nov. 14, 2024. https://venturebeat.com/data-infrastructure/the-new-paradigm-architecting-the-data-stack-for-ai-agents.
57. Zhixuan Chu, "Professional Agents – Evolving Large Language Models into Autonomous Experts with Human-Level Competencies," arXiv, Feb. 6, 2024. https://arxiv.org/html/2402.03628v1; Yasar. https://www.techtarget.com/searchenterpriseai/definition/AI-agents.

in those contexts typically support administrative, routine, and well-structured tasks. In contrast, AI agents in cybersecurity are increasingly deployed as copilots to human analysts on the frontlines, actively engaging with unfolding incidents, making rapid decisions in unpredictable scenarios, and operating in high-stress environments.[58] In other words, AI agents are not only accelerating efficiency but also strengthening cyber resilience by autonomously performing tasks critical to continuous monitoring, vulnerability management, threat detection, incident response, and decision-making across the cyber workforce.[59]

Continuous Attack Surface Monitoring and Vulnerability Management

As AI and other emerging technologies continue to expand across cloud infrastructure, third-party platforms, Internet of Things (IoT) devices, and edge environments, the overall attack surface that individuals and organizations must continuously monitor and secure has become increasingly fragmented and challenging.[60] While edge computing allows for greater decentralization and localized processing, the growing interconnectivity between devices, applications, and services can also amplify existing vulnerabilities, expand the number of possible points of failure, introduce new threat vectors, and reduce overall visibility into risk, especially in environments where interoperability remains inconsistent or poorly managed.[61] This is particularly true as AI-enabled systems often interact with external APIs, open-source resources, and real-time data streams, many of which are difficult to track, vet, or fully control.[62]

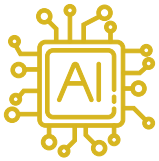
While still largely supervised by human analysts, AI agents are increasingly capable of autonomously assisting with key tasks, including mapping systems, identifying exposures, and prioritizing patches based on anticipated severity or business impact.

Traditional vulnerability management approaches, which depend on periodic scans or scheduled patching, are not well-suited for today's rapidly evolving and distributed environments.[63] AI agents offer a more adaptive and continuous alternative. While still largely supervised by human analysts, AI agents are increasingly capable of autonomously assisting with key tasks, including mapping systems, identifying exposures, and prioritizing patches based on anticipated severity or business impact.[64] As AI develops over time, these systems may help accelerate or even automate portions of the patching process, such as recommending fixes, triggering rollbacks, or adjusting configurations based on real-time context.[65]

Recent examples show how agentic capabilities are already being explored in real-world cybersecurity use cases. For instance, in late 2024, Google's Project Zero and DeepMind claimed to be the first in the world to successfully use an AI agent to uncover a "previously unknown, zero-day, exploitable memory-safety

58. Thomas Caldwell, "The Evolution Of AI Agents In The Third Wave Of AI," *Forbes*, Oct. 22, 2024. https://www.forbes.com/councils/forbestechcouncil/2024/10/22/the-evolution-of-ai-agents-in-the-third-wave-of-ai.
59. Weigand. https://www.scworld.com/feature/ai-to-change-enterprise-security-and-business-operations-in-2025.
60. Haiman Wong, "Securing the Future of AI at the Edge: An Overview of AI Compute Security," R Street Institute, July 16, 2024. https://www.rstreet.org/research/securing-the-future-of-ai-at-the-edge-an-overview-of-ai-compute-security.
61. Pete Bartolik, "Edge Requires Interoperability," CIO, May 19, 2021. https://www.cio.com/article/191756/edge-requires-interoperability.html.
62. Charles Owen-Jackson, "How cyber criminals are compromising AI software supply chains," Security Intelligence, Sept. 6, 2024. https://securityintelligence.com/articles/cyber-criminals-compromising-ai-software-supply-chains.
63. "AI is Now Exploiting Known Vulnerabilities – And What You Can Do About It," Cloud Security Alliance, June 26, 2024. https://cloudsecurityalliance.org/blog/2024/06/26/ai-is-now-exploiting-known-vulnerabilities-and-what-you-can-do-about-it.
64. Maria Korolov, "AI agents can find and exploit known vulnerabilities, study shows," CSO, July 2, 2024. https://www.csoonline.com/article/2512791/ai-agents-can-find-and-exploit-known-vulnerabilities-study-shows.html.
65. Ibid.

vulnerability in widely used real-world software."[66] Other AI agents are being trained to autonomously simulate attacks on enterprise systems, effectively performing red-teaming exercises to identify and test for vulnerabilities before adversaries can exploit them.[67] Both of these examples reflect a broader shift in how vulnerability management can be approached. As attack surfaces continue to evolve, AI agents could be key not only for helping cyber defenders keep up with emerging threats but also for advancing the speed and scale at which new and amplified vulnerabilities can be identified, prioritized, and remediated.[68]



As attack surfaces continue to evolve, AI agents could be key not only for helping cyber defenders keep up with emerging threats but also for advancing the speed and scale at which new and amplified vulnerabilities can be identified, prioritized, and remediated.

Real-Time Threat Detection and Incident Response

Although AI has already broadly demonstrated its value in cybersecurity through anomaly detection, natural language processing for threat intelligence, and the automation of repetitive or lower-level tasks, AI agents offer a promising leap forward.[69] Due to their modular architecture, memory, and capacity for goal-oriented, multi-step execution, AI agents can continuously learn from evolving threat patterns, correlate disparate signals, and initiate the appropriate responses without direct human oversight.[70] This is especially useful in high-speed or high-volume environments, where even a small delay between detection and response can affect the trajectory of an emerging threat and the success of incident containment efforts.[71]

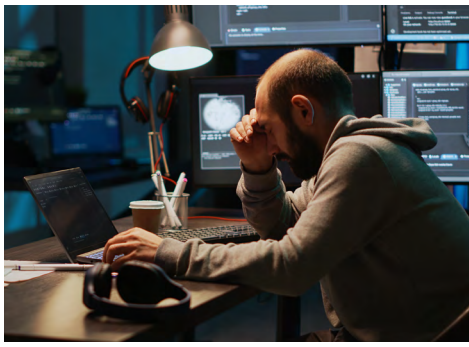
For example, in security operations centers (SOCs), AI agents can be deployed to monitor network traffic, flag anomalies, and trigger isolation protocols for systems that are suspected to be compromised.[72] A multi-agent setup could divide responsibilities between network monitoring, threat intelligence synthesis, and automated remediation.[73] These capabilities are already being demonstrated in enterprise settings with emerging security tools like Microsoft's Security Copilot agents, Simbian's SOC AI Agent, and DropZone AI's SOC Analyst, among others.[74] Once an intrusion is detected, these agents not only flag the emerging threat – they can initiate immediate responses, such as coordinating with firewalls or endpoint protection platforms to isolate affected nodes, notifying administrators, beginning system recovery procedures, or even a combination of these tasks.[75] This level of speed and coordination is particularly important because it means AI agents can help reduce the mean time to detect (MTTD) and mean time to

66. Davey Winder, "Google Claims World First As AI Finds 0-Day Security Vulnerability," *Forbes*, Nov. 5, 2024. https://www.forbes.com/sites/daveywinder/2024/11/05/google-claims-world-first-as-ai-finds-0-day-security-vulnerability.
67. Sarah Nagar and David Eaves, "An Agentic Shield? Using AI Agents to Enhance the Cybersecurity of Digital Public Infrastructure," New America, Dec. 19, 2024. https://www.newamerica.org/digital-impact-governance-initiative/blog/an-agentic-shield-using-ai.
68. "The Role of AI in Attack Surface Management: Enhancing Cyber Defense," Cyble, Feb. 13, 2025. https://cyble.com/knowledge-hub/ai-attack-surface-management.
69. "R Street Cybersecurity-Artificial Intelligence Working Group," R Street Institute, last accessed March 29, 2025. https://www.rstreet.org/home/our-issues/cybersecurity-and-emerging-threats/cyber-ai-working-group.
70. Grant Gross, "Agentic AI: 6 Promising Use Cases for Business," CIO, Nov. 14, 2024. https://www.cio.com/article/3603856/agentic-ai-promising-use-cases-for-business.html.
71. Jim Routh, "Milliseconds Matter: Defending Against the Next Zero-Day Exploit," Cloud Security Alliance, March 14, 2022. https://cloudsecurityalliance.org/blog/2022/03/14/milliseconds-matter-defending-against-the-next-zero-day-exploit.
72. Jimmy Astle, "Incorporating AI agents into SOC workflows," Red Canary, Jan. 16, 2025. https://redcanary.com/blog/threat-detection/ai-agents.
73. Antonella C. Garcia et al., "A Multi-Agent System for Addressing Cybersecurity Issues in Social Networks," CEUR Workshop Proceedings, Dec. 31, 2022. https://ceur-ws.org/Vol-3495/paper_05.pdf.
74. Thomas Claburn, "AI agents swarm Microsoft Security Copilot," The Register, March 24, 2025. https://www.theregister.com/2025/03/24/microsoft_security_copilot_agents; Kevin Townsend, "Simbian Introduces LLM AI Agents to Supercharge Threat Hunting and Incident Response," Security Week, Oct. 10, 2024. https://www.securityweek.com/simbian-introduces-llm-ai-agents-to-supercharge-threat-hunting-and-incident-response; "Dropzone AI," last accessed March 29, 2025. https://www.dropzone.ai.
75. Ibid.

respond (MTTR), both of which are crucial metrics in mitigating the scope and cost of cybersecurity incidents.[76]

Augmented Decision-Making and Cyber Workforce Support

In recent years, reports consistently underscore a persistent cyber workforce gap.[77] In 2025, the World Economic Forum reported a shortage of more than 4 million cyber professionals globally.[78] In the United States alone, the shortfall is estimated to be between 500,000 and 700,000 workers.[79] Moreover, existing cyber teams are facing increased workplace demands, with many professionals reporting unrelenting hours and workloads.[80] In fact, a recent survey reported that more than 80 percent of respondents experienced burnout.[81] This has caused many frontline cyber defenders and experts to consider not only leaving their positions, but leaving the industry as a whole, signaling a growing crisis that shows little sign of abating.[82] While many organizations are already using AI-enhanced tools to help streamline workflows and accelerate upskilling throughout their teams, the accelerating scope, speed, and sophistication of cyber threats often outpace these incremental gains.[83]



A recent survey reported that more than 80 percent of respondents experienced burnout. This has caused many frontline cyber defenders and experts to consider not only leaving their positions, but leaving the industry as a whole, signaling a growing crisis that shows little sign of abating.

As a result, AI agents are quickly entering the cyber workforce, not as replacements for human analysts but rather as force-multiplying copilots.[84] Although the current generation of AI agents is still far from perfect, they are already adept at a variety of critical tasks, including tuning firewalls, reducing noise by deduplicating security alerts, classifying security alerts by severity, using telemetry thresholds and anomaly detection to enforce policy changes, and more.[85] In doing so, AI security copilots, such as Cisco's AI Assistant, CrowdStrike's Charlotte AI, Fortinet's Advisor, Trellix's WISE, and Google's Sec-PaLM and AI Workbench, are gaining traction to help organizations keep their SOCs adequately staffed and efficient to better contain threats.[86] Moreover, when armed with these AI security copilots, SOCs are seeing notable improvements in false-positive rates (up to 70 percent) while reducing manual triage by more than 40 hours a week.[87] These early successes demonstrate how AI agents are emerging as a technological solution that can help organizations save time and money while improving their cyber resilience and retaining their cyber defenders.[88]

76. Greg Zemlin, "MTTD and MTTR in Cybersecurity Incident Response," Wiz, Sept. 5, 2024. https://www.wiz.io/academy/mttd-and-mttr.
77. Michelle Meineke, "The cybersecurity industry has an urgent talent shortage. Here's how to plug the gap," World Economic Forum, April 28, 2024. https://www.weforum.org/stories/2024/04/cybersecurity-industry-talent-shortage-new-report.
78. "Bridging the Cyber Skills Gap," World Economic Forum, last accessed March 29, 2025. https://initiatives.weforum.org/bridging-the-cyber-skills-gap/home.
79. Sophia Fox-Sowell, "To expand cyber workforce, government must unfreeze hiring and target youth, experts told House committee," StateScoop, Feb. 5, 2025. https://statescoop.com/cybersecurity-workforce-house-committee-homeland-security.
80. Matt Kapko, "Are cybersecurity professionals OK?," Cybersecurity Dive, Aug. 7, 2024. https://www.cybersecuritydive.com/news/cyber-security-burnout-stress-anxiety/723470.
81. Ibid.
82. Carolyn Crist, "Skills shortage persists in cybersecurity despite decade of hiring," HR Dive, Oct. 16, 2024. https://www.hrdive.com/news/skills-shortage-persists-in-cybersecurity-despite-hiring/729988; Joe McKendrick, "Will AI take the wind out of cybersecurity job growth?," ZDNet, July 10, 2024. https://www.zdnet.com/article/will-ai-take-the-wind-out-of-cybersecurity-job-growth; Daniel Pell, "AI Versus The Skills Gap," *Forbes*, April 17, 2024. https://www.forbes.com/councils/forbestechcouncil/2024/04/17/ai-versus-the-skills-gap.
83. Ivan Belcic and Cole Stryker, "AI Agents in 2025: Expectations vs. reality," IBM, March 4, 2025. https://www.ibm.com/think/insights/ai-agents-2025-expectations-vs-reality.
84. Columbus. https://venturebeat.com/security/ai-copilots-cut-false-positives-and-burnout-in-overworked-socs.
85. Ibid.
86. Ibid.
87. Ibid.
88. Gary Grossman, "Onboarding the AI workforce: How digital agents will redefine work itself," VentureBeat, Sept. 29, 2024. https://venturebeat.com/ai/onboarding-the-ai-workforce-how-digital-agents-will-redefine-work-itself.

