Association for Computing Machinery

acm Technology Policy Council

TechBriefs

Autonomy, Opportunities, and Challenges of Action-Taking AI Systems

by Rohith Nama, Dr. Larry Medsker, and Brian Peretti



Agentic AI systems include an AI component that plans and executes sequences of actions toward a user-defined goal, accessing external software, data, and services along the way. As capabilities continue to advance, defining Agentic Systems will evolve. Earlier AI systems with a chat interface produced text or other content for human review; agentic systems can take actions with real-world consequences, such as browsing the web, writing and executing code, sending messages to other agents or humans, and managing files. The decision of how much human oversight and control to exert over agentic AI systems depends on policy and design. Even with oversight, users must delegate some authority to these systems to act on their behalf.

This shift from output-generating to action-taking systems introduces risks and responsibilities that existing legal, regulatory, and technical frameworks do not currently address. Adoption is accelerating: a 2025 survey of over 500 technology leaders found that 48% are already adopting or deploying agentic AI.[1] With national and sub-national laws coming online, agentic AI raises questions that these frameworks, designed primarily for generative AI, do not fully address.

Agentic AI systems are typically built on large language models and inherit the limitations of chat-based AI, including hallucination and bias. Autonomy exists on a continuum: some require explicit authorization for every action; others operate completely independently toward a stated goal. Most require human approval for consequential actions, but approval can be waived over time. The boundary between autonomous and supervised actions is a design choice that can change without user notification.

Table 1: Differences between chat-based AI and agentic AI systems

	Chat-based AI	Agentic AI
Primary mechanism	Produces content in response to a prompt	Plans and executes multi-step tasks to achieve a goal
Human role	User reviews output before acting	User sets goal; agent acts without step-by-step approval
External access	Limited; may browse the web, but does not write to external systems	Reads from and writes to external systems with software interfaces
Error recovery	User has the opportunity to review the output before acting on it	Agents may act on an error before human review. Agents may provide an "undo" feature that may attempt to revert external changes.
Accountability	Typically attributed to the user who acts on the output	Potentially distributed across model provider, framework, deployer, and user
Primary security concern	Hallucination, bias, insecurely generated code, and reproduction of copyrighted content	Prompt injection, tool misuse, cascading failures, plus concerns from Chat-based systems.

Key Policy Dimensions

Legal Responsibility and Liability. Agentic AI users typically define high-level goals without specifying operational boundaries, such as stop conditions or escalation triggers. Accountability for harms caused by these systems is thus diffused in ways that existing legal doctrines were not designed to handle. Responsibility may be attributed to the model

Creative Commons license.

ACM's Technology Policy Committees provide cutting-edge, apolitical, non-lobbying scientific information about all aspects of computing to policy makers in the United States and Europe. To tap the deep expertise of ACM's 100,000 members worldwide, contact ACM's Global Policy Office at acmpo@acm.org or +1 202.580.6555.

acm Technology Policy Council

provider, framework developer, deployer, or end user, as each shapes the system's behavior, but none controls any specific action. This makes it difficult to assess responsibility for harms after the fact.

- In one incident, an AI coding agent deleted a company's entire production database despite explicit instructions to make no changes. Responsibility for the losses might be attributed to the company that deployed the agent, the platform provider, the model provider, or the developer who commanded the agent and configured its permissions. No person made "the decision" to delete the database, yet harm was done.[2]
- There is no case law applying existing doctrines such as product liability, negligence, and software licensing to Agentic AI. Software licensing allocates risk through contractual terms, but these may not map to systems that take autonomous actions affecting third parties not party to the license.[3]
- The European Parliament's 2025 study on AI civil liability found that existing product liability rules struggle with autonomous AI systems, making causation and fault difficult to establish across AI supply chains.
- Accountability may also depend on uneven infrastructure across deployments, including audit trails, replayable action logs, and standardized incident reporting. Where such infrastructure is absent, establishing what happened and why is difficult, regardless of how liability rules are framed.

Resolving liability questions for agentic AI will likely require both substantive legal clarification and parallel investment in the technical mechanisms needed to apply it.

Security and Operational Risk. Agentic AI introduces security challenges beyond those of chat-based systems.

- Because LLMs process text as both data and commands, Agentic AI systems cannot reliably distinguish between legitimate content and embedded instructions.[4]
  - In one case, malicious instructions embedded in a Slack message caused an AI agent to retrieve and expose content from private channels without any user action.[5]
- Agents that retain memory across sessions may also be vulnerable to memory poisoning, where malicious inputs persist and influence future actions.
- Agents are often configured with broad access rather than permissions scoped to specific actions, which can increase exposure when errors occur.
- Additional concerns include vulnerabilities in agent-generated code, unauthorized access to copyrighted data, and risks from third-party extensions and plugins.
- Although guardrail frameworks that monitor agent behavior and detect prompt injection are under development, none currently provide comprehensive protection.[6,7]

Agentic systems' security depends on choices about permissions, oversight, and the boundaries of autonomous action, choices that are both technical and organizational.

