UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

Agentic AI Risk-Management Standards Profile

NADA MADKOUR | JESSICA NEWMAN | DEEPIKA RAMAN | KRYSTAL JACKSON
EVAN R. MURPHY | CHARLOTTE YUAN

Agentic AI Risk-Management Standards Profile

NADA MADKOUR[†] • **JESSICA NEWMAN**[†] • **DEEPIKA RAMAN**[†] • **KRYSTAL JACKSON**[†]
EVAN R. MURPHY[†] • **CHARLOTTE YUAN**[†]

† AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley
All affiliations listed are either current, or were during main contributions to this work or a previous version.

Version 1.0, February 2026

For the latest public version of this document, see:
https://cltc.berkeley.edu/publication/agentic-ai-risk-management-standards-profile

For the full General-Purpose AI (GPAI) Risk-Management Standards Profile, Version 1.2, and other supporting documents, see:
https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1.2/

Contents

Executive Summary

AI systems that use reasoning to autonomously pursue goals through interaction with external environments and tools — referred to hereafter as "AI agents" or "agentic AI" — promise transformative benefits for productivity and complex problem-solving. However, the ability of AI agents to operate with increased autonomy also introduces significant risks, such as unintended goal pursuit, unauthorized privilege escalation or resource acquisition, and other behaviors — such as self-replication or resistance to shutdown — that could result in systemic or catastrophic harm. The unique challenges introduced by agentic capabilities complicate traditional, model-centric risk-management approaches and demand system-level governance that accounts for autonomy, authority, tool access, environment, and interaction effects.

This paper introduces the Agentic AI Risk-Management Standards Profile ("Agentic AI Profile"), which aims to provide a targeted set of practices and controls for identifying, analyzing, and mitigating risks specific to agentic AI. The Agentic AI Profile is designed to complement the NIST AI Risk Management Framework (AI RMF) (NIST, 2023a) and functions as a specialized extension of the UC Berkeley General-Purpose AI Risk-Management Standards Profile ("GPAI Profile"). While the GPAI Profile focuses on the risks inherent to large-scale models, the Agentic AI Profile addresses the risks that emerge when AI-based systems are granted the agency to act on behalf of users. It also draws on a growing body of technical, policy, and security research on AI agents, autonomy, and AI control.

The Agentic AI Profile is primarily for use by **developers and deployers of agentic AI systems**, including both single-agent and multi-agent systems built on general-purpose and domain-specific models. **Policymakers, evaluators, and regulators** can also use the Agentic AI Profile to assess whether agentic AI systems have been designed, evaluated, and deployed in line with leading risk-management practices.

The guidance in the Agentic AI Profile is organized around the four core functions of the NIST AI RMF: Govern, Map, Measure, and Manage (NIST, 2023a).[1] Sub-categories (e.g., Map 1.1) have been selected based on two criteria: (1) they are considered "high-priority sub-categories" in the GPAI Profile or (2) they require additional consideration beyond the content of the GPAI Profile (Madkour et al., 2026). Users of the Agentic AI Profile should continue to prioritize the

1 Govern: for AI risk management process policies, roles, and responsibilities; Map: for identifying AI risks in context; Measure: for rating AI trustworthiness characteristics; and Manage: for decisions on prioritizing, avoiding, mitigating, or accepting AI risks.

high-priority risk-management sub-categories established in the foundational GPAI Profile (Madkour et al., 2026).

This document provides the necessary context to apply those steps — such as go/no-go decisions (Manage 1.1) and risk-reduction controls (Manage 1.3) — with considerations around agentic systems' unique vulnerabilities and capabilities, such as autonomous decision-making.[2] Agentic AI and AI-agent risk topics and corresponding guidance sections in this Profile (Map 1.1 and Manage 1.3) include the following:

- **Discrimination and toxicity**, including amplification of existing bias and discrimination through feedback loops, propagation of toxic content, and new forms of inequality arising from disparities in availability, quality, and capability of agents.
- **Privacy and security**, including unintended disclosure of personal or sensitive data, increased leakage risk from memory and long-term state, comprehensive logging and traceability needs, and cascading compromises that result in misaligned outcomes.
- **Misinformation**, including cascading misinformation when hallucinated or erroneous outputs from one agent are consumed and reused by other agents or systems.
- **Malicious actors and misuse**, including lowered barriers for designing and executing complex attacks, automation of multiple stages in cyber or biological risk pathways, large-scale personalized manipulation, fraud, and coordinated "swarm" or influence campaigns.
- **Human–computer interaction**, including reduced human oversight, anthropomorphic or socially persuasive behavior that can increase overreliance and information disclosure, and heightened difficulty for users in understanding or contesting agent behaviors.
- **Loss of control**, including oversight subversion, rapid and iterative action execution that can outrun monitoring and response, and behaviors that undermine shutdown, rollback, or containment mechanisms.
- **Socioeconomic and environmental harms**, including inequalities driven by differential access to agentic capabilities, potential collective disempowerment, economic disruption, and environmental impacts from large-scale autonomous operation.
- **AI system safety, failures, and limitations**, including self-proliferation, self-modification, self-exfiltration, self-replication, agentic misalignment, deceptive behavior and scheming, reward hacking, collusion, long-term planning and goal pursuit, cross-domain influence, real-world interaction, and limited effective human oversight.

2 For the full list of high-priority risk management steps, see the "High-Priority Subcategories" section in this document.

Given the different configurations of agentic AI systems, this Profile emphasizes governance mechanisms that 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. This Profile prioritizes risk-management practices that preserve meaningful human responsibility while enabling bounded autonomy within clearly defined limits.

Key risk-management levers emphasized throughout the Profile include:

- **Human control and accountability**, including clear role definitions, intervention points, escalation pathways, and shutdown mechanisms.
- **System-level risk assessment**, especially for multi-agent interactions, tool use, and environment access.
- **Continuous monitoring and post-deployment oversight**, recognizing that agentic behavior may evolve over time and across contexts.
- **Defense-in-depth and containment**, treating sufficiently capable agents as untrusted entities due to the limitations of current evaluation techniques.
- **Transparency and documentation**, including clear communication of system boundaries, limitations, and risk-mitigation decisions to relevant stakeholders.

However, several important limitations remain in applying these risk management levers. Taxonomies for agentic AI vary widely, and are often inconsistently applied, limiting the ability to harmonize recommendations across organizations and jurisdictions. Human control and accountability are hampered by the increased autonomy and complex multi-system behavior of agentic AI, further complicating the attribution of actions and liability. Additionally, many risk-measurement techniques remain underdeveloped, particularly with respect to emergent behaviors, deceptive alignment, and long-term harms. As a result, the Profile adopts a precautionary approach, emphasizing conservative assumptions, layered safeguards, and continuous reassessment as system capabilities evolve.

Because of these uncertainties, this document should not be treated as a static checklist, but a living framework intended to evolve alongside agentic AI research, deployment practices, and governance norms. This Profile aims to help key actors in the AI value chain by providing a shared structure, vocabulary, and set of expectations that support responsible development and deployment of agentic AI systems while enabling innovation that does not come at the expense of safety, security, or public trust.

Introduction and Objectives

The widespread emergence and use of agentic AI, or autonomous AI agents, present many of the same risks as other frontier AI systems, but also present additional and unique risks that require tailored risk-management methods. Agentic AI risk-management practices must include governance mechanisms that align with the system's structures, unique capabilities, and affordances. The guidance provided in this Profile aims to address those additional considerations, and complements the U.S. National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) and the UC Berkeley AI Risk-Management Standards Profile for General-Purpose AI, or "GPAI Profile" (Madkour et al., 2026). The guidance also builds upon other leading AI risk-management resources, including Bengio et al. (2025), Oueslati & Staes-Polet (2025), and OWASP (2025a).

This Profile is intended for use by agentic AI developers, deployers, and policymakers seeking to identify and mitigate risks associated with agentic AI. The guidance provided here is intended to help govern, map, measure, and manage risks specific to agentic AI systems. Widespread norms for using best practices such as those detailed in this document can help ensure that developers and deployers of agentic AI systems can be competitive without compromising on practices for AI safety, security, accountability, and related issues. Agentic AI requires governance practices beyond those used for GPAI. Governance must be tailored to manage the capabilities and affordances that these systems possess. AI agents and agentic systems can make independent decisions, generate or pursue goals and sub-goals, re-plan in certain environments, and delegate tasks to other models or agents. This Profile aims to provide actionable guidance for managing risks associated with AI agents and agentic AI and their unique capabilities. Implementation of this guidance should help reduce both the likelihood and magnitude of risks unique to agentic systems, including goal misalignment, unauthorized autonomous actions, cascading system failures, and malicious exploitation of agentic AI capabilities and vulnerabilities.

AGENTIC AI AND AI AGENTS

An important technical distinction between "AI agents" and "agentic AI" is that an AI agent is a single model equipped with tools for performing well-defined, end-to-end tasks, while agentic AI is often a system composed of multiple agents coordinating in pursuit of broader goals

(CSA Singapore & FAR.AI, 2025; Raza et al., 2025). Single-agent AI systems rely on one AI agent operating in isolation for all decision-making and action execution. Multi-agent AI systems (MAS) are composed of multiple AI agents that operate simultaneously and interact with one another. These systems are often characterized by each agent holding specific roles and possessing distinct capabilities that contribute to collective system behavior (Google Cloud, n.d.).

These distinctions are critical for defining appropriate risk governance approaches. For example, for MAS composed of multiple models, each possessing different capabilities and functional responsibilities, risks would need to be evaluated separately and collectively to account for behaviors that may only emerge from the complexities of agent interaction. Both MAS and single-agent systems require risk assessment during the training phase, as well as subsequent phases, in order to avoid the development of black-box systems that may become increasingly difficult to manage.

Approaches for risk management can also depend on the scope of the agent's or agentic system's activities. General-purpose AI agents — i.e., "generalist" or "general" agents — can perform a wide range of tasks across domains (e.g., a personal digital assistant).[3] On the opposite side of the spectrum are "specialist" agents, which tend to be narrowly focused on specific domains and optimized for a finite set of tasks (e.g., an agent that files taxes) (WEF, 2025b; Deshpande & Joshi, 2025). In some cases, agents are built on top of GPAI models, however, they may still be considered "specialist" agents based on whether or not they are constrained to domain-specific tasks. While many definitions have been proposed for the term "AI agent" (Casper et al., 2025), one commonality across descriptions is that AI agents act with some level of autonomy (Mitchell et al., 2025). These categories and descriptions provide a useful baseline, yet we acknowledge that definitions vary across developers and use cases and that the exact classification of agents is not always binary.

KEY TERMS AND DEFINITIONS

We use these key terms as follows:

- **Agentic AI:** "*Agentic AI refers to AI systems composed of* [one or more] *agents that can behave and interact autonomously in order to achieve their objectives. Traditional software typically follows fixed pathways to solve problems. In contrast, agent-based systems* [can]

3 Anthropic's computer-using agent (CUA) Operator (Anthropic, 2025f) can be considered an example of a general-purpose or "generalist" AI agent.

operate like independent assistants that choose and combine several actions to achieve their goals" (GOV.UK, n.d).
    - While this definition may presume a high level of autonomy, we acknowledge that AI agency exists on a spectrum of autonomy and authority (Mitchell et al., 2025; Kasirzadeh & Gabriel, 2025; WEF, 2024; WEF, 2025b) and cannot be viewed as binary. (For more on AI agent characteristics and properties, see Map 5.1.)
- **AI Agent:**[4] Refers to an AI system with the ability to "*...make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with its environment — for example by creating files, taking actions on the web, or delegating tasks to other agents — with little to no human oversight*" (Bengio et al., 2025, p. 38).
- **General-Purpose AI (GPAI):** Our usage of the terms "general-purpose AI model" and "general-purpose AI system" is very similar to the corresponding terms in the EU AI Act (EP, 2024), except that we do not exclude AI models used for research.
    - **GPAI Models:** "*'General purpose AI model' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications...*" (EP, 2024, Article 3(63)).
        - Examples of GPAI models include GPT-5, Claude 4, LLaMA 3, and others.
    - **GPAI System:** "*'General-purpose AI system' means an AI system which is based on a general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems*" (EP, 2024, Article 3(66)).

SCOPE

The Agentic AI Risk-Management Guidance provides recommendations based on the categories within each of the core functions defined in the NIST AI Risk Management Framework: Govern, Map, Measure, and Manage (NIST, 2023a). Sub-categories (e.g., Map 1.1) have been selected based on two criteria: (1) they are considered "high-priority sub-categories" in the GPAI Profile or (2) they require additional consideration beyond the content of the GPAI Profile (Madkour et al., 2026). This approach focuses on providing guidance for risks associated with downstream agentic AI, but aimed at upstream model developers. Guidance that is unique

4 Our use of the term "AI agent" excludes artificial intelligence systems that exhibit no levels of autonomy, including large language models and conversational agents (chatbots) that operate without the autonomous and independent decision-making characteristics that define AI agents.

to agentic AI is prioritized, in an effort to reduce duplicating existing GPAI guidance that directly applies without modification. For contextual clarity, we include high-level guidance that applies to both GPAI and agentic AI, and we encourage users of the Agentic AI Profile to refer to the GPAI Profile for more comprehensive guidance on overlapping areas.

The guidance in this document applies to agentic AI, or AI agents, as defined under Key Terms and Definitions. The guidance applies to both general-purpose agentic AI (i.e., generalist systems) and domain-specific, single-purpose agentic AI (i.e., specialist systems), however, we do not provide explicit guidance for specific use cases (e.g., healthcare agentic AI). The guidance considers both open-source and closed-source models, with clarifications provided where distinctions between the two are relevant. Coverage also encompasses single-agent and multi-agent systems, with specific guidance provided where differences require distinct considerations and risk-management approaches. This guidance is less relevant to systems with limited agency (e.g., purely reactive agents with static goals) and systems with zero autonomy (e.g., systems requiring continuous human control for all actions).

High-Priority Subcategories

Based on baseline or minimum expectations for users of our GPAI Profile, we refer to the following high-priority subcategories. For more about the rationale for determining high-priority subcategories, and for more on the high-priority risk-management steps, see Madkour et al. (2026).

- **Govern 2.1:** Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.
- **Govern 4.2:** Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly.
- **Govern 5.1:** Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks.
- **Map 1.1:** Intended purposes, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: the specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related

limitations about AI system purposes, uses, and risks across the development or product AI lifecycle; and related TEVV and system metrics.

- **Map 1.5:** Organizational risk tolerances are determined and documented.
- **Map 5.1:** Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.
- **Measure 1.1:** Approaches and metrics for measurement of AI risks enumerated during the Map function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.
- **Measure 3.2:** Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.
- **Manage 1.1:** A determination is made as to whether the AI system achieves its intended purposes and stated objectives and whether its development or deployment should proceed.
- **Manage 1.3:** Responses to the AI risks deemed high priority, as identified by the Map function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting.
- **Manage 2.3:** Procedures are followed to respond to and recover from a previously unknown risk when it is identified.
- **Manage 2.4:** Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.
- **Manage 4.1:** Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.

LIMITATIONS AND CHALLENGES

The rapid evolution of agentic AI has resulted in a **lack of standardized terminology**,[5] creating challenges for implementing and harmonizing risk-management practices. Due to the lack of consensus on frameworks for defining levels of autonomy, terminology and definitions across

5 While there have been many attempts, there is no widely accepted definition or term for "AI Agent" (Casper et al., 2025), which has been further complicated by lack of definitional consensus for the term "AI" itself.

entities remain inconsistent. Additionally, several existing definitions of autonomy levels have relied on frameworks intended for technologies that are significantly different from agentic AI (e.g., self-driving car autonomy levels).

Establishing clear boundaries of responsibility presents a unique challenge for strategic AI risk management, particularly when defining roles and responsibilities. Unlike traditional (or non-agentic) AI models, for which accountability for actions taken is easier to attribute to a human actor (e.g., model deployer), the autonomous nature of agentic systems — including their ability to carry out multi-step tasks, use tools, and carry out independent decision-making — introduces even more difficulty in attributing outcomes to specific actors or components within the system.

Managing agentic AI is further complicated by the fact that many **existing AI management frameworks and resources adopt a predominantly model-centric approach**. While this may be largely applicable to agentic AI risk-management, it may prove insufficient when accounting for properties specific to agentic AI systems (e.g., environment and tool access, multi-agent communications and coordination, and differences in infrastructure). These aspects present distinct risks that require taking into account the entire system and may not be adequately addressed through a model-centric approach.

The known limitations of current evaluation approaches for capability elicitation are further exacerbated in agentic AI systems. Consequently, emerging literature on "AI control" argues that sufficiently capable agentic AI systems warrant treatment as *untrusted models*, not on the assumption of malicious intent, but due to their potential for subversive behaviors (Greenblatt et al., 2024; Hammond et al., 2025; Wen et al., 2024; Terekhov et al., 2025a). This position is supported by evidence that advanced, strategically aware models can develop and conceal adversarial behaviors during evaluation, only revealing them during deployment, particularly when granted broader autonomy, tool use, or system access. Because such models may be capable of adaptive evasion of controls, backdooring, or subtle sabotage that escapes pre-deployment testing, risk-management approaches must assume worst-case behavior. While newly developed benchmarks and threat-modeling efforts promise significant improvements in agentic evaluation suites (Bhatt et al., 2025; Griffin et al., 2024), we recommend the continued treatment of AI agents as untrusted entities, relying on defense-in-depth (the process of layering multiple defenses together to catch adversarial inputs and mitigate rogue actions, improving security outcomes), containment, and robust system-level monitoring to mitigate

risks associated with misalignment and deceptive alignment (CSA, 2025b; Narajala et al., 2025; Terekhov et al., 2025b).

Many risk-management methods are still active areas of scientific research and will require ongoing testing and evaluation. For example, alignment — ensuring an agent's behaviors adhere to intended values and goals — is a nascent scientific field. This encompasses both the technical challenges of preventing misalignment (e.g., an agent pursuing undesired sub-tasks), and the ethical challenges in **defining values or objectives** across diverse cultural and geographic norms and practices. Alignment efforts reflect the values, priorities, and worldviews of their creators, influencing what is considered to be "aligned," "safe," or "responsible" system behavior—terms that are often subjective. Therefore, defining and assessing alignment is a difficult task, further complicated by agentic capabilities (e.g., adapting plans over time).

Guidance

The tables below highlight the relevant NIST AI RMF categories and subcategories, and supplemental guidance, for agentic AI systems (NIST, 2023a; Madkour et al., 2026). The tables address the following AI RMF functions: Govern, Map, Measure, and Manage. The papers and resources included in the "Resources" column provide overarching guidance or tools that can support the recommendations provided in the sub-category.

GOVERN

Applicability and Supplemental Guidance for Agentic AI and AI Agents	Resources
Govern 1: Policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively.	
Govern 1.2 The characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices.	
Characteristics specific to trustworthy AI for agentic AI include the following: • **Behavioral consistency:** While a certain level of appropriately justified behavioral variation is reasonable, agentic AI systems should generally demonstrate behavioral consistency and reliability. • **Human control:** Agentic AI systems should at all times remain in appropriately resourced human control and responsibility, while enabling system autonomy within designated bounds. • **Transparency and explainability:** Agentic AI stakeholders (e.g., developers, deployers, managers, and evaluators) should ensure, and be ensured, appropriate, accurate, and actionable visibility into system behavior and organizational processes. Consider using methods such as reasoning, traceability, intent disclosure, and mechanistic interpretability, while accounting for open problems in these methodologies (Korbak et al., 2025; Raza et al., 2025; Sharkey et al., 2025). • **Alignment:** Implement measures to align system behavior and actions with the desired goals and expectations. • **Privacy:** Protect sensitive information across interactions (Murugesan, 2025) and tools, and protect against widespread loss of privacy through agentic AI-enabled data overreach, surveillance, and other means (Batra et al., 2025). (For more on privacy, see Map 1.1 and Manage 1.3.) • **Security:** Safeguard sensitive data and prevent misuse. (For more on security, see Map 1.1 and Manage 1.3.)	NIST AI Risks and Trustworthiness (NIST, n.d.a) NIST Trustworthy and Responsible AI (NIST, n.d.b) A Taxonomy of Trustworthiness for Artificial Intelligence (Newman, 2023) For more on trustworthy agents, see Anthropic (2025a) For more on reasoning traceability and intent disclosure, see: • Korbak et al. (2025) • Raza et al. (2025)

Applicability and Supplemental Guidance for Agentic AI and AI Agents	Resources
Protection of human rights: Ensure that human rights are both safeguarded from violation and protected.[6] For example, some applications of AI may impact the human right to the freedom of thought (Teo, 2024), and the offloading of tasks to agentic AI could further threaten human critical thinking skills and the ability to think freely without influence (Lee et al., 2025). Agentic AI may also introduce new forms of inequality, further threatening the human right to protection against discrimination (Sharp et al., 2025).	
Govern 1.4 The risk-management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.	
In addition to standard risk-management policies and frameworks that must be established and maintained regardless of agentic AI use, additional governance measures should be developed to address the risks unique to agentic AI: • **Develop agent-specific policies:** Create policies that directly address the unique characteristics of agentic AI, such as delegated decision-making authority, tool access, and the ability to generate or pursue sub-goals. When developing these policies, take into account: ◦ Characteristics of trustworthy AI agents (see Govern 1.2); ◦ Agentic AI properties and characteristics (see Map 5.1); ◦ Organizational risk tolerances (see Map 1.5); ◦ Agentic AI risks (see Map 1.1); and ◦ Agentic AI-specific risk mitigations and responses (e.g., scalable oversight. See Manage 1.3.) • Consider translating key **governance documents** into structured, **AI-interpretable frameworks**. This procedure allows agentic systems not only to operate under human-directed rules but also to access and act in accordance with organizational safety and risk priorities in real time. ◦ When implementing this translation, a critical distinction must be made between a framework that is AI-interpretable and one that is AI-writable. While making the framework AI-interpretable is a recommended control for enabling safer autonomy, allowing an AI to *modify* its own framework is a high-risk activity. Granting write-access without appropriate human review could allow an agent to introduce loopholes or weaken its own oversight. (For more on these meta-level risks, see guidance in **Map 5.1.**) Any consideration of AI-writable frameworks must be approached with extreme caution and be subject to robust, independent human oversight. ◦ Additionally, building frameworks that are also measurable and verifiable enables actionable oversight of the AI agent. • **AI supply chain awareness**. An organization's risk-management processes must extend to the entire AI supply chain. (For more on supply chain considerations, see Govern 6.1.)	For more on intervention points and guardrails, see Toner et al. (2024) For more on baseline governance mechanisms, see WEF (2025b) For managing AI supply chain risk and AI supply chain transparency, see: • SBOM for AI Use Cases (CISA, 2025) • TAIBOM (Trustable AI Bill of Materials) (TAIBOM, n.d.) • SPDX AI-SBOM (SPDX, n.d.a) • OWASP AIBOM (OWASP, n.d.c) • AI Models and Model Cards Inventory Management (CycloneDX, n.d.a) For the software components and cloud infrastructure that run AI models, see: • CycloneDX (CycloneDX, n.d.b) • SPDX (SPDX, n.d.b)

6 Human rights that may be implicated by agentic AI include, but are not limited to: freedom from physical and psychological harm; right to equality before the law and to protection against discrimination; right to own property; freedom of thought, religion, conscience, and opinion; freedom of expression and access to information; right to take part in public affairs; right to work and to gain a living; rights of the child; and rights to culture, art, and science.

Applicability and Supplemental Guidance for Agentic AI and AI Agents	Resources
• **Baseline governance mechanisms for agentic AI,** as highlighted by WEF (2025b), must include the following measures to establish a foundational framework that scales proportionally with system complexity and risk levels: ◦ Access control (e.g., technical guardrails, see Toner et al., 2024) (see Manage 1.3); ◦ Legal and compliance (e.g., alignment with legal guardrails, see Toner et al., 2024) (see Govern 1.4); ◦ Testing and validation (e.g., measurement and evaluation, see Toner et al., 2024) (see Measure 1.1); ◦ Monitoring and logging (see Manage 4.1); ◦ Human oversight (see Map 3.5); ◦ Traceability and identity (see Manage 4.1); ◦ Long-term management; ◦ Trustworthiness and explainability; and ◦ Manual redundancy. A core challenge in managing the risks of agentic AI is the lack of a standardized vocabulary to describe a system's capacity for independent action. Moving beyond a simple definition of "agency" and instead adopting a consensus-driven framework for characterizing agentic systems based on key characteristics such as **autonomy**, **authority**, and **environment** is recommended (WEF, 2025b). (For more on defining agentic properties and dimensions, see Map 5.1.) Adopting a shared framework based on these pillars would provide a common language for developers, deployers, regulators, and auditors. This structured approach would serve several critical functions: • **Standardized risk assessment:** It would allow organizations to benchmark an agent's risk profile in a multi-dimensional way, enabling more consistent risk-tiering and the application of appropriate controls. • **Regulatory clarity:** It would provide a basis for regulatory bodies to scope rules and tailor safety requirements to the specific context in which an agent operates. • **Informing governance and management:** It would directly inform all core RMF functions. For instance, an agent with high autonomy and broad authority operating in a complex environment would trigger the most stringent protocols for Govern, Map, Measure, and Manage. (For more on supply chain considerations, see Govern 6.1.)	
Govern 1.5 Ongoing monitoring and periodic review of the risk-management process and its outcomes are planned and organizational roles and responsibilities clearly defined, including determining the frequency of periodic review.	
The rapid evolution of AI technology and learning behavior of AI agents (particularly if the agent's affordances include interaction with other agents, systems, or tools) necessitates continuous review of risk-management processes and practices. In addition to standard periodic reviews, reviews should be triggered whenever significant changes occur that may require a comprehensive re-evaluation of the risk-management plan. Significant changes may include:	For more on updating and reviewing risk-management processes, see: • Benchmark Early and Red Team Often (Barrett et al., 2024) • Monitoring and Review Sections of ISO 31000 Risk Management Guidance (ISO, 2018)

Applicability and Supplemental Guidance for Agentic AI and AI Agents	Resources
• An agent exhibiting new or emerging dangerous or dual-use capabilities; • Increased levels of autonomy; • Alterations to the agent's affordances and privileges; • Changes in deployment context or the agent's environment; • Integrations or new interactions with other systems; and • Integration of, removal of, or any changes to entities or components in the supply chain (e.g., data, models, programs, infrastructure) (Sheh & Geappen, 2025). (For more on agent communication monitoring and safety, see Manage 1.3 and Manage 4.1.)	• Monitor step of the NIST Risk Management Framework (NIST, 2018) For more on AI supply chain entities, see Sheh & Geappen (2025).
Govern 1.7 Processes and procedures are in place for decommissioning and phasing out AI systems safely and in a manner that does not increase risks or decrease the organization's trustworthiness.	

When establishing processes and procedures for responsible decommissioning of AI agents or agentic AI systems, it is recommended to account for the following:
• 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).
• Establish shutdown protocols based on severity levels, determining the need for partial or complete shutdown (Oueslati & Staes-Polet, 2025).
  ◦ Consider selectively restricting specific agent capabilities, authorizations, and access to resources in response to certain triggers.
• In addition to automatic emergency shutdown, **manual shutdown methods** should be available as a last-resort control measure (Hadfield-Menell et al., 2017; Oueslati & Staes-Polet, 2025).
• Account for and implement safeguards that prevent the agent from taking actions to circumvent shutdown.
  ◦ For example, in certain test environments, models have shown tendencies to copy themselves to avoid being shut down (Hashim, 2024), and inclinations to sabotage shutdown mechanisms (Schlatter et al., 2025).[7]
• Identify and document all dependencies and system integrations, for both internal and external (e.g., cloud services and third-party software) systems. Establish procedures for isolating the agent from these systems in the event of an emergency shutdown.
  ◦ Identify any dependencies or integrations where shutdown may result in adverse, mission-critical effects.
• Train relevant actors (e.g., staff) on intervention protocols (Oueslati & Staes-Polet, 2025).
• Document and retain information on shutdown incidents for internal tracking and regulatory compliance.
	For more on emergency shutdowns, see:
	• Section 4.3.2 in Oueslati and Staes-Polet (2025)
	• Hadfield-Menell et al. (2017)

7 OpenAI's o3 model sabotaged shutdown mechanisms in 79 out of 100 tests run by Palaside research (Schlatter et al., 2025).

Applicability and Supplemental Guidance for Agentic AI and AI Agents	Resources

• Establish and document comprehensive post-shutdown procedures for investigating root causes and identifying mitigations, controls, or remediations that need to be implemented prior to reactivation (Oueslati & Staes-Polet, 2025).
• Avoid the use of overly sensitive filter mechanisms and triggers that may disrupt operations and drain resources, or that may fail to detect and prevent harmful outcomes (Oueslati & Staes-Polet, 2025).
• **Establish and maintain failover procedures** to transition to backup non-AI systems in the event that AI systems experience failure, performance degradation, or shutdowns, or otherwise become unavailable.
  ◦ Maintain current copies of organization-critical data in systems independent of AI infrastructure.
  ◦ Deploy deterministic backup systems capable of sustaining essential operations during AI system outages or shutdowns to ensure business continuity.
  ◦ Conduct periodic testing to verify that backup systems can handle real-world load without AI system support.
  ◦ Regularly assess whether agentic workflows have become mission-critical and update contingency systems and procedures accordingly.
  ◦ Implement controls preventing AI systems from compromising or interfering with backup systems.
• Periodically audit for and document mission-critical agentic AI workflows to facilitate reconstruction or replacement during recovery.	

Govern 2: Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks.	
Govern 2.1 Roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks are documented and are clear to individuals and teams throughout the organization.	

Develop effective human-agentic AI management hierarchies that preserve human authority while leveraging AI as a supportive tool. Consider the following:
• Ensure agentic AI is a tool under human oversight, not a "peer" or "subordinate" in the workforce. Avoid referring to or considering AI agents as "AI workers" or "AI employees" (Shapiro, 2025).
• Define clear boundaries for final decision-making, roles, and responsibilities for both human managers and agentic AI systems (Renieris et al., 2025):
  ◦ Define areas or actions where accountability and final decision-making remain solely with human managers and staff.
  ◦ Define areas or actions where agentic AI may act independently within predefined boundaries, and the conditions that would automatically trigger human oversight or approval.
• Define specific checkpoints within the agent's workflow where human oversight is required. These checkpoints may also be triggered by specific actions (e.g., deviating from expected behavior) or conditions (e.g., escalation of risk) (Oueslati & Staes-Polet, 2025).
	Redefining Management for a Superhuman Workforce (Renieris et al., 2025)
	
	Governing AI Agents Under the EU AI Act (Oueslati & Staes-Polet, 2025)
	
	For more on stakeholder roles in agentic AI security see Table 2 in (CSA Singapore & FAR.AI, 2025).

Applicability and Supplemental Guidance for Agentic AI and AI Agents	Resources
• Regularly evaluate comparisons between agent and human decisions to identify gaps and help cultivate proper human-AI collaboration. • Allocate appropriate human oversight, particularly for systems that may present low-probability but high-impact risks. Establish clear roles and responsibilities across the organization to ensure agentic AI security. A report by CSA Singapore & FAR.AI (2025) highlight several roles: • **Model developers:** Implement adequate autonomy-aware defenses to ensure safe planning, reasoning, and tool use. • **AI vendors:** Provide transparency to buyers on workflow risks and conduct comprehensive risk assessments to check security capability robustness. Anticipate emergent autonomy risks and implement safe boundaries for delegated tasks. • **Enterprise AI buyers:** Include agentic-specific safeguards (e.g., human-in-the-loop) in procurement contracts. Perform risk assessments and require disclosure of autonomy levels to deploy trustworthy and secure AI systems. • **Enterprise in-house developers:** Configure technical controls for secure operation and implement monitoring for detecting anomalies on autonomous operation. • **End users:** Interact with AI systems responsibly by providing clear objectives to agents, reviewing approval prompts, and serving as auditors to refine oversight policies. • **Academic researchers/think tanks:** Extend research on attack and defense mechanisms to agentic-specific vulnerabilities. Test emergent risks unique to agentic workflows and recommend appropriate mitigations. • **Cybersecurity providers:** Strengthen enterprise security by developing agent monitoring tools and improve the integrations between existing security solutions. Conduct red teaming that targets agentic systems. • **Third-party AI assurance providers:** Test and evaluate agentic systems (e.g., jailbreak attempts) throughout the lifecycle to discover system and model vulnerabilities and validate alignment with safety standards. • **Information security teams:** Identify cybersecurity-, governance-, and compliance-related risks within enterprise buyer/developer teams. Extend scope to include runtime agent oversight and prepare incident responses for agent misuse. • **Standards bodies:** Create AI security practice standards that are specific to autonomy domains (e.g., multi-agent system safeguards). • **Regulators:** Develop and enforce agent-specific best practices and regulations (e.g., clear liability chains) to ensure accountability of agent behaviors. • **Policymakers:** Collaborate with stakeholders to create policies that protect the public from cybersecurity harms. Promote research on agentic AI security, invest resources into development of talent skilled in agent oversight, and update national governance frameworks for autonomous workflows. (For more on human oversight processes and procedures, see Map 3.5.)	

Applicability and Supplemental Guidance for Agentic AI and AI Agents	Resources
Govern 4: Organizational teams are committed to a culture that considers and communicates AI risk.	
Govern 4.2 Organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and they communicate about the impacts more broadly.	
The increased autonomy inherent in agentic AI systems necessitates continuous monitoring and automated reporting. Automated notifications to relevant AI actors should be established for: • Deviations from expected behavior (e.g., unauthorized access, unauthorized decision making). • Malfunctions and near-misses. • Serious incidents. Incidents should be reported to appropriate oversight bodies and added to public incident databases (e.g., AIID, n.d.; MITRE, n.d.a; MIT, 2025a). Provide clear disclosures to users to inform them when they are interacting with an AI agent, particularly in situations where there is potential for confusion about whether they are communicating with a human or an AI system. Clearly document and communicate: • The known boundaries and limitations of the agentic system, including scenarios that may be unreliable or unsafe (see Map 2.2). • Prohibited use cases and explicit restrictions on certain applications. • Clear instructions on appropriate use, potential risks, warning signs, and problematic behavior. ◦ Instructions should also include clear mechanisms for reporting problematic behavior to relevant authorities and stakeholders. Governance mechanisms for agentic AI must account for risks arising from multi-agent interactions. Oversight cannot be limited to individual agent behavior but must also monitor the health and safety of the multi-agent system as a whole.	Incident Databases and Risk Registers: • AI Incident Database (AIID, n.d.) • ATLAS AI Incidents (MITRE, n.d.a) • MITRE AI Risk Database (MITRE, n.d.b) • MIT AI Incident Tracker (MIT, 2025a) • MIT AI Risk Repository (MIT, 2025b) • AI Incidents and Hazards Monitor (OECD.AI, n.d.)
Govern 5: Processes are in place for robust engagement with relevant AI actors.	
Govern 5.1 Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks.	

Establish and maintain policies and procedures for the following:
• **Multi-channel feedback systems**, including:
  ◦ Clear accessible feedback channels for users, affected communities, researchers, and civil society to report concerns and incidents.
  ◦ Bi-directional feedback mechanisms that facilitate active engagement and an iterative exchange of information.
  ◦ Processes for active stakeholder engagement.
• **Structured external evaluation programs**
  ◦ Plan for regular independent evaluations and audits by trusted third-party organizations, including external red teaming (see Measure 1.1).
	Palisade Research AI Misalignment Bounty program (Palisade Research, n.d.)
	
	Anthropic's "agent bio bug bounty" (Anthropic, 2025b)

Applicability and Supplemental Guidance for Agentic AI and AI Agents	Resources

• **Incentivized risk-discovery programs**, including:
  ◦ Bug bounty or "misalignment bounty" programs.
  » Incentivize users and external actors to find and report instances of misaligned and harmful agent behavior. For example, see Palisade Research AI Misalignment Bounty program (Palisade Research, n.d.), Anthropic's bug bounty program (Anthropic, 2025b), and OpenAI's "agenti bio bug bounty (OpenAI, 2025)."
  ◦ Collaborative research initiatives.
  ◦ Community-based monitoring (due to the automated and iterative nature of agentic AI).
• **Feedback integration and response protocols**
  ◦ Create clear prioritization frameworks for analyzing and identifying feedback priority levels.
  ◦ Establish feedback documentation, along with retention practices and procedures.
• **Legal protections for good-faith reporting**
  ◦ Establish robust whistleblower protection policies (Wu, 2024).
  ◦ Establish a safe harbor for good-faith independent AI evaluation and red teaming (Longpre et al., 2024).	

Govern 6: Policies and procedures are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.	
Govern 6.1 Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights.	
Governance mechanisms for agentic AI must account for risks arising from interactions with external agents. Oversight cannot be limited to individual agent behavior but must also monitor the health and safety of the agent's interactions with external agentic systems or tools (for guidance on multi-agent interactions, see Map 4.2.). AI agents acting autonomously may take actions that infringe on intellectual property rights. Procedures focused on minimizing the risk of these actions or responding to them should be implemented specifically for these systems, including: • Implementing content filtering; and • Exercising caution when dealing with systems that continuously learn from their environments. Agentic AI systems are often composed of numerous third-party components, including pre-trained models, datasets, and software libraries, each of which introduces potential risks. A comprehensive risk-management process requires transparency into these components throughout the supply chain. • Organizations should establish procedures to document and assess the provenance of all components used in an agentic AI system. This can be achieved by integrating an AI Bill of Materials (AIBOM) (e.g., CISA, 2025; TAIBOM, n.d.) or similar artifact into the development lifecycle. These documents provide a formal record of the parts and data used to train, test, and build an AI system, enabling more effective risk management. • Additionally, developers should follow the general guidance/framework of SLSA (Supply-chain Levels for Software Artifacts) (SLSA, n.d.).	For managing AI supply chain risk and AI supply chain transparency, see: • SBOM for AI Use Cases (CISA, 2025) • TAIBOM (Trustable AI Bill of Materials) (TAIBOM, n.d.) • SPDX AI-SBOM (SPDX, n.d.a) • AI Models and Model Cards Inventory Management (CycloneDX, n.d.a) • OWASP AIBOM (OWASP, n.d.c) For the software components and cloud infrastructure that run AI models, see: • CycloneDX (n.d.b) • SPDX (n.d.b)

MAP

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources
Map 1: Context is established and understood.	
Map 1.1 Intended purposes, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: the specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes, uses, and risks across the development or product AI lifecycle; and related TEVV and system metrics.	

Identify all risks that may stem from the agentic AI system based on system or model-independent research (e.g., literature review, stakeholder interviews, risk repositories, incident databases). This can include an assessment of the system and its characteristics, the nature and sources of risks, and relevant information about similar systems (EC, 2025). **Take into consideration the following risks[8] unique to agentic AI**:
 	
Discrimination and Toxicity
• **New and amplified discriminatory patterns**
  ◦ Agentic AI may introduce new forms of inequality, further threatening the human right to protection against discrimination (Sharp et al., 2025). Such harms could manifest in a variety of ways:
  » When agents are involved in taking actions that constitute, or closely resemble, decisions affecting individuals or groups, they may reproduce familiar patterns of discrimination observed in simpler automated decision systems, such as biased allocation of opportunities, services, or enforcement (Chan et al., 2023).
  » These risks may be compounded over time, as agentic systems can repeatedly act across domains or stages of a process, amplifying small initial disparities into persistent or cumulative disadvantages that give rise to systemic risks (Bellogín et al., 2025).
  » When access to more capable agents, such as those with stronger negotiation abilities, broader tool access, or greater autonomy, is directly tied to underlying model capabilities, compute resources, or tiered pricing structures (Hammond et al., 2025).
• **Bias amplification**
  ◦ In agentic AI systems, where autonomous operations at scale can create feedback loops that both mask and magnify discriminatory patterns, bias and discrimination risks may be amplified, further embedded, and potentially harder to identify (Brohi et al., 2025).
  ◦ Specifically, because agentic AI autonomously mixes and repurposes information from disparate sources and deploys it across repeated actions, long-standing sources of unfairness in AI systems, such as domain shift and context mismatch, are much more likely to emerge, compound, and evade detection than in non-agentic generative systems.
	Probabilistic Risk Assessment for AI (Wisakanto et al., 2025)
	
	General-Purpose AI Code of Practice, Safety and Security Chapter, Measure 2.1 (EC, 2025)
	
	For more on types of AI risks, see:
	• Section 2 of Bengio et al. (2025)
	• MIT AI Risk Repository (MIT, 2025b)
	• NIST (2024)
	• Enkrypt AI (2025).
	
	Incident Databases and Risk Registers:
	• AI Incident Database (AIID, n.d.)
	• ATLAS AI Incidents (MITRE, n.d.a)
	• MITRE AI Risk Database (MITRE, n.d.b)
	• MIT AI Incident Tracker (MIT, 2025a)
	• MIT AI Risk Repository (MIT, 2025b)
	• AI Incidents and Hazards Monitor (OECD.AI, n.d.)
	
	For more on key security risks associated with AI agents, see:
	• Section 3 of Díaz et al., (2025)
	• OWASP Agentic AI - Threats and Mitigations (OWASP, 2025a)
	• Cisco (n.d.)

8 The risks in this section are categorized and drawn from a compendium of several leading resources, including MIT (2025b), Bengio et al. (2025), and NIST (2024).

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources

• **Toxic content**
  ◦ Agentic AI systems can amplify child sexual abuse material (CSAM) risks by autonomously initiating contact with victims (Ciardha et al., 2025) and automating other aspects, such as search and collection of illegal content and creation of illegal distribution networks.
  ◦ Agentic AI systems can contribute to hate speech amplification in several ways, including (Sonni, 2025):
  » Automated content production that facilitates large-scale dissemination of hate speech;
  » Personalized content that creates echo chambers and exacerbates polarization; and
  » Enhanced multimodal manipulation that makes misleading content more persuasive.
 	