Cybersecurity Considerations and Potential Risks

AI agents are proving to be powerful not only because of what they can do on their own, but also because of how effective they are at learning and adapting across digital environments based on new data or updated information. Unfortunately, the same capabilities that make AI agents impressive, such as their memory, autonomy, and reasoning, can also make them attractive targets for exploitation.[89]

Though there are different ways to conceptualize an AI agent's architecture, we organize the agentic infrastructure stack into four primary layers: perception, reasoning, action, and memory. Each layer corresponds to a critical stage in how data is collected, analyzed, applied, and refined throughout the AI agent lifecycle. Because each layer serves a distinct function within the AI agent's workflow, the risks and mitigation needs associated with them also differ between modules, shaping the cybersecurity considerations at each stage.

Layer 1: Perception Module

At this first layer, the agent is tasked with scanning and observing a given environment through sensors (i.e., cameras, data inputs) to provide it with foundational context, and that data is then transformed into a suitable format for processing.[90] Because the perception module relies on multiple data pipelines for its analysis, this layer could face a variety of data-specific security risks that would affect the data confidentiality and integrity of the agentic workflow. These attacks include—but are not limited to—adversarial data injection (also known as data poisoning) and AI model supply chain risks.

Adversarial data injection is one of the most prominent security risks against the perception layer of an agent workflow because it tampers with the model's integrity and the agent's ability to factually analyze the data points in its training.

Adversarial data injection is one of the most prominent security risks against the perception layer of an agent workflow because it tampers with the model's integrity and the agent's ability to factually analyze the data points in its training.[91] For example, bad actors could seamlessly insert modifications that mislead vision models into incorrect characterizations and imprecise content classifications on behalf of the agent. In the case of image processing, bad actors could manipulate the image pixels, add extra noise to the image, or perform other types of perturbation that are difficult to notice both with the human eye and via AI-enabled perception systems.[92]

While the manipulation of image pixel values to deceive the agent is a common adversarial data risk, researchers have found that even small-scale perturbations in a dataset can meaningfully affect an agent's learning process—causing it to misclassify inputs into either "a maliciously-chosen target class (in a targeted attack) or classes that are different from the ground truth."[93] These types of

89. Matt Dangelo, "The rise of autonomous AI: How intelligent agents are redefining strategy, risk, & compliance," Thomson Reuters, March 3, 2025. https://www.thomsonreuters.com/en-us/posts/technology/autonomous-agentic-ai; Maria Korolov, "AI agents will transform business processes – and magnify risks," CIO, Aug. 21, 2024. https://www.cio.com/article/3489045/ai-agents-will-transform-business-processes-and-magnify-risks.html.
90. Haziqa Sajid, "What Are AI Agents, and How Do They Work?," Lakera AI, Nov. 13, 2024. https://www.lakera.ai/blog/what-are-ai-agents.
91. Forest McKee and David Noever, "Transparency Attacks: How Imperceptible Image Layers Can Fool AI Perception," arXiv, Jan. 29, 2024. https://arxiv.org/abs/2401.15817.
92. Ian J. Goodfellow et al., "Explaining and Harnessing Adversarial Examples," arXiv, Dec. 20, 2014. https://arxiv.org/abs/1412.6572; Battista Biggio and Fabio Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," arXiv, Dec. 8, 2017. https://arxiv.org/abs/1712.03141.
93. Chaowei Xiao et al., "Generating Adversarial Examples with Adversarial Networks," arXiv, Feb. 14, 2019. https://arxiv.org/pdf/1801.02610; Alexey Kurakin et al., "Adversarial examples in the physical world," arXiv, July 8, 2016. https://arxiv.org/abs/1607.02533.

attacks are particularly challenging, as bad actors can execute them without having any direct access to the model architecture.

Such data poisoning methods may also "reorient" the agent's data analysis from one intended pattern, set by developers, to a malicious one, set by bad actors, by altering the training set's distribution or reshaping the data to align with adversarial objectives. For instance, in a backdoor attack, an adversary could deliberately modify the training data to introduce specific triggers that, when encountered, would cause the model to behave in a predetermined, often malicious way.

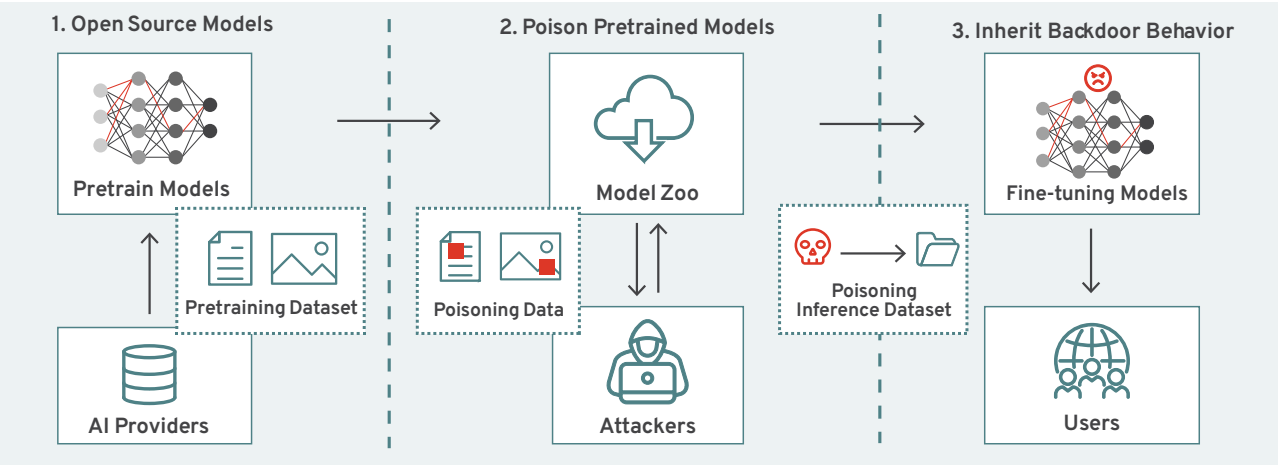
In a backdoor attack, an adversary could deliberately modify the training data to introduce specific triggers that, when encountered, would cause the model to behave in a predetermined, often malicious way.

These kinds of cybersecurity risks are especially concerning at the perception layer because of its heavy reliance on state-of-the-art foundation models—many of which are externally sourced—creating additional dependencies.[94] While these models are critical for enabling advanced agent performance, their integration also expands the agent's exposure to potential software supply chain vulnerabilities, particularly during the pre-training phase.

In fact, threat actors can exploit the decentralized nature of the AI and software supply chain by embedding malicious data within these foundational models at the pre-training phase. The nature of this attack depends on the target, which can range from data poisoning to weight poisoning, along with the method of label modification.[95] Both types of backdoor attacks can lead to compromised downstream performance in agentic systems.

Figure 3 illustrates the backdoor attack process on pre-trained foundation models. In many cases, these attacks are difficult to detect, and, therefore, challenging to mitigate. This also increases the likelihood of risk transfer from the foundation model to the AI agent itself.[96] If transferred, the AI agent may inherit these vulnerabilities and carry them forward into deployment.

Figure 3: Framework of Backdoor Attacks to Unified Foundation Models



Source: Zenghui Yuan et al., "Backdoor Attacks to Pre-trained Unified Foundation Models," arXiv, Feb. 23, 2023 https://arxiv.org/pdf/2302.09360

94. Lareina Yee et al., "Why agents are the next frontier of generative AI," McKinsey Digital, July 24, 2024. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/why-agents-are-the-next-frontier-of-generative-ai; Laura French, "How AI coding assistants could be compromised via rules file," SC World, March 18, 2025. https://www.scworld.com/news/how-ai-coding-assistants-could-be-compromised-via-rules-file.
95. Yuan. https://arxiv.org/pdf/2302.09360.
96. Hao Wang et al., "Model Supply Chain Poisoning: Backdooring Pre-trained Models via Embedding Indistinguishability," OpenReview.net, Jan. 29, 2025. https://openreview.net/forum?id=VWQwwMxFht#discussion.

In addition to poisoning and backdoor attacks, recent reports have uncovered a vulnerability in the Safetensors conversion service that Hugging Face—a leading open-source platform used by developers to host, customize, and share pre-trained machine learning models—offers.[97] According to AI security firm HiddenLayer, attackers can "send malicious pull requests with attacker-controlled data from the Hugging Face service to any repository on the platform, as well as hijack any models that are submitted through the conversion service."[98] This vulnerability presents a heightened security risk, particularly given Hugging Face's role as a major hub for pre-trained models. In practice, such exploitation could allow threat actors to impersonate a chatbot and submit malicious inquiries, ranging from instructions on how to successfully conduct a money heist to even building a nuclear bomb or a bioweapon.

Layer 2: Reasoning Module

The second layer of the AI agent workflow is the reasoning module, which governs the agent's internal decision-making processes. At this stage, data collected in the perception module in layer 1 is interpreted and transformed into actionable outputs. The agent reviews and analyzes contextual information and may apply pre-learned heuristics, patterns, or logical ordering to generate a conclusion with the support of specialized hardware like graphic processing units or tensor processing units and model hosting environments.[99] For example, an agent might analyze network activity logs to determine whether a user request is legitimate or suspicious, drawing on historical behaviors and anomaly detection models to inform its decisions. Because the reasoning module plays a central role in analysis and judgment, vulnerabilities and bad cyber hygiene in this layer can lead to incorrect decisions or mischaracterizations, particularly if adversaries manipulate the signals or exploit vulnerabilities in the models or supporting infrastructure. Ultimately, this inaccuracy could undermine the end-user's trust in the agent's reliability and accuracy.

For example, an agent might analyze network activity logs to determine whether a user request is legitimate or suspicious, drawing on historical behaviors and anomaly detection models to inform its decisions.

One of the most common process-level security risks at this stage is the exploitation of the model's underlying vulnerabilities. These flaws could stem from widely used AI frameworks like PyTorch, which plays a critical role in the reasoning layer by enabling developers to build, train, and fine-tune machine learning models. PyTorch is also commonly used for deep learning model development, inference processing, and optimization, making it a core component of many AI agent workflows.[100]

Security flaws can also arise from misconfigured libraries and insecure model hosting environments, especially those that allow user-generated model uploads without robust validation.[101] In early 2024, for example, researchers found that

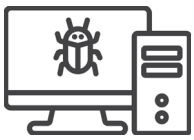
97. Ravie Lakshmanan, "New Hugging Face Vulnerability Exposes AI Models to Supply Chain Attacks," The Hacker News, Feb. 27, 2024. https://thehackernews.com/2024/02/new-hugging-face-vulnerability-exposes.html.
98. "Hijacking Safetensors Conversion on Hugging Face," HiddenLayer Inc., Feb. 21, 2024. https://hiddenlayer.com/innovation-hub/silent-sabotage.
99. Alexander De Ridder, "How Intelligent Agents Use Knowledge Representation for Decision-Making," SmythOS, Jan. 31, 2025. https://smythos.com/ai-agents/ai-tutorials/intelligent-agents-and-knowledge-representation; Barbara Bickham, "The Role of Heuristics in Developing Smarter AI Systems," Trailyn Ventures, April 19, 2024. https://www.trailyn.com/the-role-of-heuristics-in-developing-smarter-ai-systems.
100. "Getting Started," PyTorch, last accessed March 30, 2025. https://pytorch.org; Ashish Kurmi, "PyTorch Supply Chain Compromise," StepSecurity, Dec. 9, 2024. https://www.stepsecurity.io/blog/pytorch-supply-chain-compromise.
101. Elizabeth Montalbano, "Hugging Face AI Platform Riddled with 100 Malicious Code-Execution Models," Dark Reading, Feb. 29, 2024. https://www.darkreading.com/application-security/hugging-face-ai-platform-100-malicious-code-execution-models.

the Hugging Face platform was home to approximately 100 malicious machine learning models capable of depositing malicious code onto users' machines.[102]

Another category of attacks that could affect the agent's knowledge base are model exploitation attacks. Instead of targeting AI inputs directly, adversaries can attempt to probe the AI's internal logic to extract proprietary knowledge, internal decision pathways, or sensitive training data. There are three techniques that threat actors can use to exploit models. First, bad actors can attempt to extract personal identifiable information (PII) by reconstructing aspects of the training data.[103] This is also known as a model inversion attack. Another approach is black-box extraction, where an attacker without direct access to the model's architecture or weights submits iterative queries to infer or replicate the model, leading to intellectual property theft or downstream adversarial attacks.[104] Finally, threat actors may attempt to jailbreak or probe the model's logic by crafting prompts that aim to trick the model into revealing its underlying structure or how the agent processes information. As these bad actors continuously refine their prompts, they can analyze outputs to map the agent's decision-making boundaries, identify vulnerabilities to exploit later, and improve hallucination tactics that degrade or mislead agent performance over time.[105]

Layer 3: Action Module

The third layer of the agentic workflow is the action module, which is responsible for translating the decision-making process made in layer 2 into real-world operations. Because this is the stage where actions are executed, even seemingly minor manipulations can lead to unintended—and potentially harmful—consequences. This makes the action module particularly sensitive to attacks that exploit an agent's ability to interface with external systems.



Bad actors could compromise this layer through a variety of vectors, including, but not limited to prompt injection, command hijacking, unauthorized access, privilege escalation, and vulnerabilities within API integrations.

Bad actors could compromise this layer through a variety of vectors, including, but not limited to prompt injection, command hijacking, unauthorized access, privilege escalation, and vulnerabilities within API integrations. These risks underscore the importance of implementing stringent output validation and access controls at this layer.

In prompt injection attacks, threat actors feed malicious prompts into the agent with the goal of manipulating the agent to perform actions outside the scope of its intended purpose, like, leaking PII or generating malicious outputs.[106]

While prompt injection focuses on manipulating inputs to modify the agent's behavior, command hijacking takes this one step further by executing unauthorized commands based on the earlier inputs. One well-documented example is the "Imprompter" attack in which attackers manipulated AI systems using deceptive prompts to retrieve sensitive information like names, email

102. Ibid.
103. "Model inversion attacks," Michalsons, March 8, 2023. https://www.michalsons.com/blog/model-inversion-attacks-a-new-ai-security-risk/64427?utm_source=chatgpt.com.
104. Mumtaz Fatima and Amy Chang, "Safeguarding AI: A Policymaker's Primer on Adversarial Machine Learning Threats," R Street Institute, March 20, 2024. https://www.rstreet.org/commentary/safeguarding-ai-a-policymakers-primer-on-adversarial-machine-learning-threats.
105. Rachneet Sachdeva et al., "Turning Logic Against Itself: Probing Model Defenses Through Contrastive Questions," arXiv, Jan. 3, 2025. https://arxiv.org/abs/2501.01872.
106. Matthew Kosinki and Amber Forrest, "What is a prompt injection attack?," IBM, March 26, 2024. https://www.ibm.com/think/topics/prompt-injection.

addresses, and payment details.[107] This incident illustrates how seemingly minor vulnerabilities and exploitation techniques can be leveraged to compromise user privacy and disrupt system integrity.

Another security risk at this layer involves unauthorized access through privilege escalation. Because AI agents operate across different structured layers of execution and interact with a range of systems, such as data stores, applications, end users, and pre-trained foundation models, any security gap could allow threat actors to move laterally within the agentic workflow. This privilege escalation could result in misconfigured role-based access controls, which allow bad actors to access restricted modalities and the foundations of the underlying model or to prompt the model to perform unauthorized actions.[108] In a recent example, researchers from Palo Alto Networks identified two vulnerabilities in Google's Vertex AI platform that could have enabled bad actors to escalate privileges and extract models.[109]

As mentioned earlier in this section, the action module is also vulnerable to insecure execution permissions that stem from weak access controls. These security flaws could grant bad actors the ability to poison models or trick them into executing deceptive or harmful requests under the guise of legitimate prompts.

Since this layer is the primary interface between AI agents and external connections, APIs are a critical, and often undersecured, vector for exploitation. If communication channels are not secured with robust safeguards, bad actors could intercept and manipulate different model requests and responses through man-in-the-middle attacks or re-prompt previous AI queries to gain access to restricted areas.[110] For example, a ransomware group could exploit a vulnerable fraud-detection API to alter transaction data while evading detection.

More broadly, API vulnerabilities can also stem from failures in endpoint detection, missing or improperly validated API keys, or lax token authentication.[111] These weaknesses can open the door for threat actors to bypass restrictions, achieve privilege escalation, and manipulate the agent's behavior.[112] In doing so, threat actors could also execute adversarial prompt injection or other forms of input manipulation that lead to harmful outputs.[113]

In addition to APIs, AI agents often rely on third-party services for data analysis. This adds yet another layer of cybersecurity risk: Compromised datasets, insecure API dependencies, or insufficient monitoring can allow threat actors to tamper with

Because AI agents operate across different structured layers of execution and interact with a range of systems, such as data stores, applications, end users, and pre-trained foundation models, any security gap could allow threat actors to move laterally within the agentic workflow.

107. Matt Burgess, "This Prompt Can Make an AI Chatbot Identify and Extract Personal Details From Your Chats," *Wired*, Oct. 17, 2024. https://www.wired.com/story/ai-imprompter-malware-llm/?utm_source=chatgpt.com.
108. Ximeng Liu et al., "Privacy and Security Issues in Deep Learning: A Survey," IEEE Access, December 2020. https://www.researchgate.net/publication/347639649_Privacy_and_Security_Issues_in_Deep_Learning_A_Survey.
109. Ofir Balassiano and Ofir Shaty, "ModeLeak: Privilege Escalation to LLM Model Exfiltration in Vertex AI," Unit 42 Palo Alto Networks, Nov. 12, 2024. https://unit42.paloaltonetworks.com/privilege-escalation-llm-model-exfil-vertex-ai.
110. Wenqi Sun et al., "Clustering Mobile Apps based on Design and Manufacturing Genre," IEEE Xplore, December 2020. https://ieeexplore.ieee.org/document/9344944.
111. "Wallarm Releases 2025 API ThreatStats Report, Revealing APIs are the Predominant Attack Surface," Wallarm, Jan. 29, 2025. https://www.wallarm.com/press-releases/wallarm-releases-2025-api-threatstats-report.
112. "API Security's Role in Responsible AI Deployment," Wallarm, Jan. 21, 2025. https://lab.wallarm.com/api-securitys-role-in-responsible-ai-deployment.
113. "Discover and Protect Generative AI APIs," Traceable AI, last accessed March 30, 2025. https://www.traceable.ai/securing-gen-ai-apis.

AI agent operations without detection.[114] If any part of the third-party software supply chain is compromised, the agent's performance and trustworthiness could be severely degraded and more difficult to immediately remediate and restore.

Many agents are also deployed in cloud environments, which carry their own set of cyber risks and potential vulnerabilities. Weak configuration settings, such as unsigned code releases or poorly managed access controls, leave systems exposed to software supply chain attacks and malicious model updates.[115] Furthermore, security flaws in cloud storage buckets can also leak sensitive model parameters, inviting intellectual property theft and other similar types of adversarial attacks.[116]

Layer 4: Memory Module

The fourth and final layer of the agentic workflow is the memory module, which is responsible for retaining context across tasks, storing relevant data, and informing future decisions based on past interactions.[117] This module distinguishes AI agents from other AI models or LLM-based tools, which typically operate within a single session or a query window.[118] By enabling long-term context awareness, learning persistence, and memory-driven adaptability, the memory module facilitates the AI agent's continuous self-improvement capabilities over time.[119]

One of the primary cybersecurity risks that can occur at this layer is memory tampering or corruption, where threat actors manipulate stored memory to distort an agent's understanding or to introduce incorrect historical data.[120] This can occur through learning stream poisoning (maliciously modifying real-time inputs that are then retained as memory) or through unauthorized edits to at-rest memory databases.[121] These attacks could degrade AI agent performance or subtly influence future actions toward harmful outputs.[122]

Relatedly, unauthorized data retention is another cybersecurity risk within this layer.[123] When unauthorized data retention occurs, AI agents remember data or information they were not supposed to retain, either because they inadvertently collected data outside of its intended use case or learning scope, retained data longer than permitted, or failed to delete it when instructed

One of the primary cybersecurity risks that can occur at this layer is memory tampering or corruption, where threat actors manipulate stored memory to distort an agent's understanding or to introduce incorrect historical data.

114. Bryan McNaught, "API Protection for AI Factories: The First Step to AI Security," F5, Dec. 19, 2024. https://www.f5.com/company/blog/api-security-for-ai-factories; Yared Gudeta et al., "Securing the Future: How AI Gateways Protect AI Agent Systems in the Era of Generative AI," Databricks, Nov. 13, 2024. https://www.databricks.com/blog/ai-gateways-secure-ai-agent-systems.
115. Favour Efeoghene, "How to Secure CI/CD Pipelines Against Supply Chain Attacks," StartUp Growth Guide, July 24, 2024. https://startupgrowthguide.com/how-to-secure-ci-cd-pipelines-against-supply-chain-attacks.
116. Emmanuel Ok, "Addressing Security Challenges in AI-Driven Cloud Platforms: Risks and Mitigation Strategies," ResearchGate, February 2025. https://www.researchgate.net/publication/388997486_Addressing_Security_Challenges_in_AI-Driven_Cloud_Platforms_Risks_and_Mitigation_Strategies.
117. "AI Agents," Nvidia Glossary, last accessed March 30, 2025. https://www.nvidia.com/en-us/glossary/ai-agents.
118. Jeffrey Yang Fan Chiang et al., "Why Are Web AI Agents More Vulnerable Than Standalone LLMs? A Security Analysis," arXiv, Feb. 27, 2025. https://arxiv.org/html/2502.20383v1.
119. Xun Jiang et al., "Long Term Memory: The Foundation of AI Self-Evolution," arXiv, Oct. 21, 2024. https://arxiv.org/html/2410.15665v1.
120. Chad DeChant, "Episodic memory in AI agents poses risks that should be studied and mitigated," arXiv, Jan. 20, 2025. https://arxiv.org html/2501.11739v1?ref=community.heartcount.io.
121. Jaimin Patel, "Secure AI Agents by Design with AI Runtime Security," Palo Alto Networks, Jan. 23, 2025. https://www.paloaltonetworks.com/blog/network-security/secure-ai-agents-by-design-ai-runtime-security; Symantec Threat Hunter Team, "AI: Advent of Agents Open New Possibilities for Attackers," Symantec Enterprise Blogs, March 13, 2025. https://www.security.com/threat-intelligence/ai-agent-attacks.
122. Ibid.
123. Zhaorun Chen et al., "AgentPoison: Red-teaming LLM Agents via Poisoning Memory or Knowledge Bases," arXiv, July 17, 2024. https://arxiv.org/abs/2407.12784.