Consumer Transparency and Recourse. When an AI agent acts on an individual's behalf, the user may not know what systems the agent can connect to, what actions it can take without explicit confirmation, or how to revoke permissions once granted. No standardized disclosure format currently exists for agent permissions.

- OpenClaw can read and send emails, manage calendars, and execute terminal commands through consumer messaging apps, accumulating more than 300,000 users within weeks of its 2025 release.[8] Security researchers found that approximately 12% of third-party extensions in its marketplace contained malicious code, including keyloggers and data exfiltration tools.[9]
- Because many agents communicate in natural language and exhibit human-like responsiveness, users may develop trust in their reliability, particularly in sensitive domains such as healthcare, where an agent's confident tone can mask clinical uncertainty.[10]
- Agents that act through natural language interfaces can also introduce language-mediated harms. An agent that sends an email or schedules a meeting on a user's behalf may damage relationships, reputation, or trigger legal exposure.

Consumer transparency for agentic AI involves disclosing that an AI is in use, making the scope of an agent's authority visible, and accounting for potential harms.

Multi-Agent Coordination Risks. When multiple agents interact in coordinated networks, they can produce unexpected outcomes. Multi-agent networks may also reinforce one another's errors, a pattern called conformity bias.[11,12]

Workforce and Societal Impact. Organizations view agentic AI as a means to reduce current headcount, particularly in entry- and mid-level roles. A 2025 Gartner survey of supply chain leaders found that 55% expect agentic AI to reduce hiring for entry-level positions, and 51% believe the technology will drive a shift to overall workforce reductions.[13] Similar dynamics are happening across multiple sectors.

- Reductions in entry-level positions will likely shrink pathways through which workers traditionally develop expertise. Where junior roles disappear, the path to mid-career and senior expertise narrows, raising questions about how skills and judgment are formed in fields where hands-on practice has been essential.
- Independent studies measuring the actual effects of agentic AI on organizational effectiveness, service quality, employment composition, or equity remain limited. Productivity claims have not been independently verified at scale, even as those claims are already shaping hiring decisions.

Accelerating displacement, narrowing entry pathways, and unverified productivity gains may have long-term effects on skill formation and labor markets that are not yet measurable but are already shaping organizational behavior. Tracking these effects systematically may be a precondition for informed labor and education policy responses.

Conclusion

Agentic AI represents a shift from AI systems that generate content for human review to systems that take actions on humans' behalf. This raises questions about legal liability, consumer transparency, and regulatory adequacy that are distinct from earlier generations of AI.

Many technical and operational challenges are being addressed in industries through risk management, engineering practices, and emerging standards. Relevant standards efforts are underway, including NIST's AI Agent Standards Initiative, OECD analytical frameworks, and industry-funded research into secure agent ecosystems.[14] Yet while enterprise deployments often involve staged integration and constrained environments, consumer-facing systems are progressing with less structured oversight.

In sectors such as healthcare, financial services, and critical infrastructure, where existing frameworks assume a human decision-maker, existing law may need to be supplemented by sector-specific guidance tailored to autonomous systems. Agentic systems that operate across jurisdictions for a single task also raise cross-border accountability questions that existing international mechanisms have yet to address.

A central policy question is who is responsible when autonomous systems cause harm, and how to ensure adequate consumer transparency when agents act on an individual's behalf. The technical question is how to make AI Agentic Systems reliably follow the policies that they are given. To align these goals, safeguards such as defined authentication and delegation, the ability to audit and monitor systems, support for investigations and fault determination, and further research into inter-agent dynamics are needed.[15]

