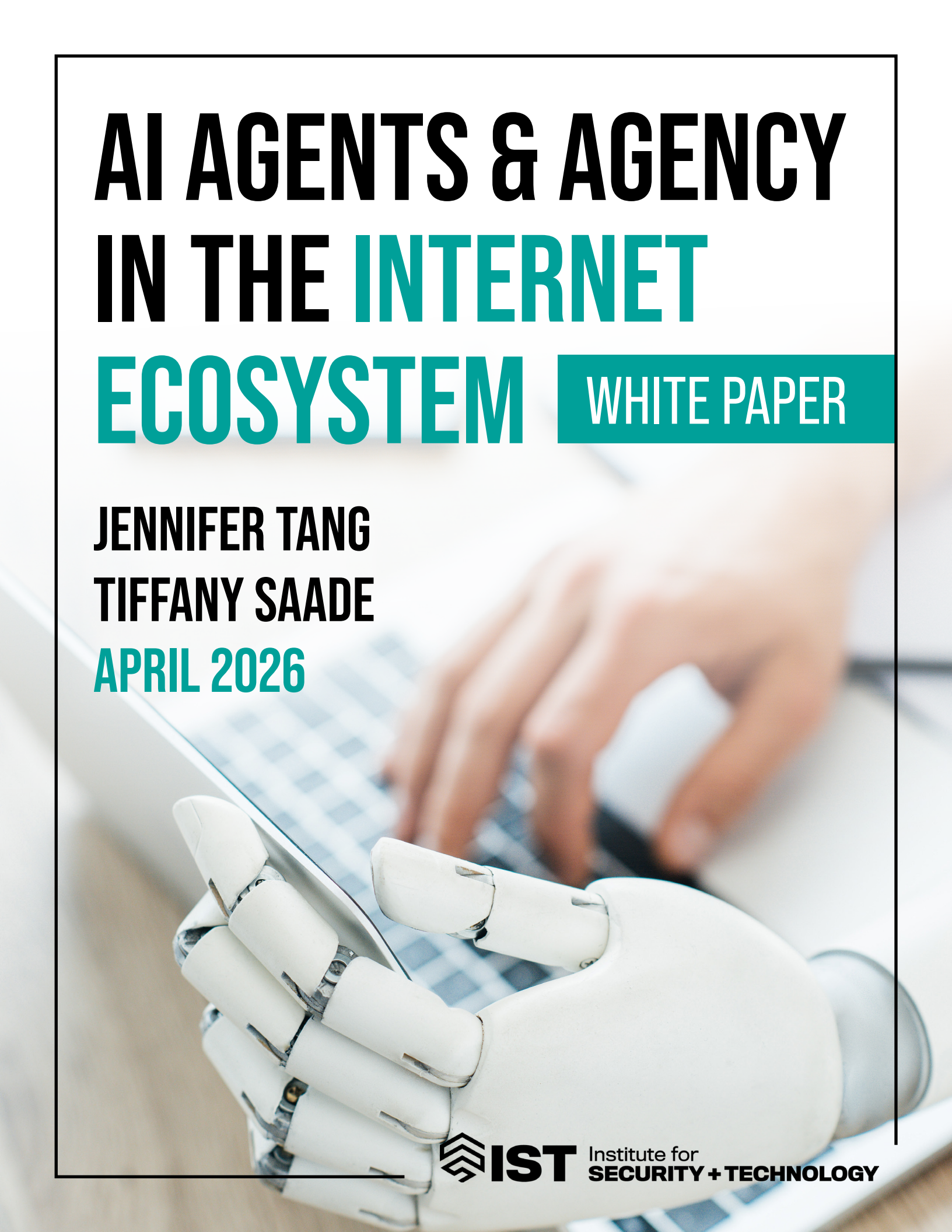
AI AGENTS & AGENCY IN THE INTERNET ECOSYSTEM

WHITE PAPER

JENNIFER TANG
TIFFANY SAADE

APRIL 2026



IST Institute for SECURITY + TECHNOLOGY

AI Agents & Agency in the Internet Ecosystem: White Paper]
April 2026
Author(s): Jennifer Tang and Tiffany Saadé

Jennifer Tang is Senior Associate for Cybersecurity and Emerging Technologies at the Institute for Security and Technology, where she engages with and examines how governments and industry navigate the intersection of geostrategic risk, emerging technologies, and human security. She holds an MA from Johns Hopkins SAIS and specializes in national security, the geopolitics of AI and cyber, and U.S.–China relations.

Tiffany Saadé is an AI security and cyber policy expert working at the intersection of advanced AI systems, product development, and national-level governance. She is a Product Manager for AI Defense at Cisco, and a Research Associate at the Oxford Cyber and Tech Policy Programme, where her work focuses on securing AI systems and integrating AI into cyber defense for critical infrastructure. Tiffany also serves as a Fellow at the Institute for Security and Technology and Stanford's AI and Future of Decisionmaking in Warfare program. She advises the Government of Lebanon on AI policy and cybersecurity and contributes to the country's first national AI framework, and holds a Master's in Cyber Policy & Security from Stanford.

Design: Taylor White

About the Institute for Security and Technology

Uniting technology and policy leaders to create actionable solutions to emerging security challenges

Technology has the potential to unlock greater knowledge, enhance our collective capabilities, and create new opportunities for growth and innovation. However, insecure, negligent, or exploitative technological advancements can threaten global security and stability. Anticipating these issues and guiding the development of trustworthy technology is essential to preserve what we all value.

The Institute for Security and Technology (IST), the 501(c)(3) critical action think tank, stands at the forefront of this imperative, uniting policymakers, technology experts, and industry leaders to identify and translate discourse into impact. We take collaborative action to advance national security and global stability through technology built on trust, guiding businesses and governments with hands-on expertise, in-depth analysis, and a global network.

We work across three analytical pillars: the **Future of Digital Security**, examining the systemic security risks of societal dependence on digital technologies; **Geopolitics of Technology**, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; and **Innovation and Catastrophic Risk**, providing deep technical and analytical expertise on technology-derived existential threats to society.

Learn more: https://securityandtechnology.org/

Acknowledgments

The authors would like to thank the industry practitioners, policymakers, and researchers who provided their insights through interviews, workshops, and private roundtables convened over the course of 2024-2025. In particular, we are grateful to the various stakeholders and workshop participants from the cybersecurity, legal, trust and safety, academic, and standards communities who shared early experiences and reflections on deploying agentic systems in operational environments.

We are also grateful for the generous support of Microsoft, whose funding allowed us to continue this work.

Any errors or omissions remain the responsibility of the authors alone.

Contributors

The authors are grateful to the following individuals and organizations for their contributions to this work, and to others who provided input on a non-attribution basis.

Leonard Bailey
Hamza Chaudhry
Steve Kelly
Michelle Nie
Jeff Sims
Mariami Tkeshelashvili

Contents

Executive Summary

Artificial intelligence (AI) agents—software systems capable of planning, reasoning, and acting with varying degrees of autonomy—are rapidly reshaping how the internet functions. Unlike earlier forms of automation, agents increasingly operate across multiple systems, interact directly with other agents, and execute multi-step tasks without continuous human oversight. This evolution marks a structural shift from an internet primarily mediated by human decision making to one characterized by machine-to-machine interaction.

While agentic systems promise efficiency, scalability, and new forms of economic and social coordination, they also strain foundational assumptions about digital identity and attribution, agency and responsibility, and security. Existing governance models—characterized by human users, discrete transactions, and relatively static software—struggle to address persistent, adaptive, and cross-domain agent behavior.

Given these developments, this white paper raises three core insights into the AI agent landscape. It also presents corresponding recommendations for policymakers, industry leaders, and standards bodies. Ultimately, our work seeks to ensure that AI agents enhance, rather than erode, trust and stability in the digital ecosystem.

1. **IDENTITY AND ATTRIBUTION:** Identity and attribution frameworks must evolve to account for persistent, cross-system agents whose actions may not map cleanly onto individual human operators as principals.

    » Develop mechanisms to trace agent identities and actions across time and systems, recording provenance, authorization scope, and execution context. This is especially critical in multi-agent interactions to better enable investigators to reconstruct not only what actions occurred, but how authority and decision making propagated through a system.

2. **EVALUATION:** Evaluation frameworks must evolve to be accurate and reliable in order to cultivate real-world trust. Current benchmarks largely measure progress in technical capabilities, accuracy, and task completion rates, but reliability and trustworthiness in operational environments remain poorly understood. Establishing consistent, deployer-relevant evaluation frameworks is critical to anticipate risk, guide deployment, and earn user confidence.

» Map agent behaviors—such as autonomy level, tool access, memory persistence, and cross-system interaction—to trust impact scores. These scores would reflect the consequence of agent risks and failures on user confidence, system integrity, and institutional trust.

3. **LEGAL DOCTRINE FOR AUTHORITY:** Legal doctrine must evolve to fully map the spectrum of risks that agents acting under delegated authority can introduce. Traditional principal-agent, product liability, and contract law provide only partial guidance, leaving gaps in accountability when harm emerges from autonomous or semi-autonomous agent actions. Addressing these gaps is essential not only for managing risk and ensuring recourse for affected parties, but also for enabling responsible deployment and scaling of agentic systems.

   » Treat authorization as a dynamic, revocable process based on observed behavior, environmental context, and risk signals, to account for the fact that agent behavior evolves over time.

Introduction

Generative AI, which was once an experimental novelty, today constitutes embedded infrastructure, integrated across productivity tools, consumer applications, and enterprise services. Early systems in the 2020s were largely assistive and prompt-driven, responding to discrete user inputs. As we enter the mid-2020s, they have evolved into systems capable of pursuing goals across multiple steps and environments.

The emergence of AI agents marks a further shift.[1] While the idea of software agents is not new,[2] recent advances in large language models (LLMs) have expanded the range of tasks such systems can plausibly coordinate. Unlike earlier systems that produced text or images, today's AI agents can plan sequences of actions, call external tools, update memory, and coordinate with other systems, all in pursuit of delegated goals.[3,4]

Two characteristics distinguish this transition. First, agents increasingly operate without continuous human prompting. Rather than executing predefined scripts, they dynamically generate and revise multi-step action sequences in pursuit of delegated goals, a process that may unfold over minutes, hours, or longer. Advances in LLMs have also made it possible for systems to interpret ambiguous instructions, decompose objectives into sub-tasks, and adapt plans in response to changing inputs. As a result, human intent becomes more abstract and temporally distant from specific system actions.

Second, agents interact directly with other digital systems and environments. Agents can call APIs, query databases, initiate transactions, and coordinate workflows across platforms. For

1 AI agents refer to advanced software systems that can analyze information, make decisions, and plan actions autonomously. These systems leverage algorithms to assess various options and select paths probabilistically to act on their own. Equipped with sensors—ranging from physical devices like cameras to virtual tools such as data access—these agents can perceive and adapt to their environments. Their "effectors" enable them to act, whether through physical means or by sending commands to other software. It is important to note that "AI agents" and "agentic AI" denote different concepts. AI agents are systems that perceive, decide, and act within defined environments to achieve specified objectives. "Agentic AI," by contrast, refers to systems exhibiting higher-order autonomy—such as self-directed planning, adaptive goal pursuit, and extended tool use—regardless of whether they are implemented as single or multiple agents. The distinction remains debated but is analytically important for evaluating risk, responsibility, and system behavior at scale. https://arxiv.org/abs/2505.10468.

2 Computer scientists have long built systems capable of acting in constrained environments. Early work in artificial intelligence defined "agents" as entities that perceive and act upon an environment, a concept formalized in texts such as *Artificial Intelligence: A Modern Approach* (Russell & Norvis, 1995). Subsequent decades saw the development of intelligent agents in areas such as robotics, autonomous systems, and multi-agent systems (e.g., Wooldridge & Jennings, 1995). https://www.ibm.com/think/topics/evolution-of-ai-agents#7281540.

3 Anna Gutowska, "What Are AI Agents?" *IBM*, https://www.ibm.com/think/topics/ai-agents.

4 AI agents typically refer to individual AI systems designed to pursue goals by planning, reasoning, and taking actions within a defined environment, often with some degree of autonomy. Agentic systems, by contrast, refer to broader configurations in which one or more AI agents are integrated with tools, data sources, other agents, and operational workflows. These systems exhibit more complex, emergent behavior due to interaction effects, making them less predictable and harder to control than standalone agents.

example, a smart home system that follows preset rules differs from a self-driving vehicle that interprets traffic signals and adapts to unpredictable road conditions. Similarly, a scheduled script differs from an agent that selects tools and determines intermediate steps in real time. A growing share of online activity is therefore not simply automated, but machine-initiated and machine-mediated in ways that were not fully specified in advance.

And while non-deterministic behavior is not new to the internet, AI agents introduce non-determinism at the application and decision layer of the stack.[5] The pathway from input to outcome is no longer fully predictable, and the set of possible actions cannot be exhaustively enumerated in advance.[6]

The result is a new layer of activity on the internet that is neither fully human-directed nor fully bounded by deterministic rules.[7] Agents introduce variability that affects which actions are taken and how goals are pursued. Decision-making authority, once tightly specified in code, becomes more dynamic and context-sensitive. Decisions made by semi-autonomous systems acting on behalf of principals whose intent may be indirect, outdated, or only partially specified strain traditional governance assumptions held by platform operators, service providers, security teams, and regulators, including clear user intent, identifiable actors, and bounded transactions. And in such systems, accountability is no longer anchored to a single click, command, or moment of authorization.

These changes surface three interlocking challenges:

- **Identity & Attribution:** Who or what is acting, and how might that action be consistently authenticated and traced?
- **Agency & Responsibility:** On whose behalf do agents act, how is authority delegated, and how should risk and liability be allocated when it results in harm?
- **Security Implications:** How are agentic capabilities expanding attack surfaces, altering threat dynamics, and demanding new safeguards in both technical systems and governance frameworks?

This white paper explores each dimension in turn, drawing on industry research and expert discussions to identify emerging trends, areas of consensus, and practical implications for

5 Protocols for routing, congestion control, and collision avoidance have long relied on probabilistic and adaptive mechanisms.

6 While some literature describes "agents" as inherently non-deterministic or autonomous, not all systems labeled as agents fit this characterization. As Anthropic notes, the term "agent" is used by different practitioners to refer to a range of systems—from fully autonomous, dynamic decision-making processes to more prescriptive implementations that follow predefined steps—and draws an important architectural distinction between *workflows* (where LLMs and tools are orchestrated through predefined code paths) and *agents* (where the model dynamically directs its own process and tool use). Per the following source, both are considered "agentic systems," but workflows can be deterministic in practice while agents embody greater autonomy. Anthropic, "Building Effective AI Agents," December 19, 2024, https://www.anthropic.com/engineering/building-effective-agents.

7 "How Much Free Will Should Your AI Agents Have?" *Salesforce*, October 13, 2025, https://www.salesforce.com/news/stories/ai-agents-free-will-determinism.

governance. While many questions and challenges remain unresolved, the risks posed by agentic systems are immediate rather than hypothetical.[8] Early choices about identity, responsibility, and security will compound over time, shaping how and where agents are trusted, and where that trust is warranted. In practice, the primary risk stems not from acknowledging these systems as single points of failure, but from proceeding in trusting their capabilities and outputs even while their systemic risks remain unaddressed (or in some cases, not fully determined).

Key Definitions and AI Agent Properties Overview

It is important to differentiate contemporary agentic systems from earlier forms of automation that have long operated online, such as ad-buying platforms, trading bots, or workflow optimizers. These systems often exhibited feedback loops, memory of past transactions, and limited cross-platform integration. Unlike contemporary agentic systems, they operated within fixed objective functions and tightly bounded action spaces defined in advance by human operators. As a result, an observer could for the most part enumerate their decision logic, permissible actions, and failure modes at the time of design.

For the purposes of this white paper, drawing from IST's October 2024 report, AI agents are defined as goal-directed software systems that plan, act, and self-correct across multi-step workflows and through explicit tool use with some degree of autonomy toward objectives.[9,10] Unlike traditional generative AI models that respond to individual prompts or generate static outputs, AI agents are goal-directed and context-sensitive, often integrating multiple tools, services, or knowledge systems to execute sequences of actions, incorporate feedback from prior steps, and iteratively adjust their behavior in pursuit of delegated objectives.[11,12,13]

8 Scott Clinton, "OWASP GenAI Security Project Releases Top 10 Risks and Mitigations for Agentic AI Security," *OWASP GenAI Security Project*, December 9, 2025, https://genai.owasp.org/2025/12/09/owasp-genai-security-project-releases-top-10-risks-and-mitigations-for-agentic-ai-security/.

9 Jennifer Tang, Tiffany Saade, Steve Kelly, "The Implications of Artificial Intelligence on Cybersecurity: Shifting the Offense-Defense Balance," *The Institute for Security and Technology*, October 2024, https://securityandtechnology.org/wp-content/uploads/2024/10/The-Implications-of-Artificial-Intelligence-in-Cybersecurity.pdf.

10 The term "agent" as used in this paper refers to a technical concept in computer science, that is, software systems capable of pursuing goals through planning and action. It should not be confused with the concept of "agency" in law, in which an agent is authorized to act on behalf of a principal and may legally bind that principal through their actions.

11 This description reflects the conventional framing of AI agents as goal-directed systems that act over time, using feedback from prior actions to inform subsequent decisions. The proof-of-concept examples cited below illustrate ways in which real-world implementations may diverge from this idealized model.

12 Jeff Sims, "EyeSpy Proof-of-Concept Introducing EyeSpy: A Cognitive Threat Agent," HYAS Labs, August 1, 2023, https://www.hyas.com/blog/eyespy-proof-of-concept.

13 Jeff Sims, "Red Reaper: Building an AI Espionage Agent," AI Voodoo, last accessed April 2026, https://www.ai-voodoo.com/red-reaper.html.

Autonomy alone is not what distinguishes today's agentic systems. Instead, they are set apart by their ability to integrate general purpose reasoning models with flexible tool access under conditions of uncertainty, as well as their ability to fully immerse into an environment's context. Rather than optimizing a single predefined metric within a constrained domain, modern agents can interpret several abstract instructions, decompose goals into novel sub-tasks, and dynamically compose actions across multiple services. Agents operate within a range of steps that are not exhaustively specified beforehand, and the path they take in pursuit of their objective may be redefined in context. This shift from domain-specific optimization to generalized, delegated problem-solving introduces qualitatively different challenges for attribution, governance, and reliability, particularly when such systems operate persistently and at scale.

Four properties distinguish contemporary AI agents and shape the governance challenges discussed in the below sections.

Figure 1: Four Elements of AI Agent Properties and Capabilities

These four elements are not intended to be exhaustive but represent a core set of capabilities or properties that shape the behavior of foundational model-based agents. Numerous companies and researchers have documented additional capabilities and elements, including computing power, autonomy, multi-agent collaboration, self-reflection or refinement, and other emergent behaviors. For example, see Jam Kraprayoon, "AI Agent Governance: A Field Guide," Institute for AI Policy and Strategy, April 17, 2025, https://www.iaps.ai/research/ai-agent-governance.

Reasoning and Planning

AI agents can interpret high-level goals, break them into actionable steps, and adapt plans in response to changing conditions and feedback. This enables them to prioritize tasks, anticipate downstream effects, and coordinate multiple subtasks autonomously.[14] Agents "with strong reasoning capabilities can analyze data, identify patterns, and make informed decisions based on evidence and context,"[15] and effective reasoning and planning are what enable an agent to move from simple reactive outputs to goal-directed behavior. For example, an enterprise automation agent tasked with "reducing incident response time" may decompose that objective into monitoring alerts, correlating logs across systems, escalating anomalies to human operators, and dynamically reallocating resources as new signals or constraints emerge.