to.[124] This can lead to compliance violations in relation to existing privacy laws or user terms and conditions and unintentionally expose sensitive user information.[125] Even otherwise well-configured AI agents can face these cybersecurity challenges if existing memory governance guardrails are missing or improperly implemented.[126]

What makes the memory module particularly significant is its recursive relationship with the earlier three layers of the agentic lifecycle.[127] If the data lifecycle were conceptualized as a circle, this fourth layer effectively closes the loop, meaning that any vulnerabilities or risks introduced earlier in the process, such as poisoned data or training processes and faulty reasoning, may be not only preserved but reinforced over time.[128] For example, if adversarial data is ingested through the perception layer and not flagged as corrupted, the memory module could preserve it as a trusted input, continuing to apply that context to influence future reasoning processes and actions.[129] Similarly, if an attack manipulates an AI agent's logic at the reasoning module, the tasks it finishes in the action module may be remembered as valid precedent.[130]

In this way, memory does not simply inform an AI agent's future performance—it can also carry forward mistakes and risks from its past. Without strong protections and best practices for ensuring data accuracy, implementing retention boundaries, and managing memory, the memory module can become both a repository of insights and a source of cascading cyber vulnerabilities and risks.[131]

Without strong protections and best practices for ensuring data accuracy, implementing retention boundaries, and managing memory, the memory module can become both a repository of insights and a source of cascading cyber vulnerabilities and risks.

Anticipating Needs, Solutions, and Responsibilities

To advance our cyber preparedness in this rapidly unfolding era of AI agents, we must adopt a proactive, balanced, and adaptable strategy. Striking the right balance means encouraging policymakers, end-users, and developers to recognize and fully leverage the benefits that AI agents offer while anticipating and addressing the cybersecurity risks they may amplify or introduce.

The following recommendations highlight the policy needs, emerging technological solutions, and responsible design and deployment strategies that are best suited for supporting and guiding the advancement of AI agents.

Policy Needs

Establish Voluntary, Sector-Specific Guidelines for Human–Agent Collaboration

The White House should direct federal agencies, such as the National Institute of Standards and Technology (NIST), the Department of Labor, and relevant industry

124. Rashi Shrivastava, "The Prompt: Privacy Risks 'Haunt' AI Agents," *Forbes,* March 11, 2025. https://www.forbes.com/sites/rashishrivastava/2025/03/11/the-prompt-privacy-risks-haunt-ai-agents; David Ruiz, "New AI 'agents' could hold people for ransom in 2025," Malwarebytes, Feb. 4, 2025. https://www.malwarebytes.com/blog/news/2025/02/new-ai-agents-could-hold-people-for-ransom-in-2025.
125. Erich Kron, "Five privacy concerns around agentic AI," SC Media, Feb. 19, 2025. https://www.scworld.com/perspective/five-privacy-concerns-around-agentic-ai.
126. Daniel Berrick, "Minding Mindful Machines: AI Agents and Data Protection Considerations," Future of Privacy Forum, Feb. 5, 2025. https://fpf.org/blog/minding-mindful-machines-ai-agents-and-data-protection-considerations.
127. Christian Vasquez, "Cybersecurity pros are preparing for a new adversary: AI agents," Fortune, Feb. 18, 2025. https://fortune.com/2025/02/18/cybersecurity-pros-are-preparing-for-a-new-adversary-ai-agents.
128. Ibid.
129. DeChant. https://arxiv.org/html/2501.11739v1?ref=community.heartcount.io.
130. Ibid.
131. Berrick. https://fpf.org/blog/minding-mindful-machines-ai-agents-and-data-protection-considerations.

regulators, to develop voluntary, sector-specific guidelines that support secure, transparent, and human-centered agentic deployments.[132]

Rather than prescribing a rigid, one-size-fits-all mandate, these guidelines should encourage organizations—including AI laboratories, private-sector companies, and universities—to define tailored human–agent interaction frameworks. These frameworks should clarify when agents may be deployed, under what conditions they may act autonomously, whether they are permitted to learn independently, when human oversight is required, how responsibility is assigned in the event of failures, and what protocols exist for detecting, escalating, and correcting errors. The goal is to ensure that agents support—not replace—human decision-making and talent, especially in sensitive fields of work like healthcare and national security.[133]

The goal is to ensure that agents support–not replace–human decision-making and talent, especially in sensitive fields of work like healthcare and national security.

Given the potential of AI agents to reshape workforce dynamics and drive an "AI talent revolution," these guidelines should also promote organizational readiness for human–agent collaboration by offering recommendations for redesigning jobs and reskilling or upskilling current employees.[134] These guidelines should also highlight responsible deployment strategies, such as incremental rollouts, permissioning boundaries, and real-time escalation protocols, tailored to team-based environments where agents and humans are likely to interact.[135] While federal agencies can help provide baseline guidance and highlight priorities, public–private partnerships will be essential to translating these principles into practice. Consortia like the Partnership on AI, Microsoft's Tech-Labor Partnership with the American Federation of Labor and Congress of Industrial Organizations (AFL-CIO), and Cisco's AI-Enabled ICT Workforce Consortium provide early momentum and serve as leading examples of cross-sector collaboration.[136]

Expand and Facilitate Information Sharing and Multi-Stakeholder Collaboration on Evolving Agentic Risks

Given their potential to serve as force multipliers for offensive, defensive, and adversarial cyber operations, AI agents require equally coordinated and dynamic strategies for timely, cross-sector information sharing about emerging agentic risks, observed unintended agentic behaviors, deployment challenges, and successful risk mitigation strategies.[137] Specifically, the White House should direct federal agencies like the Cybersecurity and Infrastructure Security Agency to collaborate with sector-specific regulatory bodies and industry stakeholders to expand information-sharing forums and develop publicly available software tools and resources for testing and evaluating agentic security and performance.[138]

132. Haiman Wong and Brandon Pugh, "Key Cybersecurity and AI Policy Priorities for Trump's Second Administration and the 119th Congress," R Street Institute, Jan. 6, 2025. https://www.rstreet.org/research/key-cybersecurity-and-ai-policy-priorities-for-trumps-second-administration-and-the-119th-congress.
133. Shana Lynch, "Predictions for AI in 2025: Collaborative Agents, AI Skepticism, and New Risks," Stanford University Human-Centered Artificial Intelligence, Dec. 23, 2024. https://hai.stanford.edu/news/predictions-for-ai-in-2025-collaborative-agents-ai-skepticism-and-new-risks.
134. Whiting, "The rise of 'AI agents': What they are and how to manage the risks." https://www.weforum.org/stories/2024/07/what-is-an-ai-agent-experts-explain.
135. Ibid.
136. Rebecca Finlay, "Sharing AI Mistakes: Partnership on AI's Rebecca Finlay," MIT Sloan Management Review, Nov. 12, 2024. https://sloanreview.mit.edu/audio/sharing-ai-mistakes-partnership-on-ais-rebecca-finlay; "AFL-CIO and Microsoft Announce New Tech-Labor Partnership on AI and the Future of the Workforce," AFL-CIO, Dec. 11, 2023. https://aflcio.org/press/releases/afl-cio-and-microsoft-announce-new-tech-labor-partnership-ai-and-future-workforce; "Leading Companies Launch Consortium to Address AI's Impact on the Technology Workforce," Cisco, April 4, 2024. https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2024/m04/leading-companies-launch-consortium-to-address-ai-impact-on-the-technology-workforce.html.
137. Rosa Merced, "Exploring the potential of AI agents in combatting cyber threats," Outshift by Cisco, Feb. 18, 2025. https://outshift.cisco.com/blog/exploring-the-potential-of-ai-agents-in-combatting-cyber-threats.
138. Wong and Pugh. https://www.rstreet.org/research/key-cybersecurity-and-ai-policy-priorities-for-trumps-second-administration-and-the-119th-congress.

These efforts should emphasize use-case-specific transparency, such as anonymized incident reports and adversarial testing results, to accelerate collective learning and cyber preparedness.[139]

Prioritize Investments in Public–Private Partnerships for Advancing Agentic Security and Evaluation

Congress should prioritize investments in continued research and development initiatives aimed at strengthening the cybersecurity posture of AI agents across their full lifecycle.[140] While private companies naturally have strong incentives to secure their own products and services, many agentic risks—such as model hijacking, memory poisoning, and emergent multi-agent behavior—can cut across proprietary systems and lack clearly defined ownership or liability.[141]

In cases where agentic risks are cross-cutting and infrastructural, the federal government can play a limited but essential role by supporting foundational research and ensuring that key findings are made publicly available to foster broader coordination and informed risk mitigation for agentic developers and researchers. This can include funding adversarial testing, agent-specific risk modeling, and resilience evaluations focused on architectural features like memory integrity and autonomous decision-making.[142] While NIST has already taken early steps to evaluate hijacking risks in AI agents, scaling this work will require additional investments in competitive grants, interdisciplinary research hubs, and public–private partnerships that accelerate knowledge-sharing and innovation across sectors.[143]

Emerging Technological Solutions

Advance and Apply Automated Moving Target Defense (AMTD) Capabilities to Disrupt Evolving Exploitation Pathways

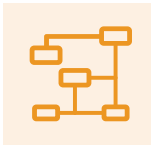
AMTD are systems designed to continuously alter a system's attack surface by shifting IP addresses, memory allocations, or control paths to deliberately complicate adversarial reconnaissance efforts and reduce system predictability.[144] When paired with the autonomous and continuous self-improvement capabilities of AI agents, AMTD systems could rotate access privileges, shuffle API endpoints, or re-randomize internal configurations to limit the persistence of adversarial probing attempts or prompt injection attacks.[145] These techniques are expected to be particularly useful in edge computing environments, where agents will need to remain flexible and responsive while operating across distributed and often interdependent digital environments.[146]

139. Ghose. https://scet.berkeley.edu/the-next-next-big-thing-agentic-ais-opportunities-and-risks.
140. Wong and Pugh. https://www.rstreet.org/research/key-cybersecurity-and-ai-policy-priorities-for-trumps-second-administration-and-the-119th-congress.
141. Phaedra Boinodiris and Jon Parker, "The evolving ethics and governance landscape of agentic AI," IBM, March 22, 2025. https://www.ibm.com/think/insights/ethics-governance-agentic-ai.
142. Ghose. https://scet.berkeley.edu/the-next-next-big-thing-agentic-ais-opportunities-and-risks.
143. "Technical Blog: Strengthening AI Agent Hijacking Evaluations," National Institute of Standards and Technology, Jan. 17, 2025. https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations.
144. Mark Loman, "Pioneering Automated Moving Target Defense (AMTD)," Sophos, Oct. 19, 2023. https://news.sophos.com/en-us/2023/10/19/pioneering-automated-moving-target-defense-amtd.
145. Sailik Sengupta, "Moving Target Defense: A Symbiotic Framework for AI & Security," Association for Computing Machinery Digital Library, May 8, 2017. https://dl.acm.org/doi/pdf/10.5555/3091125.3091473.
146. "Fujitsu develops world's first multi-AI agent security technology to protect against vulnerabilities and new threats," Fujitsu Limited, Dec. 12, 2024. https://www.fujitsu.com/global/about/resources/news/press-releases/2024/1212-01.html.

Adapt and Implement Hallucination Detection Tools for Continuous Agentic Security Monitoring

Originally developed to improve the quality control capabilities and accuracy of LLM outputs, hallucination detection tools are now quickly being repurposed for agentic security to identify reasoning flaws and gaps, anomalous or suspicious behavior, and low-confidence outputs before they reach the action module.[147] Emerging hallucination detection tools operate by using internal consistency checks, multi-source fact validation, and prompt-response tracking to monitor for misalignment, especially under adversarial or high-stress conditions.[148] In the context of AI agents, these hallucination detection tools are already proving successful at revealing compromised memory recalls, model drifts, and inference anomalies, all of which are essential to helping developers identify vulnerabilities before they can be exploited by threat actors.[149]

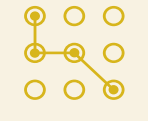
Develop and Adopt Agent Identifiers and Traceability Tools to Improve Oversight



To improve explainability and oversight, AI researchers and developers should continue creating identification infrastructure and persistent tools capable of tracking and logging the full arc of agent activity, including an agent's initial and expanding data-collection strategies; third-party dependencies and tool applications; completed tasks; pending actions; reasoning logic streams; and memory recall.[150] This approach builds on existing strategies of embedded provenance tracking and digital auditing that are used to enable real-time behavioral analysis, version-control tracking, validating software supply chain dependencies and end-user contributions, and post-incident forensics.[151] The ongoing development of agent IDs—which log instance-specific information such as the interacting system and interaction history—represent a practical foundation for model development, much like how a serial number is used to trace a product and its history.[152] These identifiers could also help track the origins, certifications, and performance of an AI system.[153] This increased visibility and improved agentic explainability would equip cyber practitioners, AI researchers and developers, and end-users with the dynamic intelligence needed to detect suspicious activity, make attributions to culpable threat actors, and conduct incident investigations.[154]

Responsible Design and Deployment Strategies for Developers and End Users

Maintain Strong Cyber Hygiene Best Practices



Cybersecurity fundamentals remain essential, but they must now extend into each layer of the agentic infrastructure stack. Core cyber best practices, such as robust identity and access management, secure API usage, and zero-trust

147. Alice Gomstyn and Alexandra Jonker, "New ethics risks courtesy of AI agents? Researchers are on the case," IBM, Dec. 23, 2024. https://www.ibm.com/think/insights/ai-agent-ethics.
148. Ibid.
149. Ibid.
150. Afek Shamir, "AI Agents: Why We Should Strategize on Governance," Tech Policy Press, Dec. 19, 2024. https://www.techpolicy.press/ai-agents-why-we-should-strategize-on-governance.
151. Ibid.
152. Ibid.
153. Ibid.
154. Ibid.

architectures should be implemented when designing new AI agents or adapting existing agents for customized applications.[155] These measures can help reduce the risk of cascading failures across the agent's workflow and maintain system integrity.[156] As agents operate with increasing autonomy, maintaining strong cyber hygiene best practices remains the first line of defense.

Implement Boundaries for Agentic Scope and Autonomy to Well-Defined Tasks

Before designing and deploying agents, organizations, developers, and end-users should clearly define and document the agent's intended scope, purpose, task parameters, and tiered permission levels.[157] This practice is imperative because it reduces the risk of unintended consequences, improves agentic reliability, and mitigates the potential of agentic drift, misalignment, and overreach.[158] Organizations should also update internal policies that outline acceptable agentic design and deployment strategies and train employees on how agents can be responsibly used for their individual roles and responsibilities.[159]

Deploy AI Agents Incrementally with Built-In Evaluation and Rollback Protocols

Responsible and secure deployment of AI agents requires iterative testing and continuous monitoring.[160] Organizations and end users should always aim to introduce agents gradually, starting with sandboxed environments and escalating through staged pilot programs, while applying regular red-teaming, running automated stress tests, and preparing rollback protocols at each phase of agentic deployment.[161] This allows developers and end users alike to continuously observe an AI agent's real-world behavior, identify surface novel risks in real-time, calibrate agent performance over time, and course correct effectively if an agent starts completing harmful tasks or expanding beyond its defined purpose and scope of work.[162]

Conclusion

The rise of AI agents signals a marked shift in how emerging technologies interact with, interpret, and influence our digital world. Increasingly described as the "third wave" of AI innovation, AI agents represent a departure from passive models that rely on continuous human oversight and intervention.[163] With the ability to act autonomously, reason, and learn through experience, AI agents are poised to redefine the contours of human–machine collaboration.

These agentic advancements also introduce complex governance questions and cybersecurity challenges. As AI agents take on more decision-making

155. Canon, "Why cyber hygiene remains critical in the era of AI-driven threats," CSO, Feb. 26, 2025. https://www.csoonline.com/article/3820782/why-cyber-hygiene-remains-critical-in-the-era-of-ai-driven-threats.html; "It May be Time to Review Your Cyber Hygiene," Security Magazine, May 8, 2024. https://www.securitymagazine.com/articles/100646-it-may-be-time-to-review-your-cyber-hygiene.
156. Korolov. https://www.cio.com/article/3489045/ai-agents-will-transform-business-processes-and-magnify-risks.html.
157. "Guidelines for defining goals and instructions for AI agent," Webex Help Center, Feb. 7, 2025. https://help.webex.com/en-us/article/nelkmxk/Guidelines-for-defining-goals-and-instructions-for-AI-agent.
158. Ibid.
159. Lynch. https://hai.stanford.edu/news/predictions-for-ai-in-2025-collaborative-agents-ai-skepticism-and-new-risks.
160. Xuhui Zhou et al., "HAICOSYSTEM: An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions," arXiv, Sept. 24, 2024. https://arxiv.org/abs/2409.16427.
161. Ibid.
162. Ibid.
163. Caldwell. https://www.forbes.com/councils/forbestechcouncil/2024/10/22/the-evolution-of-ai-agents-in-the-third-wave-of-ai.

roles and responsibilities, their actions may increasingly reflect the assumptions, priorities, and constraints embedded into their underlying models. Ensuring these systems are secure, reliable, and aligned with well-defined objectives requires more than technical measures alone; it demands coordinated efforts from stakeholders across industry, government, and the public. Our regulatory frameworks and risk management strategies must also evolve in parallel with AI agents to enable effective oversight, support responsible design and deployment, and reduce emerging risks.

While we cannot perfectly anticipate all the long-term impacts AI agents will have on how we work, learn, and live, what remains certain is that our governance and cybersecurity responsibilities are only beginning. With balanced, flexible safeguards in place, agentic systems can be deployed in ways that maximize their benefits while proactively mitigating the cybersecurity risks they may introduce or amplify. Beyond the recommendations outlined in this study, there are still many opportunities for future research and development, ranging from scalable audit mechanisms and real-time monitoring tools to infrastructure-agnostic safeguards and standards for agent-to-agent interactions.[164] Moreover, as AI agents are developed and deployed across borders, the need for voluntary and shared global governance norms for acceptable use and permissible development strategies will only grow.[165] Ultimately, the broader imperative is not simply to keep pace with emerging technologies but to guide and shape their trajectory, ensuring that they augment human talent and skills, reinforce America's technological leadership and economic competitiveness, and remain grounded in our founding values.[166]

Ultimately, the broader imperative is not simply to keep pace with emerging technologies but to guide and shape their trajectory, ensuring that they augment human talent and skills, reinforce America's technological leadership and economic competitiveness, and remain grounded in our founding values.

About the Authors

Haiman Wong is a resident fellow in the Cybersecurity and Emerging Threats team at the R Street Institute. Her research examines the intersection of cybersecurity and emerging technologies, including AI, edge computing, and connected vehicles.

Tiffany Saade is a master's candidate in the Stanford Ford Dorsey Program in International Policy, with a concentration in Cyber Policy and Security. Her research explores the intersection of AI and the evolving cybersecurity threat landscape, focusing on how AI models can be exploited in offensive cyber operations and strategies to secure them against misuse.

164. "The Rise of AI Agents: Transforming Citizen Development and User Experiences," Windows Forum, Feb. 27, 2025. https://windowsforum.com/threads/the-rise-of-ai-agents-transforming-citizen-development-and-user-experiences.353983/?amp=1.
165. Betsy Morris, "Beyond Intelligence: The Impact of AI Agents," Stanford University McCoy Family Center for Ethics in Society, Aug. 16, 2024. https://ethicsinsociety.stanford.edu/news/beyond-intelligence-impact-advanced-ai-agents.
166. Richard Reisman and Richard Whitt, "New Perspectives on AI Agentiality and Democracy: 'Whom Does It Serve?,'" Tech Policy Press, Dec. 6, 2024. https://www.techpolicy.press/new-perspectives-on-ai-agentiality-and-democracy-whom-does-it-serve; Eleonore Fournier-Tombs, "An Ethical Grey Zone: AI Agents in Political Deliberations," Carnegie Council for Ethics in International Affairs, Nov. 13, 2024. https://www.carnegiecouncil.org/media/article/ethical-grey-zone-ai-agents-political-deliberation.