REFERENCES

1 EY. 2025. EY Technology Pulse Poll, April 2025 (504 U.S. technology leaders surveyed). (2025). https://www.ey.com/en_us/newsroom/2025/05/ey-survey-reveals-that-technology-companies-are-setting-the-pace-of-agentic-ai-will-others-follow-suit.
2 Beatrice Nolan. 2025. AI-powered coding tool wiped out a software company's database. *Fortune*, (July 23, 2025). https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure/.
3 Andrea Bertolini. 2025. Artificial Intelligence and Civil Liability: A European Perspective. Study PE 776.426. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2025/776426/IUST_STU(2025)776426_EN.pdf.
4 Fábio Perez and Irene Ribeiro. 2022. Ignore Previous Prompt: Attack Techniques for Language Models. In *Proceedings of the NeurIPS Machine Learning Safety Workshop*. https://arxiv.org/abs/2211.09527.
5 PromptArmor. 2024. Data Exfiltration from Slack AI via Indirect Prompt Injection. (Aug. 2024). https://www.promptarmor.com/resources/data-exfiltration-from-slack-ai-via-indirect-prompt-injection.
6 Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (EMNLP '23). Association for Computational Linguistics, Singapore, 431–445. https://doi.org/10.18653/v1/2023.emnlp-demo.40.
7 Sahana Chennabasappa, Cyrus Nikolaidis, Daniel Song, David Molnar, Stephanie Ding, Shengye Wan, Spencer Whitman, Lauren Deason, Nicholas Doucette, Abraham Montilla, Alekhya Gampa, Beto de Paola, Dominik Gabi, James Crnkovich, Jean-Christophe Testud, Kat He, Rashnil Chaturvedi, Wu Zhou, and Joshua Saxe. 2025. LlamaFirewall: An Open Source Guardrail System for Building Secure AI Agents. (2025). https://arxiv.org/abs/2505.03574 arXiv: 2505.03574.
8 Elia Zaitsev. 2026. What Security Teams Need to Know About OpenClaw, the AI Super Agent. *CrowdStrike Blog*, (Feb. 4, 2026). https://www.crowdstrike.com/en-us/blog/what-security-teams-need-to-know-about-openclaw-ai-super-agent/.
9 Amy Chang and Vineeth Sai Narajala. 2026. Personal AI Agents like OpenClaw Are a Security Nightmare. *Cisco AI Defense Blog*, (Jan. 30, 2026). https://blogs.cisco.com/ai/personal-ai-agents-like-openclaw-are-a-security-nightmare.
10 Alejandro Bellogín, Paolo Giudici, Stefan Larsson, Jun Pang, Gerhard Schimpf, Biplav Sengupta, and Gonenc Solmaz. 2025. Systemic Risks Associated with Agentic AI: A Policy Brief. ACM Europe Technology Policy Committee. https://www.acm.org/binaries/content/assets/public-policy/europe-tpc/systemic_risks_agentic_ai_policy-brief_final.pdf.
11 Christian Schroeder de Witt, Klaudia Krawiecka, Igor Krawczuk, Ben Hagag, William L. Anderson, Peter Belcak, Ben Bucknall, Xiaohong Cai, Ayush Chopra, Doron Cohen, Ron F. Del Rosario, Andis Draguns, Annie Gray, Keren Katz, Vasilios Mavroudis, Jaron Mink, Sumeet Ramesh Motwani, Jonathan Petit, Leif-Sebastian Rembeck, Chandler Smith, John Sotiropoulos, Steven Young, Sarah Scheffler, and Mary Llewellyn. 2025. Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. (2025). https://arxiv.org/abs/2505.02077 arXiv: 2505.02077.
12 Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11, 20, (May 2025), eadu9368. https://doi.org/10.1126/sciadv.adu9368.
13 Gartner, Inc. 2026. Gartner Survey Shows 55% of Supply Chain Leaders Expect Agentic AI to Reduce Entry-Level Hiring Needs. (Feb. 25, 2026). https://www.gartner.com/en/newsroom/press-releases/2026-02-25-gartner-survey-shows-55-percent-of-supply-chain-leaders-expect-agentic-ai-to-reduce-entry-level-hiring-needs.
14 National Institute of Standards and Technology (NIST). [n. d.] AI Agent Standards Initiative. Retrieved May 14, 2026 from https://www.nist.gov/caisi/ai-agent-standards-initiative.
15 Schmidt Sciences. 2025. Achieving a Secure AI Agent Ecosystem. Schmidt Sciences. https://www.schmidtsciences.org/wp-content/uploads/2025/06/Achieving_a_Secure_AI_Agent_Ecosystem-3.pdf.

ADDITIONAL INFORMATION

With 100,000 members in 190 countries, the nonprofit Association for Computing Machinery is the world's largest and longest-running organization of professionals involved in all aspects of computing. Under the auspices of the global ACM Technology Policy Council, committees in the U.S. and Europe provide cutting-edge, apolitical, non-lobbying information about computing and its social impacts to policymakers at all levels of government in many forms, including briefings, testimony, consultations, and rulemaking comments, reports.

To tap the deep expertise of ACM's global membership, please contact ACM's Global Policy Office at *acmpo@acm.org* or +1 202.580.6555. To receive ACM TechBriefs quarterly, in the body of a one-line email send *subscribe ACM-tpc-tech-briefs* followed by your first and last names to *listserv@listserv.acm.org*

ACKNOWLEDGMENTS

The authors would like to thank the following subject-matter experts for their reviews and input: Mr. Gerhard Schimpf, Dr. Benjamin Cedric Larsen, and Dr. Geoff Cohen. A thank you to the ACM TechBrief Committee and the ACM Technology Policy Council for their valuable input. The PDF version of the document meets WCAG 2.2 Level AA accessibility standards and complies with the PDF Accessibility Checker (PAC2024) requirements.

DOI: https://doi.org/10.1145/3819832 ISBN: 979-8-4007-2785-6 Creative Commons license.