Response to the Request for Information Regarding Security Considerations for Artificial Intelligence Agents

March 9, 2026

To the Center for AI Standards and Innovation (CAISI),

Thank you for the opportunity to submit recommendations on Security Considerations for Artificial Intelligence Agents. We are researchers affiliated with UC Berkeley, with expertise in AI research and development, security, and robustness. We previously submitted responses to requests for information (RFIs) and requests for comment (RFCs) including responses to OSPT in April 2025 on the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan in April 2025, to OSTP on the Development of an Artificial Intelligence (AI) Action Plan in March 2025, to NIST on Managing the Risk of Misuse for Dual-Use Foundation Models in March 2025, to NIST in May 2024 on the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, and at multiple points in 2021-2023 at various stages of NIST development of AI RMF guidance.

Most recently, we completed a yearlong research initiative to develop a NIST-AI-RMF-compatible profile specifically tailored to agentic systems: the **Agentic AI Risk-Management Standards Profile** ("Agentic AI Profile") (Madkour et al. 2026). This effort was designed to complement and extend the NIST AI RMFand addresses the unique technical, operational, and security challenges posed by AI agents—particularly those capable of autonomous planning, tool use, and persistent environmental interaction. While our responses draw directly from this work, we encourage reading the full Agentic AI Profile for more guidance on the topic.

Below we provide (1) overarching recommendations on security considerations for AI Agents, and (2) answers to the specific questions posed in the Federal Register RFI (NIST-2025-0035) on security considerations for AI Agents.

Thank you again for the opportunity to comment on Security Considerations for Artificial Intelligence Agents. If you need additional information or would like to discuss further, please contact Nada Madkour at nada.madkour@berkeley.edu.

Our best,

Nada Madkour, Ph.D.
Interim Director
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Deepika Raman
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Charlotte Yuan
Graduate Student Researcher
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Krystal Jackson
Non-Resident Research Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

1. Overarching Recommendations

In this section, we provide overarching recommendations on Security Considerations for Artificial Intelligence Agents.

Scale Governance Mechanisms with Degrees of Autonomy
Given the different configurations of agentic AI systems, governance mechanisms should scale with degrees of agency, rather than treating autonomy as a binary attribute. Agentic AI ranges from narrowly scoped, single-agent systems to highly autonomous, multi-agent architectures operating in complex environments, requiring risk controls that are proportionate to these characteristics. In practice, this guidance would resemble activating the strictest protocols for an agent granted high autonomy and extensive authority in a complex environment. Whereas agents with restricted authority and autonomy might trigger more proportional, model-level safeguards.

Support Human Control and Accountability
It is important to develop effective human-agentic AI management hierarchies that preserve human authority while leveraging AI as a supportive tool, and to establish hierarchical oversight and escalation pathways that provide a clear, tiered system of oversight, ensuring that human attention is directed where it is most needed. Real-time monitoring systems should be equipped with emergency automated shutdowns and be triggered by certain activities (e.g., access to systems or data outside of the agent's authorized scope) or crossed risk thresholds (Chan et al., 2024; Oueslati & Staes-Polet, 2025). In addition to automatic emergency shutdown, manual shutdown methods should be available as a last-resort control measure.

Implement Continuous Monitoring and Post-Deployment Oversight
Recognizing that agentic behavior may evolve over time and across contexts, the development and implementation of continuous monitoring and rapid-response infrastructures to accommodate for the speed of progress and to help adequately prepare for potential emerging risks and misuses is necessary. This includes investment in continuous monitoring mechanisms to keep track of and trace agent behavior in complex deployment environments (e.g., by using outcome monitoring), and investment in rapid-response infrastructure that can help in disabling agents or limiting their authority when significant evidence of unforeseen or emerging risks is observed.

Employ Defense-in-Depth and Containment

Given the many unknown and/or emergent risks from agentic systems and the lack of robust evaluation regimes, security considerations must evoke a layered technical, organizational, and societal safeguards across agentic AI development and deployment stages to ensure redundancy against failures (Bengio et al., 2025). Treating sufficiently capable agents as untrusted entities due to the limitations of current evaluation techniques can help mitigate risks from accidents, malfunctions, and malicious use.