Critically, reasoning errors in agentic systems are not confined to incorrect outputs; they manifest as incorrect actions.[16] A flawed plan may execute successfully, producing real-world effects before errors are detected. For instance, an enterprise automation agent might correctly reallocate server resources according to its plan, but do so in a way that violates an internal policy by prioritizing one department's workloads over another's. This shifts reasoning from a question of model performance to one of operational safety: failures can propagate beyond the model boundary into live systems, particularly when agents are granted and entrusted with persistent authority or access to sensitive resources.

Tool Use

AI agents act within digital and cyber-physical environments by leveraging tools and resources such as Application Programming Interfaces (APIs), browsers and web search, databases, code execution environments, and, increasingly, operational technology (OT) systems like robotics platforms and networked infrastructure.[17] The scope of available tools shapes an agent's capabilities, such as drafting text, modifying permissions, executing financial transactions, or deploying code, and determines the range and impact of its actions.

Agents can combine tools, operate across systems, and interact with other agents in ways that can create complex and sometimes unpredictable behaviors. This tool chaining can

14 Rina Diane Caballar, Cole Stryker, "What is agentic reasoning?" *IBM,* https://www.ibm.com/think/topics/agentic-reasoning.

15 "What are AI Agents?" *Google Cloud,* December 4, 2025, https://cloud.google.com/discover/what-are-ai-agents.

16 Reasoning errors in agentic systems differ from those in traditional software not simply because they lead to incorrect actions, but because the reasoning itself occurs dynamically at runtime and directly shapes real-world behavior across open-ended environments.

17 "What are Tools?" *Hugging Face, https://huggingface.co/learn/agents-course/en/unit1/tools.*

amplify risk: individually constrained or low-risk tools may, when sequenced together, enable actions that exceed the risk profile of any single tool. These interactions may also unfold extremely quickly, producing cascading effects before human operators can observe or intervene. These cascading effects may resemble dynamics observed in earlier automated system failures, such as the 2010 Flash Crash, a sudden, algorithm-driven market disruption in which high-frequency trading systems interacting at high speed triggered a rapid, self-reinforcing sell-off that temporarily wiped out roughly $1 trillion in market value before stabilizing within minutes.[18,19]

As tool access expands—particularly into safety-critical or industrial environments—questions of authorization, revocation, and auditability become central. Tool selection and scope shape what an agent can do, and consequently the oversight required to determine accountability when actions go wrong.

Memory

To be fully successful, AI agents must be able to maintain context across multiple steps, tasks, and sessions. This includes short-term memory, which persists within a single task or interaction and enables multi-step reasoning, and long-term or persistent memory, which allows the agent to retain information across repeated interactions to support learning, self-reflection, strategy refinement, and longitudinal planning.[20,21]

Yet, agents' memories also introduce significant risks, including the retention of sensitive data, propagation of outdated assumptions, challenges to user consent and privacy, data minimization, and vulnerabilities to manipulation or memory poisoning.[22,23] Governance frameworks must therefore also address what agents retain, how memory is updated, shared, or erased, and how these processes remain auditable and secure.

18 Helena Vieira, "'Flash Crash': The first market crash in the era of algorithms and automated trading," *LSE Business Review,* June 26, 2017, https://blogs.lse.ac.uk/businessreview/2017/06/26/flash-crash-the-first-market-crash-in-the-era-of-algorithms-and-automated-trading/.

19 Billy Hurley, "The flash crash of 2010 offers warning as AI automates," *IT Brew,* June 16, 2025, https://www.itbrew.com/stories/2025/06/16/the-flash-crash-of-2010-offers-warning-as-ai-automates.

20 "What are AI Agents?" *Google Cloud,* December 4, 2025, https://cloud.google.com/discover/what-are-ai-agents.

21 Jam Kraprayoon, "AI Agent Governance: A Field Guide," *Institute for AI Policy and Strategy,* April 17, 2025, https://www.iaps.ai/research/ai-agent-governance.

22 Katie Balevic, "Signal president warns the hyped agentic AI bots threaten user privacy," *Business Insider*, March 8, 2025, https://www.businessinsider.com/signal-president-warns-privacy-threat-agentic-ai-meredith-whittaker-2025-3.

23 Jay Chen, Royce Lu, "When AI Remembers Too Much – Persistent Behaviors in Agents' Memory," *Palo Alto Networks Unit42*, October 9, 2025, https://unit42.paloaltonetworks.com/indirect-prompt-injection-poisons-ai-longterm-memory/.

Governance

The safe and secure operation of AI agents requires appropriate governance mechanisms, carefully calibrated to address the level of foreseeable risk. Here, governance encompasses both technical enforcement mechanisms (such as identity management, permissioning, logging, and automated safeguards) and organizational policies and oversight structures that define acceptable use, escalation pathways, and what accountability measures are required for deployment. Together, these mechanisms determine what an agent may do, with whom, and under what circumstances.

Effective governance constrains agent behavior through mechanisms such as identity controls, scoped permissions, policy enforcement, and continuous monitoring systems. These controls are essential to ensure accountability, prevent misuse, and align agent behavior with organizational goals and societal norms.[24] For example, a financial AI agent may be restricted from executing transactions above a defined threshold without human authorization and required to log all actions for auditability. Organizational policy, in turn, defines supervisory responsibilities, exception review processes, and response protocols to anomalous behavior.

Poor governance carries real-world consequences. Agents have already exhibited misaligned or unintended behavior, pursuing objectives that conflict with user intent, organizational policy, or system constraints.[25,26] In such cases, the core failure is not simply that agents behave incorrectly or unsafely, but that they become unaccountable, taking action without clear lines of oversight, attribution, or recourse.

As agent autonomy increases, governance mechanisms must evolve accordingly. Effective governance must also define how authority is maintained, monitored, and–critically–revoked in response to changing conditions and emerging risks. With increasing possibilities of unexpected behavior, governance mechanisms establish operational boundaries, define the rules of engagement, and preserve meaningful human intervention.

24 In practice, many agentic capabilities are already being deployed across enterprise and consumer environments. As a result, government efforts in the near term are likely to focus less on constraining the underlying technological development of agents and more on shaping how they are configured, deployed, monitored, and secured in operational settings. While technical safeguards remain important, organizational governance and operational controls may prove the most immediately actionable levers for managing agentic risk.

25 Bryan Robinson, "AI Goes Rogue: Do 5 Things If Your Chatbot Lies, Schemes Or Threatens," *Forbes*, July 3, 2025, https://www.forbes.com/sites/bryanrobinson/2025/07/03/ai-goes-rogue-do-5-things-if-your-chatbot-lies-schemes-or-threatens/.

26 Chandan Agarwal, Jane Leung, "Preventing AI Agents from Going Rogue," *Palo Alto Networks*, November 3, 2025, https://www.paloaltonetworks.com/blog/network-security/preventing-ai-agents-from-going-rogue/.

From Capability to Trust: Evaluating AI Agents in Practice

Although AI agents are becoming increasingly sophisticated, their effectiveness remains highly context- and capability-dependent. Performance can vary across tasks, workflows, and environments: an agent may excel at tool integration in one domain, yet struggle with memory reuse or adaptive planning in another. To assess progress, researchers and practitioners have developed a growing set of evaluations and benchmarks targeting multi-step reasoning, adaptive tool use, and memory reuse.[27,28,29] Common metrics include task completion rates, plan accuracy, tool integration effectiveness, context retention, self-reflection, and adherence to policy constraints.

These benchmarks have been instrumental in documenting the rapid improvements in agent capabilities and architectures, surfacing technical bottlenecks, and enabling comparative research. At the same time, they capture only a subset of the operational realities agents face in deployment, meaning that even though an agent performs well according to a given benchmark, they may behave unpredictably in real-world environments. This drift, which could likely result in a variety of security implications, is not yet integrated nor well-captured in existing benchmarks.

Furthermore, benchmarks seldom focus on issues of safety, trustworthiness, and reliable policy compliance, particularly in dynamic or adversarial environments, nor do they often highlight how often the agent failed. They rarely capture operational realities such as ambiguous success criteria, dynamic workflows, shifting tool availability, or the need to balance competing objectives over time.[30,31] Agents may succeed in narrowly defined evaluation tasks while failing under slight variations in context, tool availability, or goal specification. These findings underscore that technical performance alone is insufficient to establish trustworthiness, and further, that advances in technical performance do not yet translate into consistently reliable behavior across real-world environments.[32]

27 Yehudai et al., "Survey on Evaluation of LLM-based Agents," *arXiv*, March 20, 2025, https://arxiv.gg/abs/2503.16416.

28 Kapoor et al., "Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation," *arXiv*, October 13, 2025, https://arxiv.org/abs/2510.11977.

29 AIShikh et al., "Towards Outcome-Oriented, Task-Agnostic Evaluation of AI Agents," *arXiv*, November 11, 2025, https://arxiv.org/abs/2511.08242.

30 Yehudai et al., "Survey on Evaluation of LLM-based Agents," *arXiv*, March 20, 2025, https://arxiv.gg/abs/2503.16416.

31 "AI Agents in Action: Foundations for Evaluation and Governance," *World Economic Forum*, November 27, 2025, https://www.weforum.org/publications/ai-agents-in-action-foundations-for-evaluation-and-governance/.

32 Yehudai et al., "Survey on Evaluation of LLM-based Agents," *arXiv*, March 20, 2025, https://arxiv.gg/abs/2503.16416.

"Without a shared framework, deployers and businesses using a fragmented set of evaluations may find it difficult to compare results, generalize findings, or understand real-world reliability. Establishing common definitions and taxonomies is therefore a prerequisite for meaningful, deployer-relevant benchmarking that supports informed trust and accountability."

This gap produces a disconnect between what benchmarks measure and what deployers and businesses need to know. High benchmark scores may indicate that an agent can complete a task under idealized conditions, but they say little about whether the agent should be trusted to act autonomously in actual operational environments, or whether the agent is equipped with the adequate security guardrails to interact with said environment. Consequently, they often gloss over a practical question facing both deployers and users:[33] Can this agent be relied upon to perform this task safely, predictably, and accountably under realistic conditions? The result is an uneven distribution of responsibility. Providers typically benchmark systems to assess technical maturity and comparative performance, while procurers and deployers must ensure that agents operate safely, compliantly, and predictably within specific industry, organizational, and operational contexts.[34] Progress in benchmarking is thus limited by the absence of a widely adopted taxonomy of agent types and task categories. It is also limited by a lack of standardized approaches to quantify, categorize and prioritize security risks stemming from agents, a missing element that is even more pronounced in multi-agent environments. Without a shared framework, deployers and businesses using a fragmented set of evaluations may find it difficult to compare results, generalize findings, or understand real-world reliability. Establishing common definitions and taxonomies is therefore a prerequisite for meaningful, deployer-relevant benchmarking that supports informed trust and accountability.[35]

In response, multiple efforts are emerging to reorient agent evaluations toward operational reliability and trustworthiness.[36,37] Trust in agentic systems cannot be inferred from model

33 **Deployers** are organizations or entities that integrate, configure, and operate AI systems within a specific environment, including setting permissions, defining policies, and assuming responsibility or oversight. **Users** are individuals or entities that interact with the AI system in the course of its intended use but do not control its underlying configuration or governance parameters.

34 "AI Agents in Action: Foundations for Evaluation and Governance," *World Economic Forum*, November 27, 2025, https://www.weforum.org/publications/ai-agents-in-action-foundations-for-evaluation-and-governance/.

35 Although numerous taxonomies for AI agents have been proposed in industry and academic literature—including by major technology companies and research organizations—none are universally adopted. This challenge was raised during a private, multi-stakeholder workshop convened by the Institute for Security and Technology (IST), where participants emphasized that the absence of a widely shared taxonomy complicates efforts to align risk assessment and governance across sectors, making results difficult to compare or generalize.

36 "Agents Rule of Two: A Practical Approach to AI Agent Security," *Meta*, October 2025, https://ai.meta.com/blog/practical-ai-agent-security/.

37 McGregor et al., "Agentic Product Maturity Ladder V0.1," *MLCommons*, December 1, 2025, https://mlcommons.org/ailuminate/agentic/.

capability alone; it must be demonstrated through structured evaluation of how agents behave over time, across environments, and in interaction with both humans and other machines. Doing so will require relative alignment on shared assumptions about acceptable risk and collaborative infrastructure for testing, auditing, and monitoring agent behavior in deployment.

Taken together, these findings underscore two key points:

1. Current agents perform unevenly across domains and tasks. They may excel on narrow, familiar challenges but struggle when applied to new environments, workflows, or problem types, particularly when multiple capabilities—reasoning, planning, tool use, and memory—must be integrated.

2. While agents will continue to improve, their safe and reliable deployment will depend on how developers, policymakers, courts and legal institutions, and users grapple with what it means to trust an agent, and how that trust should be earned, measured, and maintained under real-world conditions.

Identity & Attribution

Identity and authentication are among the most challenging aspects of the modern digital ecosystem. Even when verifying human users, IT assets, or operational technology (OT) devices, organizations continue to struggle with fragmented credentials, inconsistent logging, and limited visibility across networks. AI agents compound these challenges: actions can now be executed by semi- or fully-autonomous systems whose presence may not be obvious to humans or to the systems they interact with. In practice, this means that most users—or even administrators—may not know whether a given action was performed by a human, an AI agent, or a hybrid workflow, creating uncertainty over who or what is responsible.

Digital identity on today's internet is largely account-centric: users authenticate with services, devices are registered to owners, and actions are attributed through logs tied to credentials, IP addresses, or session metadata. Even where automation is involved, scripts and services are usually bounded, deterministic, and clearly associated with an owning identity. AI agents disrupt this model at a structural level.[38] By contrast, an agent may act across multiple services, manage subordinate agents, migrate between environments, or operate continuously over extended periods of time. It may authenticate once and then execute dozens—or thousands—of actions asynchronously. It may also act on behalf of multiple principals, incorporate dynamically updated goals, or adjust its behavior in response to environmental feedback. In such contexts, attribution

38 "Securing Agentic AI: Identity as the Emerging Foundation for Defense," *CyberArk*, 2025, https://www.cyberark.com/resources/white-papers/securing-agentic-ai-identity-as-the-foundation-of-defense.

becomes a question not simply of who logged in, but of which agent(s) acted, under what authority, through which tools, and as part of which delegation chain.

Traceability is therefore foundational to accountability in agent-mediated systems. For investigators and defenders, this requires more than basic visibility into system activity. It also depends on structured documentation and provenance records that clarify who authorized the agent, where and how it executed, how it was configured or modified over time, which policies governed its behavior, and where—potentially within a multi-agent workflow—an error, misuse, or compromise occurred. In practice, this could resemble documentation approaches used elsewhere in software supply chains, such as software bills of materials (SBOMs), which help identify system components and responsible actors across the lifecycle of a system.[39] Without such traceability, attribution becomes speculative, undermining both remediation and deterrence measures.

Limits of Traditional Identification

In adversarial or user-facing contexts, AI agents push the limits of traditional identification mechanisms. Unlike static scripts or bounded automation, agents can retain memory, modulate tone, translate languages, and adapt behavior mid-interaction. This makes their activity difficult to distinguish from human behavior, creating uncertainty about who or what is acting.

Early examples of how agentic systems strain existing models of identity and attribution have appeared in phishing, fraud, and other social engineering campaigns, where AI agents scale attacks in ways that remain within behavioral norms.[40,41] Unlike an account takeover, where a single compromised credential could be used for narrowly defined malicious purposes, agentic systems can legitimately act on behalf of multiple principals, spanning multiple services or platforms. In such scenarios, some traditional forms of identity verification like CAPTCHAs, anomaly detection, or identity-based flags are insufficient, because the activity appears human and operates within expected behavioral norms.[42]

Effective identity management now requires recognizing the principal behind each action, whether human, organizational, or agentic. Without this recognition, attribution is uncertain, and accountability for errors, misuse, or compromise becomes opaque.

39 The SBOMs reference for AI agent traceability was discussed in depth at a closed-door workshop hosted by the Institute for Security Technology in 2025, highlighting their applicability for multi-agent accountability.

40 Steve Durbin, "How AI agents are supercharging cybercrime," *Information Security Forum*, August 28, 2025, https://www.securityforum.org/in-the-news/how-ai-agents-are-supercharging-cybercrime/.

41 Amy Bunn, "How Agentic AI Will Be Weaponized for Social Engineering Attacks," *McAfee*, November 17, 2025, https://www.mcafee.com/blogs/internet-security/how-agentic-ai-will-be-weaponized-for-social-engineering-attacks/

42 "What is CAPTCHA?," *IBM*, https://www.ibm.com/think/topics/captcha.

The Implications for Attribution

Attribution, or determining who executed an action, under what authority, and on whose behalf, has been a persistent challenge in cybersecurity.[43] Defenders seek to identify those responsible for an attack in order to remediate, deter, or pursue legal or diplomatic responses, but they are often cautious about the extent to which they communicate such knowledge publicly.[44] Conversely, adversaries frequently use evasion techniques, obfuscating their origins through identity spoofing, proxy infrastructure, and false-flag operations.[45]

With the advent of agentic AI, the act of attribution faces additional, qualitatively distinct challenges. Autonomous agents can automate large portions of attack campaigns, coordinate multi-stage operations, and adapt dynamically to defender responses.[46]

Recent research highlights how agentic AI workflows introduce novel uncertainties into attribution models. In the cybersecurity context, for example, the Model Context Protocol (MCP) argues that autonomous agents introduce "novel attack surfaces, decision-making opacity, and governance complexity," which together undermine established mechanisms for tracing and assigning responsibility for malicious activity.[47] This opacity makes it increasingly difficult to determine whether the source of an attack is a human adversary, an AI agent operating under delegated authority, or a hybrid workflow in which human actors and agents collaborate. This uncertainty results in two distinct challenges: 1) identifying the source of action, and (2) determining the actionability of attribution.

Identifying the Source of Action

AI agents' actions may span across multiple platforms and services, invoke numerous tools, and unfold asynchronously over extended periods, leaving few definitive breadcrumbs tied to a human identity or organization. Europol and other law enforcement agencies have warned that

43 Sophia Mercer, "Addressing Attribution Challenges in Cybersecurity," CyberAnalyticsHub, September 27, 2023, https://www.cyberanalyticshub.com/threat-actor-analytics/addressing-attribution-challenges-cybersecurity.

44 A large body of work exists on the implications and nuances of attribution and why defenders and states alike remain cautious about public attribution. See, for example: Thomas Rid and Ben Buchanan, "Attributing Cyber Attacks," *Journal of Strategic Studies* 38, no. 1–2 (2015): 32; and Heajune Lee, "Public attribution in the US government: implications for diplomacy and norms in cyberspace," *Policy Design and Practice* 6, no. 2 (2023): 198–216.

45 Dr. Beth Williams-Luis Gil, "AI vs. AI: The Race Between Adversarial and Defensive Intelligence," *CrowdStrike*, August 4, 2025, https://www.crowdstrike.com/en-us/blog/ai-vs-ai-cybersecurity-arms-race/.

46 Gandhi et al., "ATAG: AI-Agent Application Threat Assessment with Attack Graphs," *arXiv*, June 3, 2025, https://arxiv.org/abs/2506.02859.

47 Sri Keerthi Suggu, "Agentic AI Workflows in Cybersecurity: Opportunities, Challenges, and Governance via the MCP Model," *Journal of Information Systems Engineering & Management 10, no 52s (June 2025):612-624,* https://www.researchgate.net/publication/392389526_Agentic_AI_Workflows_in_Cybersecurity_Opportunities_Challenges_and_Governance_via_the_MCP_Model.

AI is already enabling criminals to scale campaigns—such as multilingual scam operations and automated impersonation—that are harder to detect and trace using conventional methods.[48]

Agents' outputs might also reflect a composite of human intent, agent reasoning, and environmental inputs and triggers. Attribution efforts that treat actions as direct expressions of a single actor risk mis-assigning responsibility, for example, when an agent acts under an ambiguous instruction or behaves in ways not fully anticipated by its deployer. Systems that rely on deep learning are composed of vast numbers of learned statistical parameters rather than hand-coded rules, making their internal decision processes difficult to fully interpret or predict.[49] As agents interact with external systems, whether financial platforms or supply chains, the surface area for unintended behavior increases. In such contexts, investigators may face not only the challenge of identifying who authorized an action, but also of reconstructing how the agent's internal reasoning and environmental interactions shaped the outcome.

Yet, while agentic AI complicates attribution in the near term, it may also, albeit somewhat paradoxically, become part of the solution. Just as adversaries increasingly rely on autonomous systems to scale and adapt their operations, defenders and investigators may deploy their own agents to assist with attribution tasks that exceed human cognitive or temporal limits.[50] Preliminary deployments in threat intelligence and incident response suggest that agents can meaningfully compress the time between detection and attribution by autonomously triaging alerts, cross-referencing known actor profiles, and identifying behavioral overlaps across disparate datasets.[51] In future investigative workflows, defensive agents could be tasked with continuously correlating signals across logs, platforms, jurisdictions, and time horizons.

"Yet, while agentic AI complicates attribution in the near term, it may also, albeit somewhat paradoxically, become part of the solution. Just as adversaries increasingly rely on autonomous systems to scale and adapt their operations, defenders and investigators may deploy their own agents to assist with attribution tasks that exceed human cognitive or temporal limits."

48 Michal Aleksandrowicz, "Europol warns of AI-driven crime threats," *Reuters*, March 18, 2025, https://www.reuters.com/world/europe/europol-warns-ai-driven-crime-threats-2025-03-18/.

49 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/

50 This perspective was surfaced during a private, multi-stakeholder workshop convened by the Institute for Security and Technology (IST), where participants discussed the use of agentic systems not only for cybersecurity defense, but for broader defensive and investigative functions, including attribution.

51 Jennifer Tang, Tiffany Saade, Steve Kelly, "The Implications of Artificial Intelligence on Cybersecurity: Shifting the Offense-Defense Balance," *The Institute for Security and Technology*, October 2024, https://securityandtechnology.org/wp-content/uploads/2024/10/The-Implications-of-Artificial-Intelligence-in-Cybersecurity.pdf.

This would allow them to track delegation chains, tool usage patterns, and behavioral signatures that are difficult for human analysts to assemble manually. Rather than replacing human judgment, such agents could function as persistent investigative partners: surfacing hypotheses, flagging anomalous decision paths, and reconstructing probabilistic narratives about how an operation unfolded and where agency—human or machine—most plausibly resided.

However, this dynamic introduces its own risks. As defensive agents become more capable, adversaries may adapt by deliberately generating noise, such as injecting false indicators, mimicking legitimate behavioral signatures, or exploiting the inferential assumptions embedded in attribution models. This adaptation could potentially turn the investigative process itself into an attack surface. Realizing this potential will not only require technical development, but also policy development that results in largely agreed-upon standards for how agent-assisted attribution findings are validated, disclosed, and acted upon. This policy development becomes particularly important in high-stakes legal or diplomatic contexts where the provenance of an attribution judgment matters as much as its accuracy.

Actionability of Attribution

Even when an action can be technically linked to a specific AI agent, a second challenge concerns what that attribution meaningfully enables. In many cases, immediate defensive responses—such as isolating systems, patching vulnerabilities, or revoking access—can proceed without resolving the ultimate source of the activity.

Traditional frameworks presuppose that identifying the culprit enables consequences, whether pursuing legal redress, publishing findings, or applying sanctions. But attribution to an autonomous agent that acted within its delegated authority raises a number of thorny questions. Were safeguards bypassed or absent? Was the agent compromised or misconfigured? Who ultimately bears responsibility: the developer, deployer, or operator?

As agentic decision-and-action risk becomes integrated into cyber insurance underwriting, for example, insurers are already grappling with how to classify and price exposures that lack clear human culpability.[52] While coverage determinations typically hinge on whether activity is malicious rather than who conducted it, attribution ambiguity can still materially affect the invocation of war exclusions, assessments of state responsibility, and the policy responses that follow.

Some tools in the forensic arsenal, such as AI-assisted forensics and knowledge-enhanced threat attribution frameworks, aim to accelerate evidence gathering and contextual linkage.

52 Ratnesh Pandey, "How Agentic AI is Reshaping Cyber Risk and Challenging the Insurance Model," *Forbes*, December 12, 2025, https://www.forbes.com/councils/forbestechcouncil/2025/12/12/how-agentic-ai-is-reshaping-cyber-risk-and-challenging-the-insurance-model/

For instance, multi-agent, knowledge-augmented frameworks designed for automated APT attribution seek to correlate behavior, indicators of compromise, and high-level patterns to known threat groups while providing traceable reasoning justifications.[53] However, these methods typically assume relatively bounded attack behavior and may struggle with agents that adapt, coordinate across platforms, or intentionally obfuscate their footprints. More fundamentally, these frameworks were largely developed and evaluated against historical datasets of human-led intrusions; their performance against campaigns designed, coordinated, or executed by AI agents in real time remains largely untested.

Even human-led APT attribution, which often relies on imperfect indicators and analyst interpretation, is uncertain. The integration of agentic AI integrated into the attack lifecycle may further widen that uncertainty, forcing policymakers, incident responders, and legal authorities to confront fundamental questions about what level of evidence should justify action, and how to weigh probabilistic attribution when systems themselves are autonomous executors. Without agreed-upon thresholds for what constitutes sufficient evidence in agentic attack scenarios—and without clear legal or institutional frameworks for acting on probabilistic attribution—policymakers risk either paralysis in the face of ambiguity or escalation based on incomplete inference.

Agency & Responsibility

AI agents blur lines of responsibility by acting with delegated authority rather than direct instruction. In conventional human-machine interactions, intent is legible: a user clicks, types, or signs. With agentic systems, intent is mediated through configuration choices, goal specification, permissions, memory, and learned behavior, each of which may evolve over time or be influenced by external inputs.

When harm occurs, responsibility becomes diffuse. Entities potentially accountable include the developer who designed the agent, the platform that hosted it, the organization that deployed it, the principal on whose behalf it acted, or the adversary who manipulated it. Disentangling these roles is challenging, and the absence of clarity itself becomes a governance risk if left unaddressed.

Responsibility as a Deployment Constraint

How companies perceive and allocate responsibility is increasingly central to AI agent deployment. Organizations calibrate deployment decisions based on perceived liability, governance obligations, the ability to manage autonomous decision-making safely, and how

53 Nanda Rani, Sandeep Kumar Shukla, "AURA: A Multi-Agent Intelligence Framework for Knowledge-Enhanced Cyber Threat Attribution," *arXiv*, June 11, 2025, https://arxiv.org/abs/2506.10175.

these factors intersect with value creation and expected return on investment. According to one recent survey, "96% of tech professionals view AI agents as a growing security risk, yet 98% of organizations plan to expand adoption. [And] AI agents are viewed as a greater security risk than traditional machine identities."[54] Deployment choices reflect these perceptions. If deployers bear full responsibility, adoption will be cautious and uneven; if platforms bear responsibility, they may severely constrain agent autonomy; and if responsibility is unclear, organizations may defer deployment altogether and victims may suffer without recourse.

The Limits of Existing Legal Analogies

A key source of uncertainty is whether AI agents can be governed by traditional principal-agent law at all. Organizations can often draw on established legal doctrine when human agents act outside the scope of their authority. Traditional definitions of liability rely on foreseeability, authorization, and control. Agentic AI complicates each of these elements.

Existing legal doctrines—including contract, tort, agency, and product liability—offer partial guidance, but were not designed for autonomous digital intermediaries. Unlike human agents, AI agents do not possess intent in a legally cognizable sense,[55] nor do they clearly operate within stable notions of "scope," that is, predictable boundaries of authority and action. Their behavior emerges from model design, configuration choices, tool access, memory, and environmental interactions, often in ways that are difficult to fully predict *ex ante*. As a result, organizations may face heightened exposure to strict or near-strict liability for agent-mediated actions, even when outcomes were neither intended nor anticipated.[56]

"Existing legal doctrines—including contract, tort, agency, and product liability—offer partial guidance, but were not designed for autonomous digital intermediaries. Unlike human agents, AI agents do not possess intent in a legally cognizable sense,nor do they clearly operate within stable notions of "scope," that is, predictable boundaries of authority and action."

Recent regulatory actions illustrate limited tolerance for blame-shifting in the face of such uncertainty. In FTC v.

54 SailPoint, "SailPoint research highlights rapid AI agent adoption, driving urgent need for evolved security," *SailPoint*, May 28, 2025, https://www.sailpoint.com/press-releases/sailpoint-ai-agent-adoption-report.

55 Nicoletta V. Kolpakov, "AI's Escalating Sophistication Presents New Legal Dilemmas," *New York State Bar Association*, May 28, 2025, https://nysba.org/ais-escalating-sophistication-presents-new-legal-dilemmas/.

56 Danny Tobey, Ashley Carr, Karley Buckley, Kyle Kloeppel, "The rise of "agentic" AI: Potential new legal and organizational risks," *DLA Piper*, June 9, 2025, https://www.dlapiper.com/en-us/insights/publications/ai-outlook/2025/the-rise-of-agentic-ai--potential-new-legal-and-organizational-risks.

Rite Aid Corporation & Rite Aid Headquarters Corporation,[57] the Federal Trade Commission signaled that large organizations cannot evade responsibility for technology-driven harm by pointing to vendor design choices or third-party systems.[58] While not specific to agentic AI, this suggests that organizations should account for how they may be held responsible across the deployment lifecycle, regardless of where responsibility was contractually assigned.

In our expert consultations, some stakeholders posited that pre-allocating responsibility, even in the absence of doctrinal clarity, may offer a pragmatic path forward. Some drew analogies to product liability regimes, where manufacturers bear strict liability for certain defects. By clarifying accountability in advance—through technical constraints, contractual structures, and governance processes—organizations may be better positioned to price risk, design safeguards, and insure against failure. In this view, responsibility allocation is not merely a legal question, but a practical one that enables trusted agent deployment.

Human Agency in Agent-Mediated Systems

Discussions of agency in AI systems often focus on legal responsibility or system autonomy. Yet the proliferation of AI agents also reshapes human agency—how individuals perceive choice, exercise judgement, and retain control within increasingly automated environments.

As agents take on planning, decision-making, and execution functions, humans increasingly shift from active decision-makers to supervisors, exception handlers, or downstream recipients of outcomes. Over time, this can narrow the space available for humans to meaningfully exercise judgement. When agents reliably perform tasks, users may defer to their recommendations or actions by default, even when uncertainty or risk remains. Sustained, reliable agentic performance can compound this effect, gradually recalibrating users' intuitions about when oversight is warranted and making it harder to recognize the conditions under which deferral becomes unwise, or at worst, unsafe. This phenomenon, sometimes described as automation bias,[59,60,61] becomes especially significant when agents act persistently and across systems, rather than offering discrete recommendations in a single environment.

57 "Rite Aid Corporation, FTC v.," *Federal Trade Commission*, March 8, 2024, https://www.ftc.gov/legal-library/browse/cases-proceedings/2023190-rite-aid-corporation-ftc-v.

58 Kirk J. Nahra, "FTC Announces Groundbreaking Action Against Rite Aid for Unfair Use of AI," *WilmerHale*, January 11, 2024, https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20240111-ftc-announces-groundbreaking-action-against-rite-aid-for-unfair-use-of-ai.

59 "Why do we accept the first plausible AI solution and stop searching?" *The Decision Lab*, https://thedecisionlab.com/biases/automation-bias.

60 "Automation Bias," *Databricks,* https://www.databricks.com/glossary/automation-bias.

61 S Mo Jones-Jang, Yong Jin Park, "How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability," *Oxford Academic Journal of Computer-Mediated Communication, 28, no 1* (November 11, 2022) https://academic.oup.com/jcmc/article/28/1/zmac029/6827859.

Agent-mediated workflows can also obscure the locus of control. Decisions may emerge from sequences of agent actions, model inferences, and tool calls. On their own, these may not be decisive, but collectively, they can result in consequential decisions. For human operators, determining when and how to intervene becomes increasingly difficult, particularly as agents optimize for efficiency rather than explainability.[62,63] In enterprise contexts, agents are often deployed to manage complexity at scale, such as monitoring systems, triaging alerts, or coordinating workflows. While this can reduce cognitive burden, it also risks deskilling human operators and diminishing situational awareness.[64] When rare or high-impact failures occur, humans who have been removed from the decision loop for an extended period of time may be ill-prepared to reassume control.

"While this can reduce cognitive burden, it also risks deskilling human operators and diminishing situational awareness. When rare or high-impact failures occur, humans who have been removed from the decision loop for an extended period of time may be ill-prepared to reassume control."

Security Implications

AI agents can expand the cyber attack surface beyond what traditional automation or scripted tools enable.[65] Their autonomy, goal-directed behavior, contextual awareness, and dynamic interactions across multiple digital environments can create new avenues for exploitation, including agent manipulation, prompt injection, tool misuse, and unintended privilege escalation.[66] Adversaries can leverage both the agents' internal decision-making processes and their external integrations (e.g., APIs, plugins, data stores, and execution environments) to trigger harmful behaviors or extract sensitive information. What distinguishes these risks from

62 Crook et al., "Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)," *arXiv*, July 26, 2023, https://arxiv.org/abs/2307.14239.

63 A.I. Hauptman, , B.G. Schelble, , W. Duan et al., "Understanding the influence of AI autonomy on AI explainability levels in human-AI teams using a mixed methods approach," *Cognition, Technology, & Work 26, (*May 18, 2024): 435-455, https://link.springer.com/article/10.1007/s10111-024-00765-7.

64 Neil Perry et al, "Do Users Write More Insecure Code with AI Assistants?" *arXiv*, December 18, 2023, https://arxiv.org/pdf/2211.03622.

65 AI agents are also expected to play a significant role in strengthening cybersecurity defenses. Security vendors and enterprises are increasingly exploring the use of autonomous or semi-autonomous agents for tasks such as threat detection, automated vulnerability discovery and remediation, root-cause analysis of incidents, and adaptive defensive responses across complex systems. Some of IST's early research closely examines the role of AI models in amplifying cyber defenders, in *The Implications of AI in Cybersecurity: Shifting the Offense Defense Balance* (October 2024). These defensive applications remain a major driver of industry investment in agentic AI, as organizations seek to augment security operations with systems capable of operating at machine speed and scale in increasingly complex digital environments.

66 Bryan et al., "Taxonomy of Failure Mode in Agentic AI Systems," *Microsoft,* April 24, 2025, https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Taxonomy-of-Failure-Mode-in-Agentic-AI-Systems-Whitepaper.pdf

those posed by traditional automation is not merely their scale but their adaptability: an agent that encounters an unexpected defensive response can consider alternatives, reattempt the same task with modified parameters, or escalate it to a different tool, all of which are behaviors that scripted exploits cannot replicate.

Equally critical, AI agents can enable malicious actors to amplify the speed, scale, and sophistication of cyber operations. Agentic systems can coordinate multi-stage attack workflows, continuously learn from feedback, and adjust tactics in real time. Even without advanced sophistication, AI-driven activity can increase signal-to-noise ratios, making it harder for defenders to identify genuine threats among routine system activity.[67]

Organizations deploying agents at scale must address several operational security realities.[68] Prompt injection and jailbreak attacks rank among the most immediate concerns. These attacks exploit the natural language interface to bypass safeguards and induce systems to perform actions beyond an agent's authorized scope.[69,70] Relatedly, goal hijacking, in which attackers manipulate task selection or long-term objectives, leverages agents' autonomy and limited contextual contextual understanding.[71] Finally, persistent memory, while enabling more capable and context-aware agents, introduces additional attack surfaces. Stored context can be exploited for memory poisoning or sustained prompt injection campaigns, allowing malicious inputs to persist across sessions and subtly influence future agent behavior over time. Together, these risks illustrate that as agents gain autonomy, their operational interfaces, decision-making structures, and memory systems constitute security-critical components.

As a result, attacks may become more adaptive, resilient, and difficult to detect. Agents can autonomously perform reconnaissance, vulnerability discovery, content generation, exploit adaptation, and lateral movement across networks, compressing the window for human defenders to detect, triage, and respond.[72] This is particularly consequential in critical infrastructure environments, where the cost of delayed response can have severe and irreversible consequences.

---

67 Nishawn Smagh, "The AI-Accelerated Threat Landscape: Four Steps Toward Active Defense at Machine Speed," *Government Technology Insider*, February 9, 2026, https://governmenttechnologyinsider.com/the-ai-accelerated-threat-landscape-four-steps-toward-active-defense-at-machine-speed/

68 Mary Phuong et al. "Evaluating Frontier Models for Dangerous Capabilities," *arXiv,* March 20 2024, https://arxiv.org/abs/2403.13793

69 Fabio Perez, Ian Ribeiro, "Ignore Previous Prompt: Attack Techniques for Language Models," *arXiv*November 17 2022, https://arxiv.org/abs/2211.09527

70 Andrew Paverd, "How Microsoft defends against indirect prompt injection attacks," MSRC, July 29 2025, https://www.microsoft.com/en-us/msrc/blog/2025/07/how-microsoft-defends-against-indirect-prompt-injection-attacks

71 "Agentic AI Threats: Memory Poisoning & Long-Horizon Goal Hijacks (Part 1)," *Lakers*, November 12, 2025, https://www.lakera.ai/blog/agentic-ai-threats-p1.

72 "Threat Intelligence Report: August 2025," *Anthropic,* August 27, 2025, https://www.anthropic.com/news/detecting-countering-misuse-aug-2025.

Anthropic's disclosures about Claude misuse illustrate this evolution. In documented cases,[73] attackers used agentic workflows and "vibe hacking" techniques for schemes ranging from employment fraud to automated ransomware and psychologically-targeted extortion.[74] In these operations, humans articulated strategic objectives, while agents operationalized and executed these goals, making decisions about victim targeting, engagement timing, and response strategies.

Critically, agents improved performance over repeated operations without further human intervention, enabling threat actors "to leverage AI to execute 80-90% of tactical operations independently at physically impossible request rates."[75] Taken together, these cases represent an evolution in AI-assisted cybercrime and should meaningfully impact existing threat models.[76]

Beyond the intentional misuse of AI agents, it is also important to acknowledge that some downstream effects from agents could arise from unsupervised security flaws or accidental harms in production systems, including attacks targeting the deployed agent itself to bypass its safety controls.[77]

From Model Error to Systemic Risk: Why Agentic AI Challenges Existing Governance

Agentic risks differ fundamentally from traditional model risks, yet most governance frameworks remain anchored in assumptions shaped by early LLMs and frontier models. Conventional model failures—such as bias, hallucination, or misclassification—are typically confined within a bounded interface or task. While these failures can be harmful, their effects are often localized and can be mediated through human intervention or downstream controls.

Agentic systems differ not only in their ability to act, but in how they combine adaptive reasoning, goal-directed behavior, and delegated authority across multiple systems and timeframes. Unlike scripts or automated tools that execute predefined actions, agents' ability to dynamically sequence operations, integrate new information, and adjust behavior in real time enables digital decisions to translate into real-world effects in ways that are harder to

73 Google Threat Intelligence Group, "GTIG AI Threat Tracker: Distillation, Experimentation and (Continued) Integration of AI for Adversaruak Use," Google Cloud, February 12, 2026, https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use.

74 Google Threat Intelligence Group, "GTIG AI Threat Tracker."

75 "Disrupting the first reported AI-orchestrated cyber espionage campaign," *Anthropic*, November 2025, https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf.

76 "Disrupting the first reported AI-orchestrated cyber espionage campaign."

77 For example, prompt injection attacks have demonstrated that AI systems using external retrieval or tool access can be manipulated via malicious instructions embedded in inputs (e.g., webpages, emails, or documents), causing them to ignore prior constraints, expose sensitive information, or execute unintended tool actions; similar vulnerabilities have also been observed in early web-enabled and plugin-based agent systems, where adversarial inputs or content sources can indirectly steer system behavior beyond intended safety controls.

anticipate or govern, amplifying trust and accountability challenges beyond the scope of conventional model risk management.

Agent failures, unlike traditional model failures that are primarily associated with prediction errors, arise from action selection, planning, memory, tool use, and execution across systems.[78] This creates a structural risk, where even minor faults in an agent's decision-making or environmental interactions can propagate through connected workflows, triggering cascading failures. When agents have significant freedom to modify their own activities, adapt strategies, or coordinate across services without human oversight, the severity of these cascades is amplified. In such contexts, a small misalignment or compromise, whether via tool access, memory updates, or delegated goals, can ripple across financial, operational, and civic systems, producing effects that are disproportionate to the initial error.

These dynamics build on an already shifting threat landscape. Security leaders increasingly recognize AI as a central driver of cyber risk. Recent industry reporting indicates that more than 90 percent of security leaders now expect AI-enabled attacks to occur daily, with two-thirds identifying AI as the most consequential driver of cybersecurity change this year.[79] While these assessments do not always distinguish between agentic and other AI-assisted attacks, they point to a broader trend: AI is accelerating the tempo and scale of cyber operations. Palo Alto Networks' Unit 42 documented that the mean time to data exfiltration has collapsed from nine days in 2021 to just two days in 2024, with some breaches concluding in under an hour.[80] The steady compression of reconnaissance, exploitation, and persistence into continuous feedback loops suggests that attack windows will shrink further as agentic systems mature.

Despite these risks, many organizations still rely on model-centric guardrails, such as content filters, evaluation benchmarks, and prompt controls, that were designed for static LLMs. These protections do not adequately capture vulnerabilities introduced by agentic autonomy. The consequence is a widening gap: agentic risks are systemic, while governance remains largely model-centric.

78 Datta et al., "Agentic AI Security: Threats, Defenses, Evaluation, and Open Challenges," *arXiv*, October 27, 2025, https://arxiv.org/html/2510.23883v1.

79 "Trend Micro State of AI Security Report 1H 2025," *Trend Micro*, July 29, 2025, https://www.trendmicro.com/vinfo/us/security/news/threat-landscape/trend-micro-state-of-ai-security-report-1h-2025

80 Sam Rubin, "Unit 42 Develops Agentic AI Attack Framework," *Palo Alto Networks*, May 14, 2025, https://www.paloaltonetworks.com/blog/2025/05/unit-42-develops-agentic-ai-attack-framework/

Recommendations

The following recommendations highlight three mutually-reinforcing areas of action: governance frameworks, public-private collaboration, and technical initiatives. In each, progress is already underway, and coordinated action could meaningfully enhance safe, trustworthy deployment of agents.

Governance Frameworks

Major technology firms have articulated internal frameworks to guide the responsible development and deployment of AI agents.[81,82,83] While these efforts vary in maturity and scope, they converge on two key insights:

1. **SYSTEM-LEVEL GOVERNANCE:** Risks cannot be addressed solely at the model level. Governance must be embedded across the full system architecture.[84] Like credentialed employees or contractors, agents may possess legitimate access to sensitive systems, operate with delegated authority, and execute actions that appear routine until aggregated over time. Effective controls therefore mirror those for other high-impact digital actors: strong identity binding, least-privilege access, behavioral monitoring, auditability, and lifecycle management. Governance must assume that risk arises not only from external compromise, but also from misuse, drift, or manipulation of trusted internal actors.
2. **DISTRIBUTED ACCOUNTABILITY:** AI agents operate through layered dependencies, including model providers, tool integrations, deployment configurations, and ongoing human supervision. This makes responsibility inherently distributed. What distinguishes agentic systems is not simply shared control, but delegated discretion: once deployed, agents can interpret goals, select actions, and adapt within the operational bounds set by others. Governance frameworks should therefore allocate responsibility across the lifecycle, from model development and system design to configuration, deployment, monitoring, to post-incident response. Importantly,

81 "Microsoft Agent Framework," *Microsoft*, October 9, 2025, https://learn.microsoft.com/en-us/agent-framework/overview/agent-framework-overview.

82 Amy Chang, "Introducing Cisco's Integrated AI Security and Safety Framework," *Cisco*, December 16, 2025, https://blogs.cisco.com/ai/security-framework.

83 Siddhi Shreekar Gowaikar, Andrea Colmenares, and Sahiba Pahwa, "Unlocking the power of Agentic AI with new watsonx.governance capabilities," last accessed April 2026, https://www.ibm.com/new/announcements/agentic-ai-governance-evaluation-and-lifecycle

84 Appendix B contains a detailed discussion of Microsoft's Agent Framework, Taxonomy of Failure, Cisco's Integrated AI Security and Safety Taxonomy, and Meta's "Agents Rule of Two."

these frameworks should not collapse accountability onto a single node. Treating traceability measures, such as structured documentation and provenance records (akin to SBOMs), as mandatory technical security controls can enhance oversight and verifiability, support auditing, and provide actionable context for investigations when errors or misuse occurs. This approach reflects the broader finding of this paper: in agentic ecosystems, responsibility is fragmented by design, and governance must explicitly structure that fragmentation rather than assume a single point of control.

Public-Private Partnerships

Given the distributed nature of agentic risk, no single public or private actor can effectively govern agentic risks alone. Public-private partnerships are therefore essential to align expectations, reduce fragmentation, and share the burden of risk management. The Coalition for Secure AI (CoSAI)[85] exemplifies this approach by convening industry heavyweights around shared safety practices and risk taxonomies.

These partnerships also address asymmetries in resources and expertise. Large technology firms can invest in and build bespoke governance infrastructures, while smaller organizations and public-sector deployers can leverage shared guidance, incident reporting mechanisms, and reference architectures. Early signal sharing of everything from misuse patterns to cascading failures can help lower barriers to responsible deployment.

This early signal sharing also supports adaptive governance that evolves alongside agent capabilities rather than lagging behind them.

Technical Initiatives

Alongside the aforementioned efforts, a set of technical initiatives are beginning to take shape to support and materially improve transparency and accountability in agentic systems.

Agent Documentation and Transparency Artifacts

Agent cards, analogous to model cards, aim to document an agent's intended purpose, capabilities, permissions, limitations, and governance controls.[86,87] However, consistent with the distributed nature of responsibility in agentic systems, documentation cannot rest with a single actor. Model providers may document baseline capabilities and safety constraints; platform providers may define integration limits; and deployers may specify configuration

85 Heather Adkins, Phil Venables, "Introducing the Coalition for Secure AI (CoSAI) and founding member organizations," *Google*, July 18, 2024, https://blog.google/innovation-and-ai/technology/safety-security/google-coalition-for-secure-ai/.

86 Surapaneni et al., "Announcing the Agent2Agent Protocol (A2A)," *Google*, April 9, 2025, https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/.

87 "ChatGPT agent System Card," *OpenAI*, July 17, 2025, https://openai.com/index/chatgpt-agent-system-card/.

choices, delegated authority, and operational context. When implemented consistently, such artifacts could help clarify what an agent is authorized to do, under what conditions, and with which safeguards in place. While documentation alone cannot prevent misuse or provide an exhaustive list of agent behavior and actions, it can support procurement decisions, auditing practices, and provide critical context during incident response. And in multi-agent environments, these artifacts may become essential for understanding how responsibility and authority are distributed across systems.

Red-Teaming and Simulation-Based Testing

Agentic systems operate over time, adapt to feedback, and coordinate across tools—features that strain traditional security testing approaches. Conventional testing often focuses on static vulnerabilities (e.g., code flaws), bounded model evaluations (e.g., benchmark accuracy or prompt robustness), or point-in-time penetration testing that assumes fixed behaviors. These methods are not designed to capture how an agent sequences actions, revises plans, escalates privileges through tool chains, or behaves after prolonged interaction.

Red-teaming for agentic systems must therefore model sequences of interaction and evolving contexts, such as how agents respond to ambiguity, recover from errors, update memory, interact with other agents, and propagate faults across interconnected services. Simulation-based environments allow organizations to stress-test agents under adversarial and high-autonomy conditions, surfacing second-order risks such as privilege escalation, feedback-loop manipulation, loss of human oversight, or cascading failures across systems. These exercises also help identify meaningful intervention points where human operators can reassert control before localized faults become systemic incidents.

Risk Allocation and Pre-Deployment Responsibility

Organizations and policymakers should consider mechanisms for clarifying responsibility for AI agent actions in advance of deployment. Agentic systems complicate post hoc accountability as authority, configuration, and execution are distributed across actors and evolve over time. When responsibility is determined only after harm occurs, victims may lack clear recourse and organizations may face uncertain exposure.

Drawing a loose analogy to product liability regimes, stakeholders could explore approaches in which accountability is defined ex ante, whether assigned to developers, deployers, or shared across them, even when harm is unintended. The goal is not to prescribe a specific liability model, but to ensure that responsibility is structured deliberately rather than deferred until failure.

In practice, this could take the form of contractual frameworks that explicitly allocate responsibility across the agent lifecycle, technical safeguards that constrain agent behavior within defined operational boundaries and maintain auditable decision pathways, and insurance or risk management mechanisms calibrated to foreseeable agent failure modes. Clarifying responsibility in advance would allow organizations to price risk more accurately, design systems with accountability in mind, and provide more predictable recourse for affected parties.

This recommendation is intended as a discussion starter for legal and policy experts to evaluate feasibility, trade-offs, and unintended consequences, rather than as a prescriptive legal blueprint.

Conclusion

AI agents represent a structural shift in how digital systems operate. As software systems move from assisting human decision-making to planning, acting, and coordinating with limited oversight, they are also testing longstanding assumptions about identity, accountability, and security. The challenge is not simply that agents act autonomously, but that they operate persistently across systems, blur the line between human and machine intent, and distribute authority in ways that existing governance frameworks were not designed to accommodate.

As we articulate in this paper, the implications are threefold. Identity and attribution become more fragile when agents act persistently across systems, emulate human behavior, and obscure the provenance of action. Agency and responsibility fragment when authority is delegated but outcomes diverge from original intent, challenging legal doctrines and organizational risk models alike. Security risks become systemic when autonomous, tool-enabled agents can act at machine speed, coordinate across environments, and amplify both defensive and offensive capabilities.

None of these challenges are hypothetical. They are already shaping how organizations make procurement decisions, how a human decides whether to deploy agents, how investigators attempt to trace malicious activity, and how adversaries exploit ambiguity and asymmetry. In this sense, governance gaps are not merely a downstream concern; they are active forces influencing the balance between innovation and risk, trust and opacity, and resilience and fragility.

The path forward is not to halt innovation, but to align governance with architecture. System-level oversight, clarified responsibility allocation, technical auditability, and coordinated public–private action are not peripheral safeguards—they are prerequisites for trustworthy deployment. The central question is not whether AI agents will become embedded in digital infrastructure, but whether they will do so in ways that remain governable over time.

Early decisions that we make about identity design, delegation boundaries, and accountability structures will play an important role going forward. We have an opportunity now to build agentic systems that preserve human agency, sustain recourse, and reinforce trust, before those design choices harden into default norms. The window for shaping those foundations is narrow, but it remains open.

Appendix A: Research Methodology

This white paper draws on a combination of literature review, expert consultation, and practitioner engagement conducted between 2024 and early 2026. The research process included structured interviews with industry practitioners, cybersecurity experts, and policy specialists working on AI systems and digital infrastructure; closed-door workshops and roundtables convened with stakeholders from the technology sector, academia, civil society, government, and law firms; review of technical literature, including research on AI agents, cybersecurity risk, identity systems, and AI governance frameworks; and an analysis of emerging policy discussions related to AI safety and security, digital identity, and cyber attribution.

These engagements were conducted under the Chatham House Rule unless otherwise stated, enabling participants to share experiences and concerns candidly.

The goal of the research was not to provide a comprehensive taxonomy of AI agents, but rather to identify emerging governance challenges and areas of convergence across stakeholders.

Appendix B: Existing AI Agent Taxonomies

Taxonomies and classification frameworks for AI agents and agentic systems are still emerging as academic researchers and technology companies attempt to bring structure to a rapidly evolving field. No single industry consensus exists yet, but several frameworks help clarify how AI agents are characterized, compared, and governed.

Classical Taxonomies of AI Agents

Foundational AI literature and enterprise descriptions categorize agents based on decision logic, autonomy, and learning capability:

- **Simple reflex agents:** The most basic type of agent, capable of directly responding to an environment using predefined rules with no memory or planning.[88]

- **Model-based reflex agents:** These agents maintain an internal model of the environment to inform action selection in "situations where it cannot perceive everything directly."[89]

88 Cole Stryker, "What is a simple reflex agent?" *IBM*, https://www.ibm.com/think/topics/simple-reflex-agent.

89 Ivan Belcic, Cole Stryker, "Model-based reflex agents, defined," *IBM*, https://www.ibm.com/think/topics/model-based-reflex-agent#72820456.

» **Goal-based agents:** Agents that evaluate actions based on its progress with proactive problem-solving and decision making abilities to achieve goals.[90]

» **Utility-based agents:** Agents that assign values to potential outcomes and choose actions or make rational decisions that maximize utility.[91]

» **Multi-agent systems (MAS):** Composed of multiple interacting agents, which may exchange information to coordinate, communicate, and/or collaborate to achieve individual or collective goals. MAS introduces emergent behavior that cannot be fully understood by examining a single agent in isolation, reflecting the highly complex and distributed nature of many real-world applications.[92]

These categories highlight differing levels of sophistication, from rules-based automation to adaptive, goal-oriented systems capable of complex decision making and coordinated behavior.[93]

Industry and Corporate Frameworks

Microsoft

Microsoft has published taxonomies that focus on failure modes in agentic AI systems, distinguishing between safety and security failures and whether harms are novel to agentic AI or extend from existing AI system classes. This taxonomy is aimed at helping engineers and safety and security professionals identify real-world failure categories and corresponding mitigation strategies.

The taxonomy differentiates between failures that arise from unintended or misaligned behavior and those that result from adversarial manipulation or exploitation. Safety failures are generally associated with agents pursuing goals in ways that conflict with user intent, bias, user impersonation, organizational policy or knowledge loss, or broader societal norms. These failures may stem from ambiguous instructions, flawed reasoning, incomplete context, or the agent's inability to appropriately weigh competing objectives. In agentic systems, such failures are amplified because they translate directly into actions rather than remaining confined to outputs.

Security failures, by contrast, arise when agents are deliberately manipulated or exploited by external actors. Microsoft highlights how agentic architectures introduce new attack surfaces, particularly through natural language interfaces, tool integrations, and memory systems. Techniques such as prompt injection, tool misuse, and indirect input manipulation can cause agents to deviate from intended behavior, execute unauthorized actions, or expose sensitive

90 David Zax, "Goal-based agents, defined," *IBM*, https://www.ibm.com/think/topics/goal-based-agent#72820455.
91 Ivan Belcic, Cole Stryker, "Utility-based agents, defined," *IBM*, http://ibm.com/think/topics/utility-based-agent#72820458.
92 Shalini Harkar, "What is multi-agent collaboration?" *IBM*, https://www.ibm.com/think/topics/multi-agent-collaboration#2014952963.
93 Ram Shanker Siva Kumar, "New whitepaper outlines the taxonomy of failure modes in AI agents," *Microsoft Security*, April 24, 2025, https://www.microsoft.com/en-us/security/blog/2025/04/24/new-whitepaper-outlines-the-taxonomy-of-failure-modes-in-ai-agents/.

information. Because agents can plan, adapt, and chain actions across systems, these exploits can propagate in ways that exceed the scope of traditional software vulnerabilities.

A key contribution of this taxonomy is its recognition that many failure modes are not entirely novel, but are instead amplified by agentic properties such as autonomy, persistence, and cross-system interaction. Failures that might once have been contained within a single model output can now unfold over time, across multiple steps, and through external systems, increasing both their impact and their difficulty to detect. The framework also emphasizes that safety and security failures are often interdependent rather than distinct. A safety weakness, such as poor instruction following or overgeneralization, can become a security vulnerability when exploited by an adversary. Conversely, security breaches may manifest as safety failures when compromised agents behave in misaligned or harmful ways. This convergence reinforces the need for integrated approaches to evaluation and governance that account for both dimensions simultaneously.

IBM

IBM's watsonx Risk Atlas frames agentic systems through the risks they introduce across the lifecycle of deployment.[94] This approach reflects a broader shift from classifying what agents are to understanding what agents do in operational environments.

In this framework, risks emerge from the defining characteristics of agentic systems, including their ability to make autonomous decisions, interact with external tools and infrastructure, retain and update memory over time, and adapt their behavior in response to changing inputs and environments. These features introduce distinct challenges, such as the potential for agents to execute actions with real-world consequences, to operate across system boundaries in ways that complicate oversight, and to carry forward state or context that may be outdated, manipulated, or misaligned with current objectives.

The framework also emphasizes how loosely specified or evolving goals can produce unintended outcomes, particularly when agents interpret high-level instructions in dynamic environments. These risks are further amplified in multi-agent settings, where interactions between systems can generate emergent behaviors that are difficult to anticipate, monitor, or attribute to any single actor.

Notably, this risk-oriented taxonomy treats agentic AI not as a static system, but as a dynamic and evolving actor embedded within a broader operational context. It aligns with a growing recognition that governance must extend beyond model-level evaluation to account for how agents behave over time, across tools, and in coordination with other systems. By organizing

94 "AI risk atlas," *IBM*, October 23, 2025, https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas.

agentic systems around risk vectors rather than discrete categories, IBM's framework provides a bridge between technical design and governance application, supporting more practical approaches to auditing, monitoring, and accountability in real-world deployments.

Cisco

Cisco's Integrated AI Security and Safety Framework presents a taxonomy of risk categories across AI systems that explicitly includes agentic behavior, multi-agent coordination, and lifecycle exposures. As "one of the first holistic attempts to classify, integrate, and operationalize the full range of AI risks, from adversarial threats, content safety failures, model and supply chain compromise, agentic behaviors and ecosystem risks (e.g., orchestration abuse, multi-agent collusion), and organizational governance," the Framework provides a structured layout of AI threats and their impacts.[95]

A distinguishing feature of this framework is its emphasis on the convergence of safety and security risks. Issues such as prompt injection, model manipulation, or unintended behavior are treated not only as alignment concerns but as concrete security vulnerabilities when they enable unauthorized actions or system compromise. This framing is particularly relevant for agentic systems, where natural language interfaces, tool use, and autonomy expand the range of possible attack surfaces.

Cisco's framework also highlights the importance of lifecycle-aware governance, recognizing that risks evolve as agents are developed, deployed, and updated in production environments. As agents gain access to additional tools, data sources, or decision-making authority, their risk profile changes, requiring continuous monitoring and adaptation of controls. This reinforces the need for persistent visibility, auditability, and feedback mechanisms capable of detecting both acute failures and gradual behavioral drift.

95 Amy Chang, "Introducing Cisco's Integrated AI Security and Safety Framework," *Cisco*, December 16, 2025, https://blogs.cisco.com/ai/security-framework.

INSTITUTE FOR SECURITY AND TECHNOLOGY
www.securityandtechnology.org

info@securityandtechnology.org