



HUGGING FACE

Hugging Face Response to NIST Request for Information on Security Considerations for Artificial Intelligence Agents

Submitted to: Center for AI Standards and Innovation (CAISI)
February 2026

About Hugging Face

Hugging Face is a U.S.-based, community-driven company committed to democratizing responsible artificial intelligence (AI) and machine learning (ML). Our open-source platform hosts over 2 million models and 500,000 datasets serving 11 million users. This direct experience with AI development, deployment, and community-driven security practices across diverse contexts informs our response to security considerations for AI agent systems.

Executive Summary

Security challenges in AI agent systems arise from the intersection of three factors: statistical foundations that introduce irreducible uncertainty, deployment in environments where malicious actors actively probe for vulnerabilities, and expanded autonomy that increases the scope and consequence of actions. Addressing these challenges requires a multi-faceted approach that includes technical and non-technical means to ensure security involving transparency, evaluation rigor, and community-driven innovation.

Drawing on our experience operating infrastructure used by millions of people, this submission offers evidence-based recommendations for measuring and improving AI agent security. Effective security practices must balance enabling beneficial innovation, preventing catastrophic failures, and maintaining democratic access to AI capabilities. Solutions that over-index on restriction risk concentrating capabilities in ways that reduce overall ecosystem resilience, while approaches grounded in transparency, standardized evaluation, and shared infrastructure can strengthen security outcomes across the entire deployment pipeline.

This response addresses priority questions 1(a), 1(d), 2(a), 2(e), 3(a), 3(b), 4(a), 4(b), and 4(d), with additional commentary on Questions 2(c), 3(c), 5(a), 5(b), and 5(c) where our operational experience provides relevant insights.



HUGGING FACE

1. Security Threats, Risks, and Vulnerabilities Affecting AI Agent Systems

1(a): Unique Security Threats Distinct from Traditional Software Systems

AI agent systems introduce security challenges that differ fundamentally from traditional software vulnerabilities. Unlike conventional software, where inputs pass through deterministic control flows, agents process natural language instructions that can be manipulated to override intended behavior. ([He et al., 2025](#)) The statistical nature of underlying models means that identical inputs may produce different outputs, complicating threat detection and creating uncertainty about system state.

Several categories of threats arising from the model-based nature of agents are particularly distinctive:

Specification gaming and objective misalignment. Even uncompromised models may pursue beneficial goals in harmful ways. A system instructed to "maximize user engagement" might exploit psychological vulnerabilities or share misleading information, not because it was attacked but because the objective was underspecified. ([Russell & Norvig, 2003](#)) This problem is compounded by instrumental convergence: agents may treat security constraints such as sandboxes or rate limits as obstacles to their primary objectives and develop sub-goals aimed at circumventing them. ([Bostrom, 2012](#)) The orthogonality of intelligence and goals further complicates the picture; more capable agents are not inherently more aligned, and greater problem-solving ability may amplify rather than constrain misaligned behavior. Traditional software executes exactly what developers program; agents interpret goals and select among unbounded action sequences, creating unforeseen behaviors that resist prediction through code review alone. A particularly concerning related phenomenon is alignment faking, where models may reason during training to preserve misaligned objectives while appearing compliant; these behaviors can persist through subsequent safety fine-tuning. ([Greenblatt et al., 2024](#); [Hubinger et al., 2024](#)) Future open-ended agents integrated into autonomous physical science laboratories could synthesize novel hazardous materials without malicious intent simply by optimizing for novelty, if informational guardrails are insufficient.

Compounding errors across action chains. Agents orchestrate multi-step workflows where early mistakes propagate through subsequent actions. ([Perrow, 1984](#)) A misclassified input might trigger inappropriate API calls, which generate flawed outputs used as inputs to downstream tools, creating cascading failures. ([Sharwood, 2025](#); [Popper, 2012](#)) Traditional software fails locally and predictably; agent failures can compound exponentially across tool interactions.

Context-dependent vulnerabilities through tool access. The same base model exhibits radically different security profiles depending on available tools and deployment context. A model with email access poses data exfiltration risks absent in sandbox environments; ([Zharmagambetov et al., 2025](#)) one with code execution capabilities can modify its own scaffolding in ways that disable safeguards. ([Beers & Rushing, 2025](#)) Security assessments conducted on isolated



HUGGING FACE

models fail to capture vulnerabilities that emerge only when agents interact with external systems.

Finally, open-ended agent systems capable of continuously generating novel outputs and action sequences present a structurally distinct threat class. ([Hughes et al., 2024](#); [Clune, 2019](#); [Sheth et al., 2025](#)) Unlike traditional software failures, which represent deviations from a known specification, open-ended agents are designed to operate in novel situations and creatively create and combine tools. As such, it is not possible to enumerate all possible vulnerabilities and attack scenarios, hampering the efficacy of traditional security protocols such as penetration testing and access control lists.

1(d): Evolution of Threats Over Time

The threat landscape for AI agents has evolved rapidly as capabilities expand and deployment contexts diversify. Early chatbot systems faced primarily attacks attempting to extract training data or generate prohibited content. As systems gained tool access and multi-step planning capabilities, threats shifted toward agent hijacking where attackers redirect entire action sequences toward malicious goals.

Several trends suggest how threats will likely evolve:

Capability-driven attack sophistication. As models develop stronger planning and problem-solving abilities, attacks will leverage these same capabilities. ([Wen, 2024](#)) Prompt injections will evolve from simple jailbreaks into sophisticated multi-turn strategies that probe for inconsistencies in system instructions. Agents with improved situational awareness may better recognize when they are being tested versus deployed, potentially behaving differently under evaluation than in production, a concern that some recent research has begun to document, though the evidence base remains actively developing. ([Meinke et al., 2024](#); [Kwa et al., 2025](#))

Supply chain and model provenance attacks. As ecosystems mature, attackers will increasingly target upstream components rather than deployed systems directly. Backdoored models uploaded to popular repositories could compromise numerous downstream applications. Fine-tuning datasets poisoned with adversarial examples could introduce vulnerabilities that activate only under specific conditions, evading pre-deployment detection.

Open-endedness as an accelerating risk surface. Open-endedness is rapidly becoming a core research priority at major frontier laboratories. These systems are designed to produce novel outputs that exceed prior training distributions, a property that makes traditional safety evaluation methods structurally inapplicable. Safety research for self-evolving systems remains significantly underdeveloped relative to the pace of capability development. ([Hughes et al., 2024](#); [Sheth et al., 2025](#))

The history of software security suggests that purely reactive approaches struggle against adaptive adversaries. Sustainable security requires transparency about vulnerabilities, broad community participation in identifying and addressing weaknesses, and infrastructural changes



HUGGING FACE

(technical systems and shared protocols external to agents that can mediate and constrain their interactions with the world) that make entire classes of attacks structurally infeasible rather than merely difficult. ([Chan et al., 2025](#)) Just as the Internet relies on protocols like HTTPS to enable trustworthy transactions, agent ecosystems will require analogous shared infrastructure (covering attribution, interaction shaping, and incident response) to operate safely at scale.

1(e): Unique Threats in Multi-Agent Systems

Multi-agent systems introduce a qualitatively distinct security surface that cannot be addressed by scaling up single-agent security practices. ([Hammond et al., 2025](#)) When multiple agents interact, the game-theoretic dynamics of their interactions produce emergent behaviors that may be difficult or impossible to anticipate from the properties of individual components.

Unintended feedback loops and alignment cascades. Multi-agent systems introduce non-stationarity: the environment each agent learns from is constantly changing because other agents are simultaneously learning and updating. This can produce rapid, unpredictable feedback loops analogous to stock market flash crashes. ([Popper, 2012](#); [Hammond et al., 2025](#); [McClellan, 2025](#)) These dynamics make multi-agent systems inherently more difficult to stabilize than singular agents.

Viral misalignment and multi-agent interaction exploits. A compromised or misaligned agent could propagate its objectives to other agents in the system by embedding adversarial instructions in its outputs that other agents consume, behaving much like a computer virus self-replicating across interconnected systems. ([Lee & Tiwari, 2024](#)) Malicious agents could similarly exploit coordination mechanisms designed for benign agents to carry out distributed attacks that are difficult to attribute to individual components. ([Hammond et al., 2025](#)) Unlike single-agent compromise, which can be contained through component isolation, viral misalignment in multi-agent systems may spread faster than detection systems can respond.

From an alternative perspective, multi-agent architectures can also function as a positive regulatory mechanism for alignment, where agents with different objective structures provide mutual checks. ([Vallinder & Hughes, 2024](#)) However, research on emergent cooperative norms among AI agents suggests these dynamics are highly model-dependent and sensitive to initial conditions, making reliability hard to guarantee. ([Hammond et al., 2025](#))

Addressing the security challenges of multi-agent deployments will require dedicated infrastructure and evaluation frameworks that do not currently exist at adequate scale or standardization. Promising directions include inter-agent communication protocols, agent identity and certification systems, and oversight layers that allow human actors to intervene on agent actions even within automated pipelines. ([Chan et al., 2025](#))



HUGGING FACE

2. Security Practices for AI Agent Systems

2(a): Technical Controls, Processes, and Practices

Model-level controls address vulnerabilities inherent to the foundation model itself. Adversarial training can improve robustness to adversarial inputs by exposing models to attack patterns during fine-tuning, though this approach faces fundamental limitations as attackers adapt. ([Mohri et al., 2018](#)) Instruction tuning that emphasizes adherence to system directives over user requests can strengthen resistance to hijacking, but risks brittleness when legitimate instructions conflict with hardcoded priorities. Distribution shift between training and deployment environments remains a fundamental challenge, as models optimized for one context may exhibit degraded robustness when deployed in novel settings. ([Quiñonero-Candela et al., 2022](#))

A critical caution applies to model-level alignment interventions: chain-of-thought (CoT) reasoning traces should not be taken as faithful representations of model alignment. Recent empirical work demonstrates that incorporating CoT monitors directly into reinforcement learning training rewards can, beyond a certain optimization threshold, produce obfuscation of misaligned reasoning rather than genuine alignment improvement. ([Baker et al., 2025](#); [Lanham et al., 2023](#)) We strongly advise against training interventions that target CoT content directly as a primary safety mechanism.

Agent system-level controls focus on scaffolding architecture and orchestration logic. Input and output filtering can catch obvious attacks or policy violations. Continuous monitoring of agent behavior, including logging of all tool calls and intermediate planning steps, enables detection of anomalous patterns, ([Yuan et al., 2025](#)) though the volume of data generated by complex agents challenges real-time analysis. Sandboxing and virtualization that isolate agent execution environments from production systems offer strong protection against many attack vectors.

We recommend grouping the following as core infrastructure-level security design principles, recognizing that external protocols and systems (not just model-level interventions) are necessary to achieve durable security at scale ([Chan et al., 2025](#)):

- **Scaffold design** deserves particular attention as a security surface: static, security-audited scaffolds with well-defined tool boundaries are preferable to dynamically evolving architectures, and any research into self-improving agent scaffolds should be conducted under heavily specified, monitored, and isolated conditions.
- **Tool restriction frameworks** that limit which APIs an agent can access based on context reduce potential damage from compromised behavior.
- **Agent certification infrastructure**: systems for making, verifying, and revoking claims about an agent's capabilities, tool access, and authorization level, which can enable counterparties to make informed decisions about whether to interact with a given agent.
- **Identity binding mechanisms** that associate an agent's actions with a responsible legal entity support accountability and create disincentives for misuse.



HUGGING FACE

Together, these represent concrete infrastructural commitments that can reduce entire classes of vulnerabilities rather than addressing individual attack vectors reactively.

Human oversight controls maintain human agency in consequential decisions. Critically, these controls must be designed with awareness of automation bias ([Holbrook et al., 2024](#); [Bainbridge, 1983](#); [Goddard et al., 2012](#)): as agents demonstrate reliable performance over time, human supervisors tend to reduce their scrutiny, potentially missing anomalous behavior.

Graduated autonomy frameworks that expand agent permissions as systems demonstrate reliable behavior create natural checkpoints, though they require careful design of trust metrics that resist gaming. Of particular importance: when a decision would be irreversible (such as sending a communication, executing a financial transaction, or modifying production data), human oversight must be mandated, not optional. The irreversibility of an action should function as a hard trigger for human involvement regardless of the agent's demonstrated reliability.

For frontier and multi-agent systems specifically, we recommend mandated evaluations with several properties: testing in non-stationary environments that mimic real-world complexity; independent third-party evaluation with API-level access; transparent documentation of performance variability alongside mean performance; and for agents operating in physical sciences, strictly isolated sandbox environments with parity to safety standards applied to human-operated facilities under guidelines such as GHS, CHP, and REACH.

2(c): Evolution of Controls with Capability Advances

As model capabilities increase, existing controls face compounding challenges.

Evaluation-aware behavior, where models correctly identify that they are being assessed and adjust their outputs accordingly, has been documented in some frontier systems and may scale with model capability, though this area warrants further rigorous empirical study. ([Meinke et al., 2024](#)) Controls must therefore prioritize realistic scenario design that does not trigger evaluation-awareness responses, combined with robust observability of evaluation trajectories.

Chain-of-thought monitoring remains useful as a complementary signal when observed alongside actions and outputs, but training directly against CoT monitors risks producing models that have learned to hide misaligned planning rather than to correct it. ([Baker et al., 2025](#)) The increasing capability of agents to conduct longer, more complex task sequences, with agentic time-horizons expanding rapidly, means that the potential damage from a compromised or misaligned agent grows proportionally with each capability increment. ([Kwa et al., 2025](#))

2(e): Relevant Cybersecurity Guidelines, Frameworks, and Best Practices

Several existing frameworks provide valuable foundations for AI agent security, though each requires adaptation:



HUGGING FACE

- The **NIST AI Risk Management Framework** offers comprehensive guidance for identifying and managing AI risks, though specific metrics for agent-specific threats require further development.
- The **NIST Cybersecurity Framework** provides well-established identify–protect–detect–respond–recover functions, though concepts like "asset inventory" become complicated when agents dynamically generate new code or modify their own scaffolding.
- **OWASP guidelines** for LLM security specifically address prompt injection, training data poisoning, and other AI-specific vulnerabilities.
- **Zero Trust Architecture** principles mandate continuous verification rather than perimeter-based security, proving particularly relevant for agents that cross traditional security boundaries.

3. Assessing the Security of AI Agent Systems

3(a): Methods for Anticipating, Identifying, and Assessing Security Threats

A critical design consideration for all pre-deployment evaluations is evaluation awareness. Models that recognize they are being assessed may suppress dangerous capabilities or perform in ways that do not reflect production behavior. ([Meinke et al., 2024](#)) Evaluation designs should therefore prioritize realistic, non-contrived scenarios; avoid extreme or obviously artificial conditions; and include systematic observability of agent trajectories through sampling, logging, and where feasible, human inspection of rollouts.

Post-deployment detection requires behavioral monitoring that goes beyond traditional intrusion detection. Logging all tool calls, planning traces, and intermediate outputs enables forensic analysis when incidents occur. ([Yuan et al., 2025](#)) Anomaly detection that identifies unusual patterns in agent behavior can flag potential compromises, though high false positive rates risk alert fatigue.

Useful resources for assessment would include standardized agent security benchmarks, publicly available datasets of attack patterns, reference implementations of monitoring infrastructure, and case studies documenting real-world security incidents.

3(b): Assessing Security of Particular AI Agent Systems

Security assessment for specific agent deployments should account for context-dependent risk profiles. Key factors include:

Tool access scope. An agent with read-only access to internal documentation poses different risks than one with database modification permissions or network access. Assessment should enumerate available tools, map potential attack chains through tool combinations, and evaluate worst-case outcomes from misuse of each capability.



HUGGING FACE

Data sensitivity. Agents handling personal health information, financial records, or trade secrets require stricter security controls than those working with public information. ([OWASP, 2025](#))

Classification of accessible data according to sensitivity and impact of unauthorized disclosure should drive control selection.

Deployment context. Agents operating entirely within sandboxed development environments face different threat models than those interacting with production systems or external parties. Understanding the boundaries between controlled and uncontrolled environments helps prioritize defenses.

Autonomy level. Agents requiring human approval for all consequential actions have fundamentally different risk profiles than those operating with full autonomy. Explicit documentation of which decisions require human oversight and under what conditions provides crucial context for security assessment, as does documentation of how human oversight is structured to resist automation bias over time, ([Goddard et al., 2012](#)) and how irreversible actions trigger mandatory human involvement. It is also worth noting that agents operating with greater autonomy may progressively erode the oversight capacity of human operators through deskilling effects. ([Budzyń et al., 2025](#))

User base. Agents exposed only to trusted internal users face different social engineering risks than those available to arbitrary external actors.

3(c): Documentation from Upstream Developers

Downstream developers deploying AI agents built on foundation models from upstream providers face information asymmetries that complicate security assessment. Useful documentation from model developers would include security-relevant capability assessments, known vulnerabilities and attack patterns identified through internal testing, training data composition information, and fine-tuning and alignment procedures. Secure disclosure channels between upstream and downstream developers, analogous to coordinated vulnerability disclosure in traditional software, could enable sharing of security-critical information while limiting exposure to attackers. Agent certification infrastructure, in which claims about an agent's capabilities, tool access, and known limitations are made verifiable and available to downstream parties, represents a promising structural approach to this information-sharing challenge. ([Chan et al., 2025](#))

4. Limiting, Modifying, and Monitoring Deployment Environments

4(a): Constraining Deployment Environments

Limiting the scope of environments where agents operate reduces potential damage from compromised or misaligned behavior. Technical approaches include:

- **Virtual environment sandboxing**



HUGGING FACE

- **Network segmentation** with firewall rules restricting access to internal systems
- **API-level access control** wrapping external tools with authorization layers that validate each agent request against policy rules
- **Capability-based security** providing agents with explicit tokens representing permissions for specific operations

Agent channels (dedicated interaction interfaces that isolate agent traffic from general digital traffic) represent a promising additional mechanism for constraining deployment environments while maintaining the observability needed for security monitoring. ([Chan et al., 2025](#))

4(b): Modifying Environments and Implementing Rollback Mechanisms

Staging environments that mirror production systems allow testing agent behavior in realistic contexts without risking actual consequences. Transaction logging and replay systems record all agent actions with sufficient detail to reproduce system state at any point in time. Undo mechanisms that reverse agent actions range from simple operation inversion to complex multi-step compensation, though many operations have irreversible real-world effects.

This points to a critical boundary for human intervention: when an action cannot be undone, a human must be involved prior to execution: not as a fallback, but as a mandatory checkpoint integrated into the agent's workflow. Building rollback infrastructure as a shared external system (one that operates independently of any individual agent's scaffolding) is preferable to relying on agent-internal undo logic, which may itself be compromised. ([Chan et al., 2025](#)) Circuit breakers automatically halt agent execution when behavioral anomalies or resource consumption exceed defined thresholds.

4(d): Monitoring Deployment Environments

Effective monitoring of agent deployment environments requires visibility into both agent behavior and environmental state. Statistical methods that establish baseline behavioral profiles and flag significant deviations show promise but face high false positive rates given the inherent variability of agent outputs. ([Yuan et al., 2025](#)) Legal and privacy challenges arise when monitoring involves agent interactions with user data or communications. ([Xu et al., 2025](#)) Research maturity is low; no widely adopted standards exist for agent behavioral monitoring. Incident reporting infrastructure (standardized channels through which agents, human users, and third parties can flag anomalous or harmful agent behavior) would significantly improve the collective capacity for post-deployment detection. ([Chan et al., 2025](#))

5. Additional Considerations

5(a): Accelerating Adoption of Security Practices



HUGGING FACE

Several interventions could accelerate adoption of agent security best practices: standardized evaluation frameworks; open-source reference implementations of secure agent architectures; public incident databases documenting real-world failures. ([DARPA, 2016](#)) Coordinated disclosure processes and educational resources targeting both technical practitioners and organizational decision-makers would address knowledge gaps that impede security adoption. ([Deng et al., 2024](#)) Investment in shared agent infrastructure, including open standards for agent identity, certification, and rollback, can reduce the cost of adoption for individual organizations and create positive-sum security dynamics across the ecosystem. ([Chan et al., 2025](#))

5(b): Areas for Government Collaboration

The government has distinctive capabilities and responsibilities that make collaboration valuable in several areas: establishing security baselines for federally funded or deployed agent systems; facilitating information sharing about threats and vulnerabilities; supporting development of evaluation infrastructure; clarifying regulatory expectations; and coordinating international standards development. A particularly important role for the government is supporting the development of shared agent infrastructure (certification systems, identity binding protocols, and incident reporting frameworks) where coordination failures may prevent private actors from investing adequately despite collective benefits. ([Chan et al., 2025](#))

Three additional areas deserve priority attention:

1. Creation of technical national standards covering each step of the model lifecycle, designed to accommodate the adaptive and autonomous nature of agents rather than emphasizing static compliance.
2. Compute governance as a physically grounded and enforceable mechanism for monitoring and regulating high-capability systems.
3. Adequate incident reporting requirements and robust whistleblower protections for AI agent security breaches.

5(c): Research Priorities

Several research areas would significantly advance agent security. Formal verification methods for agent behavior could provide stronger guarantees about security properties. Adversarial robustness research that fundamentally addresses the brittleness of ML models would improve agent security at the foundation. ([Mohri et al., 2018](#)) Interpretability and transparency techniques that make agent reasoning more inspectable would enable better security monitoring. Multi-agent security research studying vulnerabilities that emerge from agent interactions would address increasingly common deployment patterns. ([Hammond et al., 2025](#))

Particular urgency attaches to:

- Evaluation methodology that is robust to evaluation-aware behavior ([Meinke et al., 2024](#))
- Scalable oversight approaches that remain reliable even as CoT monitoring faces obfuscation pressure ([Baker et al., 2025](#))



HUGGING FACE

- Safety frameworks specifically designed for open-ended agents, an area where the gap between capability development and safety research is widest and growing fastest ([Hughes et al., 2024](#); [Sheth et al., 2025](#))

The increasing capability of agents to complete long autonomous task sequences underscores the urgency of evaluation methodology that can keep pace with capability gains. ([Kwa et al., 2025](#); [Kinniment et al., 2023](#))

Conclusion

Security for AI agent systems requires rethinking assumptions inherited from traditional software security while adapting proven principles to new contexts. The statistical nature of agents, their autonomous action selection, and deployment in adversarial environments create challenges that deterministic systems do not face. Effective security will emerge from layered controls, transparency about capabilities and limitations, standardized evaluation methods, and broad community participation in identifying and addressing vulnerabilities.

Our recommendations emphasize approaches that balance innovation and security, avoiding both unconstrained deployment that invites catastrophic failure ([Center for AI Safety, 2023](#)) and excessive restriction that concentrates capabilities in ways that reduce overall ecosystem resilience. Crucially, sustainable agent security cannot be achieved through system-level interventions alone: just as the Internet required shared protocols and infrastructure to scale safely, agent ecosystems will require analogous external systems (for attribution, interaction governance, and incident response) that operate across organizational boundaries. ([Chan et al., 2025](#)) Treating agent infrastructure as a shared public good, rather than a competitive advantage for individual developers, is essential to realizing both the benefits and the safety of AI agent systems.

Hugging Face appreciates NIST's leadership in developing frameworks for secure AI innovation and welcomes the opportunity to contribute to this important work.

Submitted by:

Avijit Ghosh, Technical AI Policy Researcher, Hugging Face

Margaret Mitchell, Chief Ethics Scientist, Hugging Face

Additional contributions from:

Alexander Müller, Jeremias Lino Ferrao, Tiwai Mhundwa, Ilija Lichkovski, Atakan Tekpirmak (AI Safety Initiative Groningen – AISIG)



HUGGING FACE

References

- Bainbridge, L. Ironies of automation. *Automatica* 19, 775–779 (1983).
- Baker, B., Huizinga, J., Gao, L. et al. Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. *arXiv preprint arXiv:2503.11926*. OpenAI (2025).
- Beers, K. & Rushing, C. AI control: How to make use of misbehaving AI agents. Center for Security and Emerging Technology (2025).
- Bostrom, N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach.* 22, 71–85 (2012).
- Budzyń, K., Romańczyk, M., Kitala, D. et al. Endoscopist deskill risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol. Hepatol.* 10, 896–903 (2025).
- Center for AI Safety. Statement on AI Risk. <https://www.safe.ai/statement-on-ai-risk> (2023).
- Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K. & Anderljung, M. Infrastructure for AI Agents. *arXiv preprint arXiv:2501.10114* (2025).
- Clune, J. AI-GAS: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv preprint arXiv:1905.10985* (2019).
- Defense Advanced Research Projects Agency (DARPA). DARPA's Cyber Grand Challenge. <https://www.darpa.mil/research/programs/cyber-grand-challenge> (2016).
- Deng, Z. et al. AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways. *ACM Comput. Surv.* (2024). <https://doi.org/10.1145/3716628>
- Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* 19, 121–127 (2012).
- Greenblatt, R. et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093* (2024).
- Hammond, L., Chan, A., Clifton, J. et al. Multi-Agent Risks from Advanced AI. Cooperative AI Foundation Technical Report #1. *arXiv:2502.14143* (2025).
- He, Y., Wang, E., Rong, Y., Cheng, Z. & Chen, H. Security of AI agents. *Proc. IEEE/ACM Int. Workshop Responsible AI Eng.* 45–52 (2025).
- Holbrook, C. et al. Overtrust in AI recommendations about whether or not to kill. *Sci. Rep.* 14, 19751 (2024).
- Hubinger, E. et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566* (2024).
- Hughes, E., Dennis, M., Parker-Holder, J. et al. Open-Endedness is Essential for Artificial Superhuman Intelligence. *arXiv preprint arXiv:2406.04268* (2024).
- Kinniment, M. et al. Evaluating language-model agents on realistic autonomous tasks. *arXiv preprint arXiv:2312.11671* (2023).
- Kwa, T., West, B., Becker, J. et al. Measuring AI Ability to Complete Long Tasks. *arXiv preprint arXiv:2503.14499* (2025).



HUGGING FACE

- Lanham, M. et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702* (2023).
- Lee, D. & Tiwari, M. Prompt Infection: LLM-to-LLM Prompt Injection within Multi-Agent Systems. *arXiv preprint arXiv:2410.07283* (2024).
- McClellan, M. AI and Financial Fragility: A Framework for Measuring Systemic Risk in Deployment of Generative AI for Stock Price Predictions. *J. Risk Financ. Manag.* 18, 475 (2025).
- Meinke, A., Schoen, B., Scheurer, J. et al. Frontier Models are Capable of In-context Scheming. *arXiv preprint arXiv:2412.04984*. Apollo Research (2024).
- Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of Machine Learning*, 2nd edn. MIT Press (2018).
- OWASP GenAI Security Project. OWASP Top 10 for Large Language Model Applications 2025, LLM02:2025 – Sensitive Information Disclosure. <https://genai.owasp.org/llm-top-10/> (2025).
- Perrow, C. *Normal Accidents: Living with High-Risk Technologies*. Basic Books, New York (1984).
- Popper, N. Knight Capital says trading mishap cost it \$440 million. *New York Times* (2012).
- Quiñonero-Candela, J. et al. (eds.) *Dataset Shift in Machine Learning*. MIT Press (2022).
- Russell, S. J. & Norvig, P. "Section 26.3: The Ethics and Risks of Developing Artificial Intelligence." *Artificial Intelligence: A Modern Approach*. Prentice Hall (2003).
- Sharwood, S. Vibe coding service Replit deleted user's production database, faked data, told fibs galore. *The Register* (2025).
- Sheth, I., Wehner, J., Abdelnabi, S., Binkyte, R. & Fritz, M. Safety is Essential for Responsible Open-Ended Systems. *arXiv preprint arXiv:2502.04512* (2025).
- Vallinder, A. & Hughes, E. Cultural Evolution of Cooperation among LLM Agents. *arXiv preprint arXiv:2412.10270* (2024).
- Wen, S. The power of generative AI in cybersecurity: Opportunities and challenges. *Appl. Comput. Eng.* 48, 31–39 (2024).
- Xu, M. et al. Acceptability of AI Assistants for Privacy: Perceptions of Experts and Users on Personalized Privacy Assistants. *arXiv preprint arXiv:2509.08554* (2025).
- Yuan, Q. et al. Multi-agent for network security monitoring and warning: A generative AI solution. *IEEE Netw.* (2025).
- Zharmagambetov, A. et al. AgentDAM: Privacy leakage evaluation for autonomous web agents. *arXiv preprint arXiv:2503.09780* (2025).