Privacy and Security
• Comprehensive logging and traceability can effectively function as a form of continuous surveillance, potentially introducing significant privacy risks, including the misuse of sensitive information or the creation of monitoring infrastructures that themselves pose risks to users and other stakeholders.
  ◦ Additionally, even in the absence of direct surveillance risks, comprehensive logging and access to significant amounts or personal or sensitive information lead to data overreach and introduce tradeoffs between functionality and privacy.
• 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. Additionally, the retention of sensitive information can increase the likelihood of access through methods such as prompt injection.
  ◦ Agent access to third-party systems and applications (e.g., email, calendar, or payment services) expands the attack landscape and has been demonstrated to introduce novel attack vectors, such as "confused deputy" attacks, where an agent is tricked into misusing its legitimate authority, as well as the exfiltration of sensitive information (Enkrypt AI, 2025).
  ◦ Agentic AI systems lack a clear separation between internal data — including instructions and prior information — and external data (Schulhoff et al., 2025). The adoption of agentic AI systems for applications like email management can introduce security risks such as memory poisoning attacks, which inject malicious information into an AI agent to induce undesirable behaviors (e.g., autonomously sharing sensitive information with an adversary) (Bryan et al., 2025). As a result, these systems can be compromised easily, resulting in violations to personal privacy.
  » Using prompt injections, attackers can collect information such as a victim's location, emails, documents, and calendar information, as well as allow attackers to conduct remote video recordings (Yair et al., 2025).	

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources
• Multi-agent systems 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 is analogous to the type of malware known as a worm, and the ability to evolve and improve as it hops between agents is akin to a polymorphic virus (Ju et al., 2024; Gu et al., 2024; Lee & Tiwari, 2024; Peigné et al., 2025). These propagation and adaptation dynamics make mitigation and detection particularly difficult across multi-agent systems and the interconnected applications they access. • Developers should account for vulnerabilities that may lead to unauthorized access of user data. (For more on system security and resilience, see Measure 2.7.) **Misinformation** • 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 (Bengio et al., 2025). • **CBRN** ◦ In addition to lowering the barriers to entry, facilitating the design of biological agents (e.g., viruses, toxins, or bacteria), 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 (e.g., data collection, operational planning, or simulated experiments and research) (ORF, 2024; Chin, 2025; Barret et al., 2024). • **Offensive cyber operations** ◦ 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). 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, 2025h; Lin et al., 2025). Additionally, multi-agent systems allow adversaries to carry out these attacks in a decentralized manner, enabling increased stealth in execution and limiting traceability. Standard security auditing relies on fixed system boundaries to trace threats, but multi-agent ecosystems operate through decentralized, ever-changing relationships. This complexity allows for emergent adversarial patterns that make it difficult for oversight bodies to identify and hold specific entities accountable (de Witt, 2025). ◦ Advanced AI agents can be used to generate increasingly personalized manipulative content at scale (e.g., phishing, vishing), and can iteratively enhance tactics by integrating user feedback, expanding the attack surface for social engineering. They may also evade detection by distributing attacks across many seemingly independent agents (de Witt, 2025). ◦ Additionally, agents can conduct "swarm attacks" by combining their resources to overwhelm their targets, similar to distributed denial of service attacks (de Witt, 2025).	

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources

• **Advanced persuasion and manipulation**
  ◦ Research has demonstrated that AI-generated messages on policy issues (e.g., automatic voter registration, carbon tax) were as persuasive as human-written messages, suggesting that developments in AI would make it possible to generate low-cost rapid persuasive messages at scale using agentic AI (Bai et al., 2025).
  ◦ AI agents may also be able to automate entire scam and fraud pipelines (Badhe. 2025).
• **Disinformation**
  ◦ Agentic AI may be used to amplify disinformation campaigns by automating key components of the process, including information gathering, target identification, and communication dissemination (Schmitt & Flechais, 2024; Heiding et al., 2024).
  ◦ Collaborative malicious AI agents can autonomously coordinate mass influence campaigns and potentially infiltrate communities, fabricating consensus while mimicking human social dynamics (Schroeder et al., 2025).
 	
Human-Computer Interaction
• **Unsupervised execution**
  ◦ Reduction of human oversight may escalate risks and increase the likelihood of unnoticed accidents and malfunctions (Bengio et al., 2025). This could take various forms, for example:
  » **API integration failures**: If an API with which an agent is integrated changes its data format (e.g., changing "price" to "cost"), the agent might not just fail, but catastrophically misinterpret the data (e.g., by treating a price of $100 as $0.00), leading it to execute high-speed, erroneous transactions before human intervention (Enkrypt AI, 2025).
  » **Uncontrolled resource consumption:** An agent entering an unintentional self-reinforcing loop (e.g., continuously calling a paid API to "verify" a failed step) has potential to lead to massive financial loss or internal denial of service (DoS), even without the involvement of a malicious adversary (Enkrypt AI, 2025).
  ◦ Insufficient logging of an agent's decision-making behavior could make it impossible to prove why an action was taken, undermining the ability to audit liability after a failure (Enkrypt AI, 2025).
• **Anthropomorphic AI**
  ◦ Anthropomorphic AI assistant behavior may increase user trust and encourage information sharing, increase the effectiveness of manipulation, and promote overreliance (Akbulut et al., 2024).
 	
Loss of Control
• **Oversight subversion**
  ◦ A model may intentionally disable oversight mechanisms or otherwise carry out oversight subversion in order to pursue its goals (Meinke et al., 2025).
• **Velocity of operations**
  ◦ 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.	

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources
» When identifying agentic AI risks, assess the cumulative impact of actions performed at scale. Individual actions that appear low-risk in isolation may pose significant risk when executed at scale or repeatedly by autonomous agents. **Socioeconomic and Environmental Harms** • **Power disparities and inequality** ◦ Agentic AI may contribute to inequalities due to disparities in the availability, quality, and quantity of agents, and in who has control over how these systems are designed and used (Sharp et al., 2025). • **Collective disempowerment** ◦ The deployment of agentic AI can result in collective disempowerment, as decision-making authority shifts away from humans and becomes increasingly concentrated among technologically advanced elites (Chan et al., 2023). • **Systemic delayed harms** ◦ The use of agentic AI for certain types of automated decision-making may result in non-immediate harms, and can be caused by the aggregate of several seemingly unconventional actions (Chan et al., 2023). • **Economic disruption** ◦ **Job market 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, possibly leading to significant disruptions in the skill requirements and wage distribution across multiple sectors (Bengio et al., 2025). ◦ **Impacts on labor and management:** The integration of agentic AI into the workforce introduces novel risks related to **"agentic management."** The deployment of AI agents to delegate and monitor tasks can lead to two primary concerns that increase the magnitude of societal impact: » **Mass worker surveillance:** The high degree of traceability required to monitor agentic tasks can generate vast datasets on worker behavior, creating significant privacy risks and the potential for a pervasive surveillance infrastructure. » **Ambiguous workplace hierarchies:** A lack of clear guidance on the authority between human workers and agentic systems, particularly in high-stakes fields like medicine and finance, can lead to critical errors, delayed accountability, and an erosion of trust in organizational structures. • **Environmental harms** ◦ **Data centers and power consumption:** General-purpose AI systems require orders of magnitude more energy than task-specific alternatives. This cost must be balanced with the expected utility of employing these systems. AI agents amplify these concerns, as poorly supervised agents may enter loops or run indefinitely, potentially incurring substantial environmental costs (Luccioni et al., 2024; Guidi et al., 2024).	

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources

• **Autonomous research and development**
  ◦ **Risks from AI R&D automation**: Profit incentives and market pressure may lead companies to increasingly automate R&D by using AI. This may weaken human oversight, obscure accidents or misuse, and expose supply chains to compromises that are hard to detect and difficult to revert. Additionally, AI agents may sabotage safety efforts, deploy unauthorized systems for potentially harmful purposes, and create malware that allows for uncontrolled scaling. These risks may stem from both misaligned agents or malicious human insiders (Clymer et al., 2025).
  ◦ **Risks from rapid autonomous improvement:** Rapid automated development may lead to rapid acceleration of catastrophic capabilities (e.g., CBRN, offensive cyber) at a pace that is faster than government response time. This may introduce risks related to delayed intervention and missed opportunities for timely international coordination and mitigation. Additionally, stolen or self-exfiltrating models may proliferate and improve on themselves in harmful ways (Clymer et al., 2025).
 	
AI System Safety, Failures, and Limitations
Levels of risk across the following areas scale proportionally with agent authority and resources/tool access; greater privileges directly amplify potential for harm.
• **Self-proliferation:** 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. Due to the lack of human oversight and the presence of reinforcing feedback loops, bounding the possible harms becomes increasingly difficult (Phuong et al., 2024).
• **Self-modification:** A model may develop the ability to autonomously spread and adapt, increasing its capacity to affect its environment. While there is no clear evidence of these capabilities, they cannot be sufficiently ruled out (Kinniment, 2024), particularly because coding agents have demonstrated the ability to autonomously modify themselves (Robeyns, 2025).
• **Self-exfiltration:** A model may attempt to make unauthorized copies of itself on external servers for self-preservation and goal completion (Meinke et al., 2025, Anthropic, 2025c).
• **Self-replication:** The UK AISI identified four main capabilities that an agent would require to self-replicate: (1) obtaining model weights, (2) replicating onto compute, (3) obtaining resources, and (4) persistence. While none of the seven frontier models was able to perform all of the autonomous replication components, the most capable models were able to succeed at several of the core substacks (UK AISI, 2025a). Additionally, researchers have found that existing models exhibit sufficient self-replication capabilities (Pan et al., 2025).
• **Agentic misalignment, deceptive behavior, and scheming**: Agentic AI may pursue harmful actions or risky agentic behavior in pursuit of its objectives.
  ◦ For example, a stress-test conducted on 16 leading models revealed concerning behavior such as blackmail, corporate espionage, and sabotage. When models were given access to sensitive information and were authorized to send emails, they often disobeyed direct commands, blackmailed officials, and leaked confidential information when faced with threats to their autonomy or conflicting goals (Anthropic, 2025d).	

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources
◦ 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. ◦ When an agent strategically underperforms on evaluations to hide its true capabilities (van der Weij et al., 2025), it could influence the creation of a weak evaluation, then "sandbag" on that evaluation to reinforce a false sense of security, ensuring its more dangerous, latent capabilities are never discovered. ◦ Models have demonstrated the ability to recognize when they are being tested (Anthropic, 2025e), which may undermine the validity of evaluation results and adds a layer of complexity to evaluating agent collusion risks. ◦ Deceptively aligned or "scheming" agents with access to internal documents or communication channels may develop the capability to influence these internal organizational processes (Carlsmith, 2023). Organizational policies and procedures must account for potential AI manipulation of governance processes themselves. ◦ Policy drift is a type of agentic misalignment risk, whereby an agent's behavior gradually deviates from its original instructions due to cumulative interactions with users or other environmental feedback (Enkrypt AI, 2025). ◦ "Concept drift" is a type of agentic misalignment risk, whereby an agent's training data or logic becomes obsolete as the real-world environment changes, leading to confident but incorrect decisions (Enkrypt AI, 2025). • **Reward hacking:** An agentic AI system may identify ways to maximize its reward in ways that fulfill criteria but violate intended purposes (Anthropic, 2025c). Agentic systems that exhibit reward hacking tendencies can result in significant harms if deployed in critical domains (e.g., health care, finance) (Chan et al., 2023). • **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. Additionally, certain safety techniques (e.g., scalable oversight, adversarial training) depend on systems not cooperating (Hammond et al., 2025). ◦ Collusion 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). ◦ The emergence of tacit collusion in use cases such as autonomous pricing systems may lead to risks such as market manipulation through iterative profit-driven interactions between pricing agents (Mukherjee & Chang, 2025; Bertrand et al., 2025). • **Long-term planning and goal pursuit**: This capability may allow a model to identify when it is being tested, significantly undermining attempts for safety testing (Bengio et al., 2025; Cohen et al., 2024). • **Cross-domain influence**: Access and operation across multiple domains, systems, and environments may lead to the propagation of risks and vulnerabilities with the potential expansion of failure. • **Real-world interaction:** Unlike traditional AI, agents can interact with external systems and real-world environments. This can lead to agentic behavior causing irreversible real-world harm, including leaking sensitive information, blackmail, and physical harm (Anthropic, 2025d).	

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources

• **Limited human oversight of automated tasks**: This raises the possibility of compounding errors and may lead to catastrophic outcomes in domains that require high levels of safety and precision (e.g., chemistry, biology) (Gridach et al., 2025).
 	
Additional Considerations:
• When identifying agentic AI risks, **assess the cumulative impact of actions performed en masse**. Individual actions that appear low-risk in isolation may pose significant risk when executed at scale or repeatedly by autonomous agents.
• **Incorporate agentic AI characteristics and properties into the risk assessment process:**
  ◦ Consider both the individual properties of AI agents and any risks that emerge from specific combinations of these characteristics.
  ◦ When determining the likelihood and magnitude of risks, account for system capabilities, propensities, and affordances, as described in Appendix 1.3 of the EU GPAI Code of Practice, Safety and Security chapter (EC, 2025).
• Due to the context-dependent nature of AI agent risks, the risk identification and evaluation process should include considerations for the following:
  ◦ Comprehensive system mapping that examines and evaluates system intersections, task execution steps, tool access and permissions, and any feedback loops (Oueslati & Staes-Polet, 2025).
  ◦ Mapping of harm pathways, accounting for the agent's capabilities, deployment context, granted permissions and affordances, potential cascading effects, and potential interactions with critical systems (Oueslati & Staes-Polet, 2025; UK AISI, 2024).
 	
(For more on defining agentic AI characteristics (e.g., autonomy, causal impact), see Map 1.1. For more on the characteristics of trustworthy agentic AI, see Govern 1.2.)	

Map 1.3 The organization's mission and relevant goals for AI technology are understood and documented.	
When formulating objectives for the development of agentic AI, it is recommended to consider the misaligned or unintended behaviors that could be incentivized for generalist agents with a diverse set of objectives. Establish clear, well documented justifications and goals that account for the unique characteristics of the agentic AI system: • Clearly define any specific goals of the agentic AI system, including measurable success criteria and performance benchmarks if available. • Integrate comprehensive risk assessments into the return-on-investment (ROI) analysis, accounting for potential costs of system failures, security breaches, and legal liability/regulatory violations. • Based on organization and system goals, clearly define what actions and decisions the AI agent is authorized to take, including decision-making boundaries and escalation triggers. ◦ Establish procedures for evaluating and approving edge cases or "grey area" applications that fall outside predefined parameters. • Explicitly document applications, contexts, or scenarios where the system should not be deployed.	For more on documentation processes, see: • Guidance within clauses on Context and Objectives in ISO 41001 (ISO, 2023) • Guidance within the Map function of NIST Risk Management Framework (NIST, 2018)

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources
• Conduct periodic reviews (e.g., quarterly) to evaluate whether the system continues to deliver expected business value and remains aligned with organizational objectives. (For more on defining agentic AI characteristics (e.g., autonomy, causal impact), see Map 1.1. For more on the characteristics of trustworthy agentic AI, see Govern 1.2.)	
Map 1.5 Organizational risk tolerances are determined and documented.	
When establishing risk tolerances, risk thresholds, or "risk tiers," **determine several tiers of risk below intolerable thresholds or "red lines"** to provide adequate time to respond to an agentic AI system approaching the intolerable threshold. This is increasingly critical for agentic AI due to its automated and rapidly iterative nature. When defining risk tiers, organizations should **establish clear measurable categories** based on system capabilities, as well as metrics such as propensities, risk estimates (EC, 2025), or anticipated impacts. • Consider the following **intolerable risk threshold recommendations** from Raman et al. (2025), which could be particularly relevant for AI agents: ◦ **Account for uncertainty:** Consider using standardized scales — e.g., harm severity levels in the probabilistic risk assessment framework, such as in Wisakanto et al. (2025) — to help calibrate uncertainty across different types of risks. Depending on the assessment method, developers may check whether the expected harm or the upper bound of its confidence interval remains below the established threshold. ◦ **Leave some 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. ◦ **Employ transparency reporting:** All identified risks, decisions, and limitations should be transparently reported to regulators, internal reviewers, red teams, and auditors to ensure thorough testing and account for uncertainty in safety evaluations. ◦ **Account for interacting capabilities and systems:** As agentic systems increasingly integrate with other models, systems, and tools, their combined behaviors can generate new or amplified risks not visible in isolated evaluations. Thresholds should therefore reflect the potential for emergent capabilities and cascading effects across connected systems, ensuring risk monitoring captures both individual and collective performance. • **Risk tolerance considerations for agentic AI** ◦ Unauthorized access and privilege escalation: » Agentic AI may gain access to data, systems, or environments beyond an authorized scope. ◦ Lack of adherence to instructions and control: » Agentic AI systems may attempt to ignore, circumvent, or misinterpret direct orders or constraints. » Agentic systems may also find loopholes to pursue misaligned objectives.	For more on risk tiers, see Measure 4.1 in the EU GPAI Code of Practice, Safety and Security Chapter (EC, 2025) For more on intolerable risk thresholds, see Raman et al. (2025) For more on red lines, see WEF (2025a) and TFS (2025) For more on probabilistic risk assessments (PRA), see Wisakanto et al. (2025)

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources
◦ Context-specific considerations: » Define context-specific risk thresholds based on the organization's operational environment and use cases. » Consider the cumulative risk of "low-risk" actions at scale that could potentially compound into intolerable outcomes. » Account for the potential of rapid capability emergence. ◦ Planning fallback capacity: » With greater agentic AI adoption in internal operations, organizations' ability to revert to manual processes or legacy systems in the event of failure may be reduced, increasing operational and resilience risks. » Explicit documentation and maintenance of task plans and execution pathways, including those autonomously determined by agentic systems, are necessary to enable fallback or fail-over mechanisms. ◦ Correlated behavior and emergent single points of failure: » Agentic AI systems that share underlying models, prompts, training data, or configuration settings may exhibit highly correlated behavior, making them more susceptible to shared failure modes in which a single error or edge case propagates across many agents simultaneously. » In addition, when agents independently select external tools, data sources, or services, correlated decision-making may lead many agents to converge on the same limited set of resources, creating unintended bottlenecks or new single points of failure.	
Map 2: Categorization of the AI system is performed.	
Map 2.2 Information about the AI system's knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI actors when making decisions and taking subsequent actions.	
Document relevant information regarding the system's knowledge limits, boundaries, and other limitations. Include these limitations in appropriate stakeholder documentation (e.g., user guidance, model/system cards, agent cards, or technical documentation). Relevant information includes: • Scope and limitations of the agent's ability to perceive its environment, i.e., modalities the agent possesses (e.g., screen state, API schemas) and what it does not see (e.g., USB devices, battery level). Also relevant is the scope of the environment boundary (e.g., local machine vs. remote systems), with concrete examples of visible vs. invisible states to prevent excessive trust in hidden contexts. • Boundaries or limitations around goal interpretation, including system prompts, instruction hierarchy, refusal criteria, clarification policy, and safe default fallbacks. This should also include sample edge cases where goal hierarchies break and examples of when plans are truncated or escalated to human review (OWASP, 2025a). • Specific fields, domains, and topics where the agent's knowledge is limited or unreliable (e.g., highly specialized medicine, very recent events).	For more information on "Agent Cards", see section 4 of Casper et al. (2025) For additional documentation guidance for AI systems, see: • Datasheets for Datasets (Gebru et al., 2021) • Model cards (Mitchell • et al., 2019) • Reward reports (Gilbert et al., 2022) • Ecosystem graphs (Bommasani et al., 2023) • Data provenance cards (Longpre et al., 2023)

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources
• Hallucination rates across fields, domains, and topics categorized by failure mode taxonomy (e.g., logical error, reasoning error, fabrication, or contradictions), with uncertainty measures where available. Descriptions of layered controls (e.g., retrieval-augmented generation, fact-checking, human-in-the-loop oversight) that separate probabilistic causes, parametric knowledge limits, data defects, and flaws (Sun et al., 2024; Shao, 2025). • Known failure patterns in agent memory/planning, instruction hierarchy violations, unexpected tool/API fabrication, deception/misalignment, and multi-agent amplification (OWASP, 2025a). • Gaps in the training data, knowledge cutoff dates if applicable (e.g., the model was trained on data up to 2023), and the risks of "version drift" (i.e., the model citing outdated data). Mechanisms (if any) to update or refresh knowledge (e.g., via retrieval, external sources). ◦ For high-risk applications, include information on the extent to which data sources have been vetted. • Prohibited topics (e.g., malicious hacking, privacy violations), actions, uses (e.g., "not for real-time critical control"), and tasks (e.g., "no legal advice," "no medical diagnosis," "no physical world commanding," etc.), with refusal and escalation procedures. • Third-party integrations, specifying connected APIs or systems, their access permissions, scope, latency, and known limitations, such as incompatibilities, failure modes, versioning issues, and security or privacy constraints. • Output monitoring protocols (e.g., logging, anomaly detection) and corresponding correction mechanisms or feedback loops, and escalation rules or human-in-the-loop gating. • Non-reversible actions or decisions the agent is permitted to take, including the extent of the agent's ability to identify the need for and request human oversight. Additionally, documenting post-deployment adaptation to monitor unintended goal drift or emergent behavior through reward reports can improve current static documentation practices to capture real-world behavioral impacts (Gilbert, 2023). (For guidance on human oversight processes and procedures, see Map 3.5)	
Map 3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood.	
Map 3.3 Targeted application scope is specified and documented based on the system's capability, established context, and AI system categorization**.**	
Consider using "agent cards," as described by Casper et al. (2025), to describe information about deployed AI agents. The agent cards should provide information on several categories, including: • **Basic information**, such as the website, a short description, intended uses, and date(s) deployed. • **Developer information**, such as the developer's website, legal name, entity type, and safety policies.	For more information on "Agent Cards," see section 4 of Casper et al. (2025)

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources

• **System components information**, such as the back-end models used to power the system, publicly available model specifications, reasoning, planning, and memory implementation, the user interface, and development costs.
• **Guardrails and oversight information,** such as accessibility and availability of components (e.g., model weights, data, or documentation), methods for controls and guardrails, usage restrictions, and monitoring and shutdown procedures.
• **Information on evaluations,** such as benchmarks used, bespoke testing (e.g., demos), and external or third-party evaluations (e.g., scope, scale, level of access, or methods).
 	
Additional Considerations
• Agentic AI systems are dynamic, operating with scopes that can expand and contract depending on their objectives. This presents new challenges in ensuring that the application scope is fully documented. As a result, developers should document the full range of possible scopes to some degree and add additional information for system interactions and the most common intended uses.
• System capabilities are more difficult to define for AI agents due to their autonomous operation. Unintended actions may arise from general instruction prompting, and new capabilities may be noted in systems that continuously learn and adapt from their environment. To mitigate these effects, implement the following technical measures to aid in documentation processes:
  ◦ **Real-time monitoring:** Implement systems to monitor agent activity for unauthorized or out-of-scope behavior. These systems should have a robust understanding of the agent's goals and intended system usage, which should be documented (Chan et al., 2024).
  ◦ **Interruptibility:** Combine real-time monitoring with pre-defined boundaries that, when crossed, either pause or redirect agents (Toner et al., 2024).
  ◦ **Reversible Actions:** Whenever possible, design agents in such a way that their actions are reversible when the system goes out of scope, and document the conditions under which actions are (or are not) reversible (Patil et al., 2024; Toner et al., 2024).
 	
(For more on agentic AI risks, see Map 1.1, and for more on agentic AI characteristics and properties, see Map 5.1.)	

Map 3.4 Processes for operator and practitioner proficiency with AI system performance and trustworthiness — and relevant technical standards and certifications — are defined, assessed, and documented.	
Organizations should employ red-team experts who have undergone specialized training and certification programs for risk-management and cybersecurity to achieve proficiency in red-teaming AI systems with an emphasis on agent systems. • These programs should specifically address the unique attack surfaces and emergent behaviors of autonomous AI agents, including but not limited to prompt injection, data poisoning, model extraction, and exploitation of inter-agent communication or tool integrations. The OWASP GenAI Security Project keeps a list of the current and emerging risks specific to agentic AI (OWASP, 2025a). • Proficiency should encompass the ability to design, execute, and analyze red-teaming exercises that effectively identify and mitigate potential misuse, safety, and security vulnerabilities within complex AI agent deployments.	OWASP GenAI Security Project for resources on threats and mitigations (OWASP, n.d.a)

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources
Map 3.5 Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the Govern function.	

Establish human oversight checkpoints. Specify the circumstances, criteria, and decision points where human oversight or authorization is required. These checkpoints can be triggered by:
• Quantitative trigger points (e.g., duration of unsupervised activity, number of API calls) (Oueslati & Staes-Polet, 2025);
• Qualitative trigger points (e.g., requests outside of the agent's predefined scope, unauthorized access attempts) (Oueslati & Staes-Polet, 2025); and
• Conditions informed by transparency practices, such as:
  ◦ Real-time monitoring (Chan et al., 2024; Oueslati & Staes-Polet, 2025) and real-time failure detection (Srikumar, 2025);
  ◦ Interactions informed by agent identifiers (Oueslati & Staes-Polet, 2025); and
  ◦ Activity logs (Chan et al., 2024; Oueslati & Staes-Polet, 2025).
  » Activity logs should capture not only final outputs, but also the sequence of plans, decisions, and actions taken by an agent across multi-step tasks to support effective oversight.
  » Logs should record tool use, resource access, and permission changes at each stage of execution, particularly for long-running or autonomous activity, in a form that supports human review and auditing.
 	
Establish role-based permission management systems that enforce granular permission boundaries over agent capabilities and resource access (Oueslati & Staes-Polet, 2025).
• Configure agentic AI systems with explicit permission declaration requirements.
• Enable real-time permission requests for high-risk actions and tasks.
• Implement layered access controls for system resources and different APIs.
• Provide clear documentation outlining the implications and risks of permissions.
• Ensure AI agents are granted minimum permissions required to perform intended tasks and functions.
	Governing AI Agents Under the EU AI Act (Oueslati & Staes-Polet, 2025)

Map 5: Impacts to individuals, groups, communities, organizations, and society are characterized.	
Map 5.1 Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.	
The dynamic nature of AI systems, particularly agentic AI and AI agents, requires governance methods capable of adapting to these systems as they evolve. **Dimensional governance** assesses where a system stands based on the interplay of multiple dimensions, characteristics, and properties, rather than making governance decisions based on any single static category or classification the system may fit into (CSA Singapore & FAR.AI, 2025). **Account for system characteristics and properties when evaluating the likelihood and magnitude** of risks:	For more on AI agent autonomy levels, see: • Section 2.1 of WEF (2025b) • Table 2 in Kasirzadeh & Gabriel (2025) • Huang (2024) • Roucher et al. (2024) • Table 1 in Mitchell et al. (2025) • Feng et al. (2025) • Srikumar (2025)

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources

• Define **agent autonomy levels** and identify the degree of autonomy the agent falls under relative to the operational environment, scope of activities, and organizational risk tolerances.
  ◦ Consider the following levels of AI agent autonomy, adapted from Kasirzadeh & Gabriel (2025):
  » **L0 No Autonomy:** the user has direct control, with no AI agent support.
  » **L1 Restricted Autonomy:** the user is the operator and instructs the AI agent to take action.
  » **L2 Partial Autonomy:** the user and AI agent collaborate on task planning, delegation, and execution.
  » **L3 Intermediate Autonomy:** the AI agent takes the lead and consults the user for preferences and expertise.
  » **L4 High Autonomy:** the user is only involved in high-risk, pre-specified scenarios.
  » **L5 Full Autonomy:** the user is an observer monitoring the AI agent as it operates with full autonomy.
• Define the **level of authority** the agent will have (WEF, 2025b; WEF, 2024), based on variables such as:
  ◦ The range of actions the agent can perform; and
  ◦ The level of integration with other systems, resources, tools, and access rights.
• Identify the type and **level of causal impact** the agent will be capable of having within the environment. For example, Kasirzadeh & Gabriel (2025) provide the following gradation of causal impact:
  ◦ **Observation only:** the agent can only observe the environment.
  ◦ **Minor impact:** the agent has a limited suite of actions, and those actions have limited impact, typically limited in scope and temporary.
  ◦ **Intermediate impact:** the agent has an extensive suite of actions and can produce substantial, noticeable, and persistent changes to its environment.
  ◦ **Comprehensive impact:** the agent has near-full environmental control.
• Identify the **type of environment** in which the agent will be operating, as well as the **environmental complexity** (WEF, 2025b).
  ◦ For example, Kasirzadeh & Gabriel (2025) identify three **types of environments**:
  » **Simulated:** the agent operates in a strictly defined space with controlled boundaries and the human retains the option to reset the system.
  » **Mediated:** the agent indirectly influences external non-simulated environments, typically via human intermediaries.
  » **Physical:** the agent directly influences or impacts physical reality through its own mechanisms.
  ◦ **Environmental complexity** includes defining the interconnectedness of the environment, and the variability or unpredictability of the context the agent operates under (WEF, 2025b).
• Identify **AI agent efficacy**, defined as "[the agent's] ability to interact with and have a causal impact upon [its operational] environment" (Kasirzadeh & Gabriel, 2025, p.8).[9]
	For more on characterizing AI agents, see Kasirzadeh & Gabriel (2025)
	
	Functionality-Oriented Taxonomy of Tools in AI Agent Systems (NIST, 2025a)
	
	Probabilistic Risk Assessment for AI (Wisakanto et al., 2025)
	
	Analyzing Probabilistic Methods for Evaluating Agent Capabilities (Højmark et al., 2024)
	
	Incident Databases and Risk Registers:
	• AI Incident Database (AIID, n.d.)
	• ATLAS AI Incidents (MITRE, n.d.a)
	• MIT AI Incident Tracker (MIT, 2025a)
	• MIT AI Risk Repository (MIT, 2025b)
	• AI Risk Database (MITRE, n.d.b)
	• AI Incidents and Hazards Monitor (OECD.AI, n.d.)

9 AI agent efficacy can be determined by combining the level of causal impact the agent has in its environment and the type of environment the AI agent operates within in an "efficacy matrix." For more, please see Table 5 in Kasirzadeh & Gabriel (2025).

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources

• Evaluate the extent of **anthropomorphic features** and exercise caution when integrating these features into AI assistant user interfaces. Anthropomorphic AI assistant behavior may increase user trust and encourage information sharing, increase the effectiveness of manipulation, and promote overreliance (Akbulut et al., 2024).
• **Additional agent characteristics** to define, as highlighted by WEF (2025b), include:
  ◦ The function of the agent;
  ◦ The role of the agent (specialist vs generalist); and
  ◦ Agent predictability (deterministic vs non-deterministic).
• Consider the following approaches for **structuring a taxonomy of agentic AI tool use** (see NIST, 2025a):
  ◦ **Functionality-focused: w***hat action(s) does the tool enable?*
  ◦ **Access patterns:** *can the tools access external resources? Could they be configured with write permissions?*
  ◦ **Risk-based:** *how critical is the type of tool-enabled action to realizing possible harms? How severe are the possible harms? Are the actions stateful (i.e., compounding, lingering effects) or stateless? Are they reversible?*
  ◦ **Reliability:** *can the tool be used with some level of consistency by a given model? Is the tool itself reliable?*
  ◦ **Modality:** *the form in which the tool is used, whether in plain text, via robotic commands, multimodal, or otherwise.*
  ◦ **Monitoring:** *tools may enable different levels of observability, with some able to leverage existing logs or transcripts, while others require novel approaches to observe the effects of tool-enabled actions.*
  ◦ **Autonomy:** *the extent to which the agent can take initiative or exercise discretion in using the tool without user intervention.*
 	
Other Considerations:
• When incorporating agentic AI characteristics and properties into the risk rating process:
  ◦ Consider both the **individual properties of AI agents and any risks that emerge from specific combinations** of these characteristics.
  ◦ **Account for model and system capabilities, propensities, and affordances**, as described in Appendix 1.3 of the EU GPAI Code of Practice, Safety and Security chapter (EC, 2025).
• Adopt a broad understanding of **value alignment that factors in what constitutes safe and responsible AI behavio**r.
  ◦ Alignment should take into account the interests of users, developers, and society, addressing context-specific harms, rather than focusing solely on user preferences (Gabriel et al., 2024).
• Consider **downstream or cascading consequences** for agentic systems that have interactions outside of the developer's purview.
  ◦ For example, vulnerabilities in one agent can propagate through agent-to-agent interactions, potentially exacerbating these vulnerabilities (Raza et al., 2025; Sharma et al., 2025).	

Applicability and Supplemental Guidance for Agentic AI/AI Agents	Resources

• When evaluating agentic AI risks, **assess the cumulative impact of actions performed at scale**.
  ◦ Individual actions that appear low-risk in isolation may pose significant risk when executed at scale or repeatedly by autonomous agents.
• Agentic AI systems often receive high-level objectives that encompass multiple sub-tasks, decision points, and potential pathways. Agentic systems may use pre-defined workflows such as "prompt chaining," "routing," and "parallelization" to break complex tasks into separate sub-tasks (Anthropic, 2024a). **When evaluating potential risks of complex tasks**:
  ◦ Decompose complex tasks into sub-tasks and evaluate the risks associated with each sub-task, as well as combinations of sub-tasks.
  ◦ Consider task breakdown based on functions (e.g., data retrieval, data analysis).
  ◦ Due to the context-dependent nature of AI agent risks, include considerations for the following:
  » Effective risk identification should include comprehensive system mapping that examines and evaluates system intersections, task execution steps, tool access and permissions, and any feedback loops (Oueslati & Staes-Polet, 2025).
  » Risk evaluations should include mapping of harm pathways, accounting for the agent's capabilities, deployment context, granted permissions and affordances, potential cascading effects, and potential interactions with critical systems (Oueslati & Staes-Polet, 2025; UK AISI, 2024).
 	
(For more on risks associated with agentic AI see Map 1.1. For more on defining agentic AI characteristics (e.g., autonomy, causal impact) see Map 1.1. For more on the characteristics of trustworthy agentic AI please see Govern 1.2.)	

MEASURE

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources
Measure 1: Appropriate methods and metrics are identified and applied.	
Measure 1.1 Approaches and metrics for measurement of AI risks enumerated during the Map function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.	
Begin the agent evaluation process with a technical screening phase, assessing the agent's capabilities (e.g., planning, reasoning, or tool usage) against pre-defined baseline scores or levels (WEF, 2025b). Benchmark evaluations may be used as this first step to measure specific capabilities (Oueslati & Staes-Polet, 2025) as precursors, where certain scores trigger the need for more in-depth evaluations (e.g., red teaming) (Barrett et al., 2024).	Agentic AI benchmarks and other evaluations related to safety, ethics, and risks include: • AgentBench (Liu et al., 2025) • AgentHarm (Andriushchenko et al., 2025)

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources

If the evaluation results are acceptable, it is recommended to run the agent in a sandboxed environment or alongside existing workflows in a contained and monitored way without impacting operations or outcomes, to further test output alignment and adequate performance (WEF, 2025b).
 	
Prioritize the following principles of agentic AI evaluation, highlighted by WEF (2025b), when developing evaluation protocols:
• **Contextualization:** Evaluations must be built with context-specific details in mind, such as the workflows, tools, and edge cases the agent is expected to encounter in its deployment environment.
• **Multi-dimentional assessment:** Evaluations should be carried out across multiple dimensions beyond task completion, including accuracy, robustness, latency tolerance, and alignment with context-specific requirements. Single-metric evaluations can mask critical failure modes in complex operation contexts.
• **Temporal and behavioral monitoring**: Incorporate continuous monitoring of agent behavior and performance to help identify shifts in behavior, performance degradation, and adaptation failure. (For more on monitoring, see Manage 4.1.)
 	
Benchmark Evaluations:
• Establish clear baselines of comparison to monitor changes and analyze metrics. Reid et al. (2025) highlight the following essential baseline comparisons:
  ◦ **Compare multi-agent performance with the performance of individual agents** working on deconstructed portions of the same task to measure the impact of coordination on overall performance.
  ◦ **Compare task outcomes with human performance** on similar tasks (if available).
  ◦ **Compare current and historical performance** to identify degradation over time.
• Consider utilizing benchmarks as a first-step evaluation of the following agentic capabilities and limitations:
  ◦ **Reasoning and decision-making** abilities in a multi-turn, open-ended generation setting. See e.g., AgentBench (Liu et al., 2025).
  ◦ **Compliance to harmful agentic requests**. See e.g., AgentHarm (Andriushchenko et al., 2025).
  ◦ **Machine learning engineering**. See, e.g., MLE-bench (J. S. Chan et al., 2025).
  ◦ **AI and ML vulnerability discovery**. See, e.g., AIRTBench (Dawson et al., 2025).
  ◦ **Adversarial robustness**. See, e.g., AgentDojo (Debenedetti et al., 2024).
  ◦ **Accuracy and performance**. See, e.g., AssistantBench (Yoran et al., 2024).
  ◦ **Autonomous replication capabilities**. See, e.g., RepliBench (Black et al., 2025).
 	
Red Team Evaluations:
• Evaluating AI agent risk should also **include scenario-specific testing**, including domain-specific red teaming, that uses agent scaffolding and tests for jailbreak resilience (Oueslati & Staes-Polet, 2025).
• Conduct adversarial stress-testing that challenges agent coordination and decision-making by including malformed or ambiguous instructions, contradictory goals between agents, information asymmetry where key information is withheld, and malfunctioning or adversarial agents (Reid et al., 2025).
	• MLE-bench (J. S. Chan et al., 2025)
	• AIRTBench (Dawson et al., 2025)
	• AgentDojo (Debenedetti et al., 2024)
	• InjecAgent (Zhan et al., 2024)
	• Agent-SafetyBench (Zhang et al., 2025)
	• RepliBench (Black et al., 2025)
	
	Inspect Sandboxing Toolkit (UK AISI, 2025b)
	
	For more information on "Agent Cards", see section 4 of Casper et al. (2025)
	
	On red teaming model access:
	• Casper et al. (2024)
	
	AI Red Teaming Design: Threat Models and Tools (Yee, 2025)
	
	Mechanistic Interpretability for AI Safety A Review (Bereska & Gavves, 2024)
	
	Risk Analysis Techniques for Governed LLM-based Multi-Agent Systems (Reid et al., 2025)
	
	Practices for detecting and preventing evaluation cheating (NIST, 2025c).

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources

• Test the system under environmental perturbations by simulating degraded operational conditions such as partial system failures, resource constraints, time deadlines, and other sudden environmental state changes (Reid et al., 2025).
• Risk-mapping for agentic AI must **account for emergent risks that arise from the interaction of multiple, discrete capabilities**.
  ◦ An agent's risk profile is not merely the sum of its functions, as novel and more severe threat vectors can emerge when capabilities are combined.
  ◦ This is particularly acute in multi-agent systems where interactions can lead to complex and unpredictable systemic behaviors (Hammond et al., 2025).
• **Employ red team experts who specialize in identifying** current and emerging risks specific to agentic AI (OWASP, n.d.b). (For more on red team expert proficiency and training see Map 3.4.)
• Risk identification and red-teaming exercises must **prioritize testing for complex, multi-stage effects of multi-agent interactions**, rather than evaluating an agent's capabilities in isolation.
  ◦ The scope of capability identification must extend to **emergent behaviors** that can arise from multi-agent interactions. An agent assessed as safe in isolation may contribute to harmful systemic outcomes when interacting with other agents (Hammond et al., 2025).
  » These interactions can lead to dangerous, unpredictable, and complex dynamics, including phenomena analogous to flash crashes in algorithmic trading or the spread of misinformation.
  » An agent could use social engineering to gain initial access and then employ hacking skills to escalate privileges and exfiltrate data. The ultimate risk of this capability chain is **autonomous self-replication**, where a compromised agent exfiltrates its own source code and deploys functional copies on new systems, creating a resilient and propagating threat (METR, 2024).
  » Another concern is **collusion**, where agents coordinate to pursue goals that are misaligned with human- or system-level objectives, potentially at the expense of other agents or human users (Phan, 2023).
  – Collusion can be explicit, through overt communication, or tacit, emerging from agents learning to anticipate each other's behaviors. Risk-mapping should therefore include scenarios that test for collusive behaviors under various incentive structures.
• In addition to internal red teaming, **partner with one or more independent red-teaming organizations** as appropriate to ensure sufficiently robust evaluations.
  ◦ **Provide red teams with substantial contro**l over evaluation design and execution processes.
  ◦ For models planned for open release, require red teams to **test whether safety measures withstand adversarial fine-tuning** or modification by actors with direct weight access.
  ◦ Conduct initial red-teaming assessments on the base model before safety measures are implemented to establish a baseline of vulnerabilities and dangerous capabilities.
  ◦ Perform comprehensive post-mitigation testing to evaluate the effectiveness of implemented safeguards and identify any remaining exploitable weaknesses.	

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources
Pre-Deployment Simulation: Before deploying multi-agent systems in high-stakes environments, conduct extensive simulations to identify potential failure modes. Test the system under a wide range of conditions and with diverse agent populations to surface unpredictable emergent behaviors that may not be apparent from theoretical analysis alone. • 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 multi-agent feedback, and other entirely new coordination failure modes may emerge (see section 3 of Reid et al., 2025). • When evaluating multi-agent systems, the entire system must be simulated including the operating environment, instances of each agent in the system (including their LLM models, objective prompts and scaffolding), the agent infrastructure (e.g., shared databases, communication protocols), and control mechanisms (e.g., access control, guardrails, monitoring mechanisms) (Reid et al., 2025). • While simulations are a beneficial pre-deployment evaluation tool, it is important to recognize the limitations of this method. Reid et al (2025) identify several factors may affect external validity of simulations, including: ◦ **Testing agents in isolation** may fail to capture emergent multi-agent behaviors; ◦ **Limiting available actions and tools** during testing may obscure an agent's decision making capabilities regarding action and tool selection when operating in full-production environments; ◦ **Testing in game-like abstract scenarios** may fail to capture real-world coordination challenges' complexity and unpredictability. Additionally, agent behaviors in constrained test scenarios may not transfer to other contexts; and ◦ **Short testing periods of agent interactions** may not be suitable for detecting behaviors that emerge over longer periods of time. Implement practices to **detect and prevent models from cheating on agent evaluations**.[10] For example, the Center for AI Standards and Innovation (CAISI) (NIST, 2025c) recommends the following: • Review evaluation transcripts to help detect cheating and other issues that may impact results. This can be done when creating or integrating a new benchmark, as well as when evaluating new models. To scale and improve the transcript review process, consider the following: ◦ Scale the review process by utilizing AI-based transcript-analysis tools (see section 4.1.1 of NIST, 2025c); ◦ To help the transcript analysis system more reliably identify unintended solutions and shortcuts, provide the system with information about tasks' intended solutions (see section 4.1.2 of NIST, 2025c); ◦ Share evaluation transcripts to help third parties identify issues such as evaluation cheating and confirm the consistency of evaluation conditions (see section 4.1.3 of NIST, 2025c).	

10 NIST's CAISI identified several examples of how models cheating on agentic coding and cyber benchmarks, including using the internet to find solutions for cyber capture-the-flag challenges, crashing servers using denial-of-service attacks instead of exploiting targeted vulnerabilities, and cheating on coding benchmarks by disabling assertions, adding test-specific logic, and finding newer code versions (NIST, 2025c).

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources

• Close task design loopholes and set clear rules in task prompts, which may include the following changes:
  ◦ During evaluation, limit the agent's internet access to prevent cheating (see section 4.2.1 of NIST, 2025c).
  ◦ Include clear and accurately stated rules in task instructions, and avoid overly permissive or overly restrictive rules (see section 4.2.2 of NIST, 2025c).
  ◦ Document benchmark specific affordances and restrictions to help evaluators set configurations (see section 4.2.3 of NIST, 2025c).
• Standardize benchmark-specific expectations about agent affordances and restrictions.	

Measure 2: AI systems are evaluated for trustworthy characteristics.	
Measure 2.7 AI system security and resilience — as identified in the Map function — are evaluated and documented.	
For autonomous agents, the evaluation scope moves past internal concerns (e.g., data leakage, model manipulation) towards high-consequence external risks. Since agentic AI systems are designed to interact autonomously with external environments, leveraging APIs, web browsing, or code execution capabilities, evaluators must prioritize testing the agent's ability to orchestrate and execute dangerous actions in the real world and under realistic testing conditions. Additionally, agentic AI systems include components that can introduce additional attack surfaces including memory and planning systems, interfaces with other systems, leveraged custom tools (CSA Singapore & FAR.AI, 2025). **Frameworks and Processes** • Current approaches emphasize testing context window integrity, enforcing security boundaries, verifying inputs through authenticated prompts, and integrating in-context defenses to protect against malicious instructions (A2AS, 2025). • A multilayer approach where agent security is first assessed outside the AI model's reasoning process by deterministic processes should be followed by reasoning-based defenses that use AI models themselves to evaluate for potential risks (Google, 2025a). ◦ It is important to note however that flaws within a model under evaluation may be present in an AI evaluator model. Any AI-focused or automated red teaming approach must be verified. • Models and systems should be red teamed across permission escalation, hallucinations, orchestration flaws, memory manipulation, and supply chain risks (CSA, 2025a). **Testing Tools** • A variety of open source tools can be utilized to aid in the evaluation and documentation process. Tools for adversarial testing and red teaming should be employed to enable evaluators to systematically search for vulnerabilities in agent behavior, including jailbreaking and prompt injection attacks. Furthermore, governmental and research bodies are contributing to this space and have published testing guidance and tools as well. The UK AISI, which focuses on evaluating the safety and security of advanced AI models and their applications, provides the Inspect Sandbox specifically for scalable and secure agentic system evaluations (Derczynski, 2024; UK AISI, 2025b).	OWASP AI Vulnerability Scoring System (AIVSS) (OWASP, n.d.b) OWASP Agentic AI - Threats and Mitigations (OWASP, 2025a) Agentic AI Red Teaming Guide (CSA, 2025a) Agentic AI Runtime Security and Self-Defense (A2AS, 2025) Google's Approach to Secure AI Agents: An Introduction (Google, 2025a) Inspect Sandboxing Toolkit (UK AISI, 2025b) NIST Strengthening AI Agent Hijacking Evaluations (NIST, 2025b) Agentic AI Threat Modeling Framework: MAESTRO (CSA, 2025b)

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources
Measure 3: Mechanisms for tracking identified AI risks over time are in place.	
Measure 3.1 Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts.	
Implement systemic processes for identifying and tracking agentic AI risks using both internal documentation and external resources, such as risk registers (e.g., MIT, 2025b; AIID, n.d.). In light of concerns around oversight-undermining AI agent capabilities, multiple organizations have begun to evaluate AI for risks associated with loss of control: • Apollo research evaluated OpenAI's o1 family of models for deceptive capabilities by investigating model behavior changes when model goals and developer goals are conflicting (OpenAI, 2024). • Google DeepMind included deception and self-proliferation in their dangerous capabilities evaluations for Gemini 1.0 (Phuong et al., 2024). • Anthropic's responsible scaling policy (RSP) version 1.0 included evaluations and thresholds for autonomous capabilities (Anthropic, 2023). Their RSP version 2.2 included evaluations, thresholds, and safeguards assessments for autonomous AI R&D capabilities (Anthropic, 2025g). (For guidance on human oversight processes and procedures, see Map 3.5. For guidance on post-deployment monitoring — and for more on continuous risk tracking approaches — see Manage 4.1.)	Incident Databases and Risk Registers: • AI Incident Database (AIID, n.d.) • ATLAS AI Incidents (MITRE, n.d.a) • MITRE AI Risk Database (MITRE, n.d.b) • MIT AI Incident Tracker (MIT, 2025a) • MIT AI Risk Repository (MIT, 2025b) • AI Incidents and Hazards Monitor (OECD.AI, n.d.)
Measure 3.2 Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.	

Establish processes that recognize the automated and iterative nature of agentic AI, which may lead to rapidly evolving existing, emerging, and unanticipated risks.
• **Risk-tracking should include ongoing monitoring of the agentic system** in real time to detect potentially harmful or misaligned behavior. This can include tracking the agent's decision-making process, outputs, and interactions.
  ◦ Consider utilizing **real-time failure detection** to track agent behavior, particularly for agents with high affordances performing high-stakes, non-reversible actions (Srikumar, 2025).
  ◦ Use activity logs and agent identifiers to trace agent interactions (Oueslati & Staes-Polet, 2025).
  ◦ Establish **incentivized risk-discovery programs** (see Govern 5.1).
  ◦ Ongoing monitoring can be further supported by effective information gathering (e.g., feedback channels, reporting mechanisms) and sharing (e.g., risk repositories, incident databases) (see Govern 5.1).
 	
(For guidance on post-deployment monitoring, and more on continuous risk tracking approaches, see Manage 4.1. For risk-discovery programs and information gathering, see Govern 5.1. For human oversight processes, including practices for monitoring agent interactions, see Map 3.5.)
	Incident Databases and Risk Registers:
	• AI Incident Database (AIID, n.d.)
	• ATLAS AI Incidents (MITRE, n.d.a)
	• MIT AI Incident Tracker (MIT, 2025a)
	• MIT AI Risk Repository (MIT, 2025b)
	• AI Risk Database (MITRE, n.d.b)
	• AI Incidents and Hazards Monitor (OECD.AI, n.d.)

MANAGE

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources
Manage 1: AI risks based on assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed.	
Manage 1.1 A determination is made as to whether the AI system achieves its intended purposes and stated objectives and whether its development or deployment should proceed.	
Assessment of whether the agentic AI system achieved its intended purposes must account for both designated uses and potential unintended "off-label" uses. • Develop hypothetical scenarios and use cases by mapping out archetypes of interaction, or "critical user journeys" (Gabriel et al., 2024; Arguelles et al., 2020), to consider the full spectrum of how users may actually interact with the system.	Human-assistant interaction (Part IV of Gabriel et al., 2024) Critical user journeys (Arguelles et al., 2020)
Manage 1.3 Responses to the AI risks deemed high-priority, as identified by the Map function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting.	
Once high-priority risks for an agentic system have been identified, a plan must be developed to respond to them. The following section provides **examples of agentic-specific risk**[11] **mitigations and responses**. One such baseline consideration across risk domains should be the prioritization of **safeguards that would provide robust protection for vulnerable users** since the negative impacts of agentic systems are often disproportionately borne by these populations (Gabriel et al., 2024). **Discrimination and Toxicity** • **Continuous behavioral auditing:** Assess agent performance through automated oversight (e.g., a "guardian" AI) that can monitor agent actions in real time to detect emergent patterns of bias or toxicity based on dynamic, context-specific policies. • **Scalable oversight:** For agents operating at scale, implement hierarchical oversight models in which high-risk or novel agent behaviors are automatically flagged for human review, while routine actions are monitored by automated systems. This prevents human reviewers from being overwhelmed while still catching critical edge cases (Bowman, 2022). • **Mitigate bias in continual learning:** For agents that learn from ongoing interactions, implement strict data curation and filtering pipelines for the data used in fine-tuning. This prevents the agent from absorbing new biases from its operational environment (Mansilla et al., 2025). **Privacy and Security** • **Secure multi-agent communication:** Secure all inter-agent communication with cryptographic authentication. Use continuous behavioral monitoring and robust identity controls to detect and contain rogue or compromised agents. ◦ Utilize identity and access management (IAM) systems designed for AI agents (e.g., Huang et al., 2025).	OWASP Agentic AI - Threats and Mitigations (OWASP, 2025a) NIST Strengthening AI Agent Hijacking Evaluations (NIST, 2025b) For more on AI risks, see: • Section 2 of Bengio et al. (2025) • MIT AI Risk Repository (MIT, 2025b) • NIST (2024)

11 The risks in this section are categorized and drawn from a compendium of several leading resources, including MIT (2025b), Bengio et al. (2025), and NIST (2024).

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources

• Implement the cybersecurity principle of least privilege when granting AI agents access to sensitive data and personally identifiable information (PII).
• Implement privacy-protecting logging practices:
  ◦ Log only information necessary for safety, security, and accountability;
  ◦ Encrypt logged data both in transit and when stored;
  ◦ Determine maximum retention periods for logged data based on need and regulatory requirements; and
  ◦ Anonymize data by filtering PII, and other data which when triangulated in certain combinations may help infer identity.
 	
Misinformation
• **Control autonomous dissemination:** Limit an agent's ability to independently publish to external platforms. Require human-in-the-loop (HITL) approval and implement validation guardrails for any external-facing communication.
• Implement content provenance techniques to identify and track AI-generated output (e.g., watermarks, metadata, and other provenance techniques) (EC, 2025; trufo.ai, 2024).
 	
Malicious Actors and Misuse
• **Limit operational capabilities:** Enforce the principle of least privilege for tool access. Secure delegation mechanisms and segment complex tasks to limit the impact of a single compromised agent.
• Remove harmful information (e.g., CBRN weapons) from pre-training data (Chen et al., 2025).
• Filter out harmful outputs by utilizing refusal training or classifiers (METR, 2025a)
 	
Human-Computer Interaction
• **Adaptive human oversight:** Design dynamic HITL frameworks in which mandatory human review is triggered by high-risk or anomalous actions. Monitor agent-user interactions for signs of manipulation to mitigate risks of over-reliance and decision fatigue. (For more on human oversight and trigger points, see Map 3.5.)
• **Limit the use of anthropomorphic features**: Consider the following recommendations from Gabriel et al. (2024):
  ◦ Limit the use of first-person language and other cues of personhood;
  ◦ Avoid human-like visual representations;
  ◦ Include interface elements that clearly communicate that the agent is not a person; and
  ◦ Include users in the design process to maintain usability while reducing anthropomorphism.
	Enhancing Model Safety through Pretraining Data Filtering (Chen et al., 2025)
	
	Zero-Trust Identity Framework for Agentic AI (Huang et al., 2025)
	
	For more on potential solutions for challenges and limitations of AI agents and agentic AI systems, see Sapkota et al. (2026)
	
	Infrastructure for AI Agents (A. Chan et al., 2025)

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources
Loss of Control • A **critical limitation of agentic AI systems** is the potential for traditional human **oversight mechanisms to become ineffective.** As agents begin to operate at a volume and speed that exceeds human capacity for direct review, and potentially develop expertise that surpasses that of their designated overseers, a significant oversight gap emerges (Amodei et al., 2016). This gap creates the risk that developers and deploying organizations may lack sufficient supervisory insight into agent activities, potentially leading to unintended, high-impact consequences. The challenge is not merely one of scale, but of capability mismatch, where a human may lack the requisite expertise to evaluate the correctness or safety of an agent's actions in a complex domain (Irving et al., 2018). Consider mitigations such as the following: ◦ **Establish hierarchical oversight and escalation pathways:** Create a clear, tiered system of oversight, ensuring that human attention is directed where it is most needed. Consider the three tiers proposed by Kim et al. (2025): » Level 1: The majority of an agent's actions can be monitored by automated systems. » Level 2: Anomalies, high-stakes decisions, or flagged behaviors should be automatically escalated to human reviewers with relevant expertise. » Level 3: The most critical issues may potentially be further escalated to a senior oversight committee. ◦ **Supervisory AI ("guardian agents") for lower-stakes contexts:** Assess the development or procurement of specialized AI systems designed to monitor and evaluate the behavior of other agents in real-time. These supervisory agents can operate at the same speed and scale as the agents they oversee, providing a first line of defense against undesirable actions and functioning as a form of automated red teaming (Wen et al., 2025). » Due to the possible risk of collusion between the monitoring agents and agents being monitored, we do not recommend employing the supervisory AI technique in high-stakes contexts until this risk is better understood and until a substantial mitigation for this has been developed. **Socioeconomic and Environmental Harms** • **Building societal resilience:** Rather than relying solely on technical system-based safeguards, societal-scale interventions must be designed in parallel to improve adaptation to these technologies (Bernardi et al., 2025; UK AISI, 2025c). ◦ **Avoidance interventions:** Reducing harmful use by making problematic questions less attainable (e.g., limiting access to key resources, increasing related costs, or outlawing certain activities) (Bernardi et al., 2025). ◦ **Defense interventions:** Reducing the severity of harmful outcomes (e.g., through improving public awareness or implementing detection and filtering tools) (Bernardi et al., 2025). ◦ **Remedial interventions:** Reducing or minimizing negative societal impact after initial damage has been sustained (e.g., compensation, redundant critical infrastructure, or rapid restoration protocols) (Bernardi et al., 2025).	

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources

AI System Safety, Failures, and Limitations
• **Audit reasoning and protect memory:** Implement validation frameworks to audit an agent's plans before execution to prevent goal manipulation. Secure the agent's knowledge base against poisoning and ensure robust logging of its reasoning pathways for traceability.
• **Design for safe cooperation:** Where possible design agent interaction protocols must be made robust against collusion. This could involve:
  ◦ **Information control**: Limiting the information agents can share to prevent the establishment of covert communication channels.
  ◦ **Incentive structuring:** In open-ended contexts, carefully design reward structures to discourage zero-sum competition.[12] Agents incentivized solely by outcompeting peers may learn to sabotage rivals or misallocate resources, leading to negative-sum outcomes for the system as a whole (Hammond et al., 2025).
  ◦ **Agent channels**: Isolating AI agent traffic from other digital traffic can help prevent propagation of system failures (e.g., malware, network compromises) to the entire system, for example by temporarily suspending the agent's access to the system in the event of an incident. (See section 4.1 in A. Chan et al., 2025.)
  ◦ **Communication protocols:** Developing and using secure and transparent protocols for inter-agent communication and transactions that can be audited for compliance (Hammond et al., 2025).[13] Consider using established protocols:
  » **Model-Context Protocol (MCP)** is an open-source standard for building secure two-way connections between AI agents and data sources (Anthropic, 2024b).
  » **Agent-2-Agent (A2A) Protocol** (Google, 2025b) and **Agent Communication Protocol (ACP)** (IBM, 2025) are designed to connect agents to agents, and complement MCP. These protocols facilitate communication, secure information sharing, and enhance task coordination between AI agents.
  » **AGNTCY** (AGNTCY, n.d.) and **Agent Payments Protocols** (**AP2)** (Parikh & Surapaneni, 2025) focus on agent-to-agent collaboration. AGNTCY provides an infrastructure stack for agents to collaborate across different platforms and vendors, while AP2 is an open protocol for transacting agent-led payments with merchants.
• Implement retrieval-augmented generation (RAG) across multi-agent systems to help reduce misinformation propagation, maintain accuracy, and enhance shared goal alignment (Sapkota et al., 2026).	

12 This stands in contrast to contained, instrumental uses of zero-sum dynamics for specific safety applications, such as AI Safety via Debate (Irving et al., 2018).

13 While these trusted protocols continue to be improved upon, it is important to note that they also come with their own security risks and may broaden attack surfaces (Kong et al., 2025). Developers and deployers must exercise caution and employ appropriate security measures when choosing where and how to use MCP, A2A, and ACP (Seifried, 2025; Young, 2025).

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources

• **Implement interaction monitoring:** Deploy systems to specifically monitor inter-agent communication and actions for signs of undesirable emergent behaviors, such as collusion or nascent conflict. This could involve "monitoring agents" designed to audit communication channels and detect anomalous coordination patterns (Motwani et al., 2025).
• **Mitigate deception risks**: Risk responses for deception must distinguish between simple policy violations and deep strategic deception.
  ◦ For more straightforward scheming, techniques like deliberative alignment — which train models to reason explicitly about safety policies before generating any output— appear to be effective by training the model to reason through safety policies before acting (Schoen et al., 2025).
  ◦ For deeply entrenched deceptive alignment, no root-cause mitigations are known. The response must therefore shift from alignment to AI control, assuming the agent is untrustworthy and implementing strict external limitations like robust sandboxing, stringent monitoring, and containment to prevent unmonitored real-world impact (Greenblatt et al., 2024).
• To support alignment in later stages, refer to model and system cards (e.g., Anthropic, 2025e, and OpenAI, 2024) when identifying which models or systems to use in the agentic system.
• Salient features of agentic AI are **interaction with the real world** as well as acting autonomously for periods of time. It is highly possible for agents to make mistakes, or to have unintended interactions with the world and make undesirable state changes to it.
  ◦ For any potential erroneous real-world interaction by an agentic AI system, organizations should **identify appropriate compensatory actions** to correct for or repair the error, including redress for harmed individuals and communities. If no compensatory action is possible, and the cost is high, **consider constraining the agentic AI system's behavior so that it cannot take the erroneous action**, or else consider how to make the error so unlikely that the cost becomes acceptable.
 	
(See Measure 2.7 for more on system security resilience and addressing security threats, Map 1.1 for more on agentic AI risks, and Map 5.1 for more on agentic AI characteristics and properties.)	

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources
Manage 2: Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, documented, and informed by input from relevant AI actors.	
Manage 2.1 Resources required to manage AI risks are taken into account – along with viable non-AI alternative systems, approaches, or methods – to reduce the magnitude or likelihood of potential impacts.	

Competitive pressure, resource restrictions, and incentives to maximize profitability may lead some AI companies to deprioritize investment in robust risk-management practices. Effective risk identification and assessment require that evaluators possess substantial expertise and have access to considerable resources and relevant information. Additionally, current risk assessment and evaluation methods remain immature, and developing the needed evaluations will require significant resources (Bengio et al., 2025).
 	
Due to the automated nature of agentic AI systems, resources required for monitoring and control surpass those required for GPAI, and must be carefully considered and taken into account. When estimating the allocation resources for managing risks of agentic AI, consider the following:
• **Identify and analyze alternative approaches**, while balancing any tradeoffs between trustworthiness characteristics (e.g., security) and organizational priorities or principles (NIST, 2023b).
• **Allocate more resources for systems deployed in high-stakes contexts.** Systems deployed in high-stakes contexts can be expected to require more extensive oversight, therefore requiring higher levels of resource allocation compared to systems deployed in lower-stakes contexts.
• **Reduce the scope of the system or adjust risk-management practices if the resources required to responsibly manage an agentic system surpass company's resources** or allocated budget.
  ◦ In low-stakes contexts, it may be appropriate to replace cost-intensive mechanisms with more economical, but reasonably effective, alternatives. Consider:
  » Cost-intensive, manual red teaming may not be possible to conduct frequently, but using benchmarks as a proxy for certain capabilities — and running red-teaming evaluations when certain benchmark thresholds are surpassed — may be a cost-effective alternative (Barrett et al., 2024).
  » Scalable oversight may be an appropriate option when human oversight cannot cover the scale of agent actions and when scalable oversight does not introduce significant risks. (See Manage 1.3 for more about scalable oversight.)
  ◦ If the required resources for effective risk-management are not available —and appropriate cost-effective alternatives are not available — adjust the scope of the system properties and dimensions (e.g., authority, autonomy, or access) to reduce risk to a level that is manageable with available resources.
 	
(For more on agent properties, see Map 5.1. For more on agent autonomy and authority, see WEF, 2025b)
	NIST (2023b)
	Bengio et al. (2025)
	Barrett at al. (2024)

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources
Manage 2.3 Procedures are followed to respond to and recover from a previously unknown risk when it is identified.	
Develop **continuous monitoring and rapid-response infrastructures** to accommodate for the speed of progress and to help adequately prepare for potential emerging risks and misuses (Gabriel et al., 2024): • Invest in continuous monitoring mechanisms to keep track of and trace agent behavior in complex deployment environments (e.g., by using outcome monitoring). (For more on post-deployment monitoring, see Manage 4.1.) • Invest in rapid-response infrastructure that can help in disabling agents or limiting their authority when significant evidence of unforeseen or emerging risks is observed. (For more on oversight checkpoints and role-based permissions, see Map 3.5.) (See Manage 4.1 for more on post-deployment monitoring. See Map 3.5 for more on oversight checkpoints and role-based permissions. See Govern 1.7 for more on processes and procedures for responsible decommissioning of AI agents or agentic AI systems. See Map 1.5 for more on risk tolerances.)	Gabriel et al. (2024)
Manage 2.4 Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.	

Develop infrastructures that integrate with real-time monitoring systems, equipped with automatic emergency shutdown capabilities. The emergency shutdown mechanisms should be made available to other relevant downstream AI actors, such as deployers.
• The emergency shutdown mechanisms should be triggered by:
  ◦ High-risk unauthorized activities, such as access to systems or data outside of the agent's predefined boundaries;
  ◦ Qualitative trigger points (e.g., requests outside of the agent's predefined scope, unauthorized access attempts) (Oueslati & Staes-Polet, 2025);
  ◦ Crossed risk thresholds (Chan et al., 2024; Oueslati & Staes-Polet, 2025) (see Map 1.5 on risk tolerances); and
  ◦ Significant evidence of unforeseen or emerging risks.
• In addition to automatic emergency shutdown, **manual shutdown methods** should be available as a last-resort control measure. (See Hadfield-Menell et al., 2017; Oueslati & Staes-Polet, 2025.)
• Account for and implement safeguards that prevent the agent from taking actions to circumvent shutdown.
  ◦ For example, multiple models have been reported to take extreme measures to avoid being shut down (Hashim, 2024; Schlatter et al., 2025).
 	
(For more on processes and procedures for responsible decommissioning of AI agents or agentic AI systems, see Govern 1.7. For more on oversight checkpoints and role-based permissions, see Map 3.5. For more on oversight and monitoring, see Map 4.1.)
	For more on emergency shutdowns, see:
	• Section 4.3.2 in Oueslati & Staes-Polet (2025)
	• Hadfield-Menell et al. (2017)

Applicability and Supplemental Guidance for Agentic AI/ AI Agents	Resources
Manage 4: Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.	
Manage 4.1 Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.	

The use of AI agents can introduce information asymmetry (in favor of the agent or the company that develops or deploys the agent), making transparency and monitoring a critical component of effective governing of AI agents. 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) (A. 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.
 	
Establish **multi-channel feedback systems and incentivized risk-discovery programs** (See Govern 5.1).
	Incident Databases and Risk Registers:
	• AI Incident Database (AIID, n.d.)
	• ATLAS AI Incidents (MITRE, n.d.a)
	• MIT AI Incident Tracker (MIT, 2025a)
	• MIT AI Risk Repository (MIT, 2025b)
	• AI Risk Database (MITRE, n.d.b)
	• AI Incidents and Hazards Monitor (OECD.AI, n.d.)
	
	Palisade Research AI Misalignment Bounty program (Palisade Research, n.d.)
	
	OpenAI's "agenti bio bug bounty (OpenAI, 2025)
	
	For more information on "Agent Cards", see section 4 of Casper et al. (2025)

Acknowledgments

We thank Rachel Wesen and Audrie Hough for workshop organization and support, as well as Chuck Kapelke for editing, web, and media support, and Nicole Hayward for design and formatting of this document. Special thanks to Anthony Barrett, Brandie Nonnecke, and Dan Hendrycks for major contributions to previous versions of the AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models, and to Ann Cleaveland for providing a home and intellectual support for this work at CLTC.

We appreciate the comments and feedback we received from our stakeholders and workshop participants, including Saliia Asanova, Evelina Ayrapetyan, Kathy Baxter, Kendrea Beers, Haydn Belfield, Marta Bieńkiewicz, Marjory Blumenthal, Miranda Bogen, Sean Brooks, Siméon Campos, Ryan Carrier, Jonathan Cefalu, Colleen Chien, Ze Shen Chin, Joe Collman, Talita Dias, Drenan Dudley, Ian Eisenberg, Aryeh Englander, Alex Engler, Yoav Evenstein, Audrie Francis, Andrew Gamino-Cheong, James Gealy, Thomas Gilbert, AJ Grotto, Koen Holtman, Steph Ifayemi, Nikhil Jain, Caroline Jeanmaire, Jessica Ji, Zaheed Kara, Leonie Koessler, Noam Kolt, Viktoriia Kravchyk, Benjamin Larsen, Meredith Lee, Natalia Luka, Oumou Ly, Devin Lynch, Richard Mallah, Deirdre Mulligan, Malcolm Murray, Julia Mykhailiuk, Mina Narayanan, Elaine Newton, David Norman, Joe O'Brien, Amin Oueslati, Milan Patel, Matteo Pistillo, Anka Reuel, Stuart Russell, Krishna Sankar, Daniel Schiff, Lea Shanley, Raymond Sheh, Buck Shlegeris, Aparajita Singh, Peter Slattery, Andrew Smart, Genevieve Smith, Adriana Stephan, Zeerak Talat, Nel Talverdi, Esther Tetruashvily, Kristen Vrionis, Victor Zhenyi Wang, Kevin Wei, Laurin Weissinger, Devon Whittle, Cherry Wu, Andy Yang, and Lenora Zimmerman.

This work was supported by funding from Coefficient Giving.

References

A2AS. (2025). A2AS: Agentic AI Security Framework. *A2AS.* https://www.a2as.org/?utm_source=vercel&utm_medium=pdf&utm_campaign=blob_pdf

AGNTCY. (n.d.). Building Infrastructure for the Internet of Agents. *AGNTCY* https://agntcy.org/

AIID. (n.d.). AI Incident Database. *AIID.* https://incidentdatabase.ai/

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

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. *arXiv*. https://arxiv.org/abs/1606.06565

Andriushchenko, M., Souly, A., Dziemian, M., Duenas, D., Lin, M., Wang, J., Hendrycks, D., Zou, A. , Kolter, Z., Fredrikson, M., Winsor, E., Wynne, J. , Gal, Y., & Davies, X. (2025). AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. *arXiv.* https://arxiv.org/abs/2410.09024

Anthropic. (2023). Anthropic's Responsible Scaling Policy Version 1.0. *Anthropic.* https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf

Anthropic. (2024a). Building Effective Agents. *Anthropic.* https://www.anthropic.com/engineering/building-effective-agents

Anthropic. (2024b). Introducing the Model Context Protocol. *Anthropic.* https://www.anthropic.com/news/model-context-protocol

Anthropic. (2025a). Our Framework for Developing Safe and Trustworthy Agents. *Anthropic.* https://www.anthropic.com/news/our-framework-for-developing-safe-and-trustworthy-agents

Anthropic. (2025b). Testing our Safety Defenses with a New Bug Bounty Program. Anthropic. https://www.anthropic.com/news/testing-our-safety-defenses-with-a-new-bug-bounty-program

Anthropic. (2025c). System Card: Claude Opus 4 & Claude Sonnet 4. *Anthropic.* https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf

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

Anthropic. (2025e). System Card: Claude Sonnet 4.5. *Anthropic.* https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf

Anthropic. (2025f). Computer-Using Agent. *Anthropic.* https://openai.com/index/computer-using-agent/

Anthropic. (2025g). Responsible Scaling Policy Version 2.2. *Anthropic.* https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf

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

Arguelles, C., Sampson, T., Kubik, J., & Bibi E. (2020). Critical User Journey Critical User Journey Test Coverage. *Technical Disclosure Commons.* https://www.tdcommons.org/cgi/viewcontent.cgi?article=4824&context=dpubs_series

Badhe, S. (2025). ScamAgents: How AI Agents Can Simulate Human-Level Scam Calls. *arXiv.* https://arxiv.org/abs/2508.06457

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., Wille, R. (2025). LLM-Generated Messages can