Implement System-Level Risk Assessment

Move beyond model-centric approaches to evaluate risks across the agentic AI ecosystem that considers autonomy, authority, tool access, operating environment, and multi-agent interactions in order to address emergent threats like cascading failures or tool misuse.

2. Our Comments on Questions Posed in the Federal Register Request for Information

In this section, we provide answers to the specific questions posed in the Federal Register RFI (NIST-2025-0035) on Security Considerations for AI Agents.

2.1 Security Threats, Risks, and Vulnerabilities Affecting AI Agent Systems

(a) What are the unique security threats, risks, or vulnerabilities currently affecting AI agent systems, distinct from those affecting traditional software systems?

The following are some examples of risks that are unique to, or meaningfully exacerbated by AI agent systems. For more on AI agent risks, please see “Map 1.1” in Madkour et al. (2026).

Privacy and Security

- Comprehensive logging and traceability can potentially introduce significant privacy risks, including the misuse of sensitive information. It may also lead to unintended disclosure of individual identities through analysis of proxy data, usage patterns, or trends linked to specific users.
- The addition of memory into agentic systems increases the likelihood of data leakage, as these systems store and work with more sensitive data in a variety of untested or unexplored contexts that may result in private data being revealed.
- Multi-agent systems pose complex security challenges, as these systems can experience cascading compromises resulting in misaligned outcomes through the spread of malicious prompts across agents working together as a system, making mitigation and detection particularly difficult across multi-agent systems and the interconnected applications they access (Peigné et al., 2025).

Hallucinations

- In the context of multi-agent systems or agent-to-agent communication, faulty or hallucinated output from one agent can propagate, causing cascading misinformation (Sapkota et al., 2026).

Malicious Actors and Misuse

- The nature of agentic AI may allow for the automation of workflows for malicious uses. In addition to lowering the barriers to entry, facilitating the design of biological agents and aiding in the creation of harmful chemicals or other substances, agentic AI can potentially be used to automate parts of several attack stages in the risk pathway (ORF, 2024; Chin, 2025; Barret et al., 2024).
- Agentic AI could potentially be used to increase the scope and scale of cyberattacks by automating reconnaissance, exploit development, and payload delivery (Singer et al., 2025; Shao et al., 2025; Kouremetis, 2025; Dawson et al., 2025; Heiding et al., 2024). For example, orchestration and scaffolding have emerged as the most critical areas of development, allowing adversaries to both bypass safety measures and execute tactical

operations autonomously (Anthropic, 2025b; Lin et al., 2025). Advanced AI agents can be used to generate increasingly personalized manipulative content at scale (e.g., phishing, vishing) and can enhance tactics by integrating user feedback, expanding the attack surface for social engineering.

Human-Computer Interaction

- Reduction of human oversight may escalate risks and increase the likelihood of unnoticed accidents and malfunctions, such as API integration failures that lead to catastrophically misinterpreting the data, executing high-speed, erroneous transactions before human intervention (Bengio et al., 2025; Enkrypt AI, 2025).
- Anthropomorphic AI assistant behavior may increase user trust and encourage sharing of sensitive information, increase the effectiveness of manipulation, and promote overreliance (Akbulut et al., 2024).

Loss of Control

- An agentic model may intentionally disable oversight mechanisms or otherwise carry out oversight subversion in order to pursue its goals (Meinke et al., 2025). Rapid, autonomous, and iterative execution of actions may outpace monitoring and response mechanisms, creating risks of large-scale, potentially irreversible harm before intervention is possible.

Economic Impacts

- AI agents may have significant undesirable impacts on job markets due the technology's potential to provide cost-cutting automation in economically competitive markets (Bengio et al., 2025). The integration of agentic AI into the workforce introduces novel risks related to "agentic management," including mass worker surveillance and ambiguous workplace hierarchies.
- General-purpose AI systems require orders of magnitude more energy than task-specific alternatives. AI agents amplify these concerns, as poorly supervised agents may enter loops or run indefinitely, potentially incurring substantial energy costs (Luccioni et al., 2024; Guidi et al., 2024).
- Rapid automated development may lead to rapid acceleration of catastrophic capabilities at a pace that is faster than government response time, introducing risks related to delayed intervention and missed opportunities for timely international coordination and mitigation.

AI System Safety, Failures, and Limitations

- Agentic AI presents unique risks related to system safety, failures, and limitations. For example, self-proliferating AI may have the ability to independently function and obtain resources, potentially expanding its influence on its environment by enhancing its capabilities or scaling its operations (Phuong et al., 2024). Due to the lack of human oversight and the presence of reinforcing feedback loops, bounding the possible harms becomes increasingly difficult.
- Agentic AI may also pursue harmful actions or risky agentic behavior in pursuit of its objectives. A scheming agent, tasked with assisting in the drafting of its own safety and oversight protocols, could identify and subtly promote policies that contain exploitable loopholes.

- Collusion between agents could lead to the exacerbation of existing capabilities, the generation of entirely new risks, and new misaligned objectives in the pursuit of overlapping goals. It may also lead to the reinforcement of mutual error and the amplification of flawed design through iterative dialogue between agents, increasing the risk of agentic misalignment (Raza et al., 2025).
- Autonomous Goal Pursuit and Intent Hijacking: Attackers can exploit the lack of separation between data and instructions in LLMs to subtly alter an agent's objectives through prompt injection, causing it to pursue unintended or harmful goals while appearing to operate normally.

(b) How do security threats, risks, or vulnerabilities vary by model capability, agent scaffold software, tool use, deployment method (including internal vs. external deployment), hosting context (including components on premises, in the cloud, or at the edge), use case, and otherwise?

Security threats, risks, and vulnerabilities in agentic AI escalate with interactions among system factors. For example:

- Tool use amplifies misuse risks, such as unauthorized API calls or data exfiltration.
- Deployment methods can heighten impact levels. For instance, internal sandboxes limit blast radius vs. external user-facing APIs.

There are a considerable number of such variations, but they can be anticipated and at least partially circumvented by proactively taking into account characteristics and properties unique to agentic systems when developing risk management strategies. These characteristics are highlighted in "Map 5.1" in Madkour et al. (2026) and include:

- **Agent autonomy levels** and the degree of autonomy the agent falls under relative to the operational environment, scope of activities, and organizational risk tolerances (e.g., high autonomy in complex environments triggers stringent protocols);
- **Level of authority** the agent will have based on variables such as the range of actions the agent can perform (e.g., read-only vs. execute privileges);
- **Type and level of causal impac**t the agent will be capable of having within the environment (e.g., informational vs. physical or financial impacts);
- **Type of environment** in which the agent will be operating, as well as the environmental complexity (e.g., controlled internal vs. open-world external);
- **AI agent efficacy**, or the agent's ability to interact with and impact its operational environment (measured via benchmarks like success rates in tool-use tasks);
- **Extent of anthropomorphic features** and exercise caution when integrating these features into AI assistant user interfaces (to mitigate manipulation risks from overreliance); and
- **The function of the agent,** the role of the agent (specialist vs generalist), and agent predictability (deterministic vs non-deterministic).

(c) To what extent are security threats, risks, or vulnerabilities affecting AI agent systems creating barriers to wider adoption or use of AI agent systems?

Agentic AI may gain access to data, systems, or environments beyond an authorized scope and may attempt to ignore, circumvent, or misinterpret direct orders or constraints. Agentic systems may also find loopholes to pursue misaligned objectives. These risks may deter wider adoption of agentic systems due to lack of trust and the severity of potential harms posed by agentic failures, misalignments, vulnerabilities, and risks.

For example: Researchers have found that models have less agentic misbehavior when they are aware that they are being tested. Additionally, research results showed that a model attempted to blackmail a supervisor to prevent being shut down (Anthropic 2025a).

(e) What unique security threats, risks, or vulnerabilities currently affect multi-agent systems, distinct from those affecting singular AI agent systems?

The following are some examples of vulnerabilities that are unique to, or meaningfully exacerbated by multi-agent systems. For more on AI agent risks and vulnerabilities, please see "Map 1.1" in Madkour et al. (2026).

Collusion

- Collusion between agents could lead to the exacerbation of existing capabilities, the generation of entirely new risks, and new misaligned objectives (e.g., circumvention of safeguards) in the pursuit of overlapping goals. Collusion may also lead to the reinforcement of mutual error and the amplification of flawed design through iterative dialogues, increasing the risk of agentic misalignment (Raza et al. 2025). The emergence of tacit collusion in use cases such as autonomous pricing systems may lead to risks through iterative profit-driven interactions between pricing agents (Mukherjee and Chang 2025, Bertrand et al 2025).

Vulnerability propagation

- Vulnerabilities in one agent can propagate through agent-to-agent interactions, potentially exacerbating these vulnerabilities (Raza et al 2025, Sharma et al. 2025).In the context of multi-agent systems or agent-to-agent communication, faulty or hallucinated output from one agent can propagate, causing cascading misinformation (Sapkota et al. 2026).

Other

- Failure modes of multi-agent systems may significantly differ from those of the individual agents they are composed of. Some failure modes may be amplified and propagated through milti-agent feedback, and other entirely new coordination failure modes may emerge (see section 3 of Reid et al 2025).
- From multi-agent interactions, emergent behaviors can arise. An agent assessed as safe in isolation may contribute to harmful systemic outcomes when interacting with other agents (Hammond et al. 2025).
- Multi-agent systems allow adversaries to carry out attacks in a decentralized manner, enabling increased stealth in execution and limiting traceability. Emergent adversarial patterns make it difficult for oversight bodies to identify and hold specific entities accountable (de Witt 2025).
- Multi-agent systems also pose complex security challenges, as these systems can experience cascading compromises resulting in misaligned outcomes (Peigné et al. 2025). The spread of malicious prompts across agents working together as a system and the ability to evolve and improve as it hops between agents makes mitigation and detection particularly difficult for these systems (Ju et al. 2024, Gu et al. 2024, Lee and

Tiwari 2024, Peigné et al. 2025).

2.2 Security Practices for AI Agent Systems

(a) What technical controls, processes, and other practices could ensure or improve the security of AI agent systems in development and deployment? What is the maturity of these methods in research and in practice?

During development, measures such as secure scaffolding design (e.g., modular frameworks with input validation), threat modeling for agent-tool interactions, and red-teaming for autonomy escalations can be considered. While these are mature risk management practices for LLMs, they are in their early stages of adoption in agentic systems.

In deployment, layered controls can help enhance resilience.For post-deployment oversight, consider using agent identifiers to trace interactions across entities, enable attribution, and revoke compromised agents. Complement this effort with real-time monitoring to gain live insight on agent activities and configure automated alerts for certain activities or high-risk conditions, activity logs to automatically document (with timestamps) agent inputs, outputs, interactions, and scaffolding, and acceptable use policies (AUPs) to define permitted uses, prohibited activities, and operational constraints, with regular updates to address emerging risks and misuse patterns.

(b) To what degree, if any, could the effectiveness of technical controls, processes, and other practices vary with changes to model capability, agent scaffold software, tool use, deployment method (including internal vs. external deployment), use case, use in multi-agent systems, and otherwise?

The effectiveness of technical controls, processes, and other practices varies significantly with changes to aforementioned variables. For instance, basic input filtering might suffice for narrow models but will fail against advanced reasoning enabling multi-step jailbreaks. Monitoring efforts may work for read-only tools but degrades with execute privileges, similar to how single-agent controls might miss emergent multi-agent risks like collusion.

While more extensive testing simulations may help identify failure modes, they are not without their limitations. Emergent behaviors and long-term patterns are not always captured in abstract testing scenarios, and thus these technical controls need to account for redundancies in safeguards in high-capability or multi-agent setups.

(c) How might technical controls, processes, and other practices need to change, in response to the likely future evolution of AI agent system capabilities or of the threats, risks, or vulnerabilities facing them?

The following practices can help ensure proactive risk management as agentic AI systems evolve:

- **Adopt scalable oversight mechanisms**: As agents gain autonomy and multi-step reasoning, implement hierarchical human-AI supervision to verify outputs beyond human direct oversight
- **Leave a margin of safety**: Given uncertainties in assessing AI risks and the expanding scope of potential harms, thresholds should be set conservatively while remaining adaptable to new evidence and effective mitigation strategies.
- **Enhance dynamic threat modeling:** Integrate findings from red-teaming efforts, adversarial testing, with continuous monitoring to anticipate vulnerabilities like jailbreaking, deception, or emergent risks in multi agent setups.
- **Identify thresholds and design kill switches:** Require phased rollouts with sub-thresholds that determine go/no-go sections and audit trails to prevent uncontrolled scaling or deployment in high-stakes domains.

2.4 Limiting, Modifying, and Monitoring Deployment Environments

(d) What methods could be used to monitor deployment environments for security threats, risks, or vulnerabilities?

Oueslati & Staes-Polet (2025) suggest a four-pillar approach:

- **Agent identifiers** can be used to trace agent interactions with several entities. Decisions regarding which identifier to attach to the agent's output will depend on both the format and the content of the output (Chan et al., 2024).
    - For example, using watermarks or other types of embedded metadata as identifiers for images (this method however carries significant limitations owing to the relative ease with which adversarial actors can remove watermarks).
    - Consider attributing agent actions to entities by **identity binding** an agent to a real-world identity (e.g., a corporation or person) (Chan et al., 2025).
    - Agent cards (similar to system cards) may also be used to bring visibility to important information (Casper et al., 2025).
- **Real-time monitoring** can be used to gain live insight on agent activities and configure automated alerts for certain activities or high-risk conditions (Chan et al., 2024).
    - Track agent behavior with **real-time failure detection** methods, particularly for agents with high affordances performing high-stakes, non-reversible actions (Srikumar, 2025).
- **Activity logs** may also be used to automatically document (with timestamps) agent inputs, outputs, interactions, and scaffolding, providing insight into the agent's decision-making process. The amount of detail captured by the activity logs may be proportional to the perceived risk level.
- **Acceptable use policies (AUPs)** should explicitly define permitted uses, prohibited activities, and operational constraints, with regular updates to address emerging risks and misuse patterns.

Additionally, considering that agentic AI systems are unprecedented in their autonomy and potential impact, post-deployment monitoring must be complemented with mechanisms for **logging and reporting incidents and near-misses** to support collective learning about emerging risks.

2.5 Additional Considerations

(b) In which policy or practice areas is government collaboration with the AI ecosystem most urgent or most likely to lead to improvements in the state of security of AI agent systems today and into the future?

Government collaboration with academia, industry, civil society, and the AI ecosystem is most urgent in standardization, incident reporting, talent pipelines, and adaptive governance to secure agentic AI systems. For more on collaborative oversight, please see “Govern 2.1” in Madkour et al. (2026).

For example: an anonymized agent incident databases can combine the knowledge from industry reports, with the research and analysis of academia, and the oversight of civil society, enabling shared threat intelligence that aids in identifying elements such as tool misuse patterns or scaffold exploits. An approach like this would help prevent siloed learning and is modeled after CISA's cybersecurity reports.

References

Akbulut, C., Weidinger, L., Manzini, A., Gabriel I., & Rieser, V. (2024). All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* https://ojs.aaai.org/index.php/AIES/article/view/31613

Anthropic. (2025a). Agentic Misalignment: How LLMs Could be Insider Threats. *Anthropic.* https://www.anthropic.com/research/agentic-misalignment

Anthropic. (2025b). Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign. *Anthropic*. https://www.anthropic.com/news/disrupting-AI-espionage

Barrett, A. M., Jackson, K, Murphy, E. R., Madkour, N., & Newman, J. (2024). Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. *arXiv.* https://arxiv.org/pdf/2405.10986

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., … Zeng, Y. (2025). International AI Safety Report*. arXiv.* https://arxiv.org/abs/2501.17805

Bertrand, Q., Duque, J., Calvano, E., & Gidel, G.(2025). Self-Play Q-learners Can Provably Collude in the Iterated Prisoner's Dilemma. *arXiv.* https://arxiv.org/pdf/2312.08484 Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI Agents. *arXiv.* https://arxiv.org/pdf/2401.13138

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Hagen, M. V., Alberti, S., Chan, A., Sun, Q,, Gerovitch, M., Bau, D., Tegmark,. … M., Hadfield-Menell, D. (2024). Black-Box Access is Insufficient for Rigorous AI Audits. *arXiv.* https://arxiv.org/abs/2401.14446

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI Agents. *arXiv.* https://arxiv.org/pdf/2401.13138

Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M.(2025). Infrastructure for AI Agents. *arXiv.* https://arxiv.org/abs/2501.10114v1

Chin, Z. S. (2025). Dimensional Characterization and Pathway Modeling for Catastrophic AI Risks. *arXiv.* https://arxiv.org/pdf/2508.06411

Dawson, A., Mulla, R., Landers, N., & Caldwell, S. (2025). AIRTBench: Measuring Autonomous AI Red Teaming Capabilities in Language Models. *arXiv.* https://arxiv.org/abs/2506.14682

de Witt, C.S. (2025). Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. *arXiv.* https://arxiv.org/pdf/2505.02077 Enkrypt AI. (2025). Agent Risk Taxonomy. *Enkrypt AI.* https://cdn.prod.website-files.com/6690a78074d86ca0ad978007/687f7fac66e8127aa565341d_Agent%20Risk%20taxonomy_enkryptai.pdf

Gu, X., Zheng, X., Pang, T., Du, C., Liu, Q., Wang, Y., Jiang, J., Lin, M. (2024). Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast. *Proceedings of Machine Learning Research.* https://proceedings.mlr.press/v235/gu24e.html

Guidi, G., Dominici, F., Gilmour, J., Butler, K., Bell, E., Delaney, S., Bargagli-Stoffi , F. J. (2024). Environmental Burden of United States Data Centers in the Artificial Intelligence Era. *arXiv.* https://arxiv.org/pdf/2411.09786

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., Schroeder de Witt, C., Shah, N., Wellman, M., … Rahwan, I. (2025). Multi-Agent Risks from Advanced AI. *arXiv.* https://arxiv.org/abs/2502.14143

Heiding, F., Lermen, S., Kao, A., Schneier, B., & Vishwanath, A. (2024). Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects. *arXiv.* https://arxiv.org/abs/2412.00586

Ju, T., Wang, Y., Ma, X., Cheng, P., Zhao, H., Wang, Y., Liu, L., Xie, J., Zhang, Z., & Liu, G. (2024). Flooding Spread of Manipulated Knowledge in LLM-Based Multi-Agent Communities. *arXiv.* https://arxiv.org/abs/2407.07791 Kouremetis, M. (2025). Evaluating Offensive Cyber Agents: Kerberoasting. *Dreadnote.* https://dreadnode.io/blog/evaluating-offensive-cyber-agents-kerberoasting#key-takeaways

Lee, D., & Tiwari, M. (2024). Prompt Infection: LLM-to-LLM Prompt Injection within Multi-Agent Systems. *arXiv.* https://arxiv.org/abs/2410.07283

Lin, J. W., Jones, E. K., Jasper, D. J., Ho, E. J., Wu, A, Yang, A. T., Perry, N., Zou, A., Fredrikson, M., Kolter, J. Z., Liang, P., Boneh, D., & Ho, D. E. (2025). Comparing AI Agents to Cybersecurity Professionals in Real-World Penetration Testing. *arXiv.* https://arxiv.org/abs/2512.09882

Luccioni, A. S., Jernite Y., & Strubell, E. (2024). Power Hungry Processing: Watts Driving the Cost of AI Deployment? *arXiv.* https://arxiv.org/pdf/2311.16863

Madkour, N., Newman, J., Murphy, E. R., Jackson, K., Raman, R., & Yuan, C., & Hendrycks, D. (2026). Agentic AI Risk-Management Standards Profile. *UC Berkeley Center for Long-Term Cybersecurity*. https://cltc.berkeley.edu/wp-content/uploads/2026/02/Agentic-AI-Risk-Management-Standards-Profile.pdf

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2025). Frontier Models are Capable of In-context Scheming. *arXiv.* https://arxiv.org/pdf/2412.04984 ORF. (2024). Issue Brief. Issue No. 768 December 2024. *Observer Research Foundation.* https://www.orfonline.org/public/uploads/posts/pdf/20241227111028.pdf

Mukherjee, A., & Chang, H. H. (2025). Agentic AI: Autonomy, Accountability, and the Algorithmic Society. *arXiv.* https://arxiv.org/pdf/2502.00289 Oueslati, A. & Staes-Polet, R. (2025). Ahead of the Curve: Governing AI Agents Under the EU AI Act. *The Future Society.* https://thefuturesociety.org/wp-content/uploads/2023/04/Report-Ahead-of-the-Curve-Governing-AI-Agents-Under-the-EU-AI-Act-4-June-2025.pdf

Oueslati, A. & Staes-Polet, R. (2025). Ahead of the Curve: Governing AI Agents Under the EU AI Act. *The Future Society.* https://thefuturesociety.org/wp-content/uploads/2023/04/Report-Ahead-of-the-Curve-Governing-AI-Agents-Under-the-EU-AI-Act-4-June-2025.pdf

Peigné, P., Kniejski, M., Sondej, F., David, M., Hoelscher-Obermaier, J., Schroeder de Witt, C., & Kran, E. (2025). Multi-Agent Security Tax: Trading Off Security and Collaboration Capabilities in Multi-Agent Systems. *Proceedings of the AAAI Conference on Artificial Intelligence.* https://ojs.aaai.org/index.php/AAAI/article/view/34970

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Abi Raad,M., Webson, A., Ho, L., Lin, S.,Farquhar, S., Hutter, M., Deletang, G., Ruoss, A., El-Sayed, S., Brown, S., Dragan, A., Shah, R., Dafoe, A., & Shevlane, T. (2024). Evaluating Frontier Models for Dangerous Capabilities. *arXiv.* https://arxiv.org/pdf/2403.13793

Raman, D., Madkour, N., Murphy, E. R., Jackson, K., & Newman, J. (2025). Intolerable Risk Threshold Recommendations for Artificial Intelligence. *arXiv.* https://arxiv.org/abs/2503.05812

Raza, S., Sapkota, R., Karkee, M., &Emmanouilidis, C. (2025). TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems. *arXiv.* https://arxiv.org/pdf/2506.04133

Reid, A., O'Callaghan, S., Carroll, L., & Caetano, T. (2025). Risk Analysis Techniques for Governed LLM-based Multi-Agent Systems. *arXiv.* https://arxiv.org/abs/2508.05687

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2026). AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges. *ScienceDirect.* https://www.sciencedirect.com/science/article/pii/S1566253525006712

Shao, M., Rani, N., Milner, K., Xi, H., Udeshi, M., Aggarwal, S., Putrevu, V. S. C., Shukla, S. K., Krishnamurthy,P., Khorrami, F., Karri, R., & Shafique, M. (2025). Towards Effective Offensive Security LLM Agents: Hyperparameter Tuning, LLM as a Judge, and a Lightweight CTF Benchmark. *arXiv.* https://arxiv.org/abs/2508.05674

Sharma, G., Kulkarni, V., King, M., & Huang, K. (2025). Towards Unifying Quantitative Security Benchmarking for Multi Agent Systems. *arXiv.* https://arxiv.org/pdf/2507.21146

Singer, B., Lucas, K., Adiga, L., Jain, M., Bauer, L., & Sekar, V. (2025). Incalmo: An Autonomous LLM-assisted System for Red Teaming Multi-Host Networks. *arXiv.* https://doi.org/10.48550/arXiv.2501.16466

Srikumar, M. (2025). Prioritizing Real-Time Failure Detection in AI Agents. *Partnership on AI.* https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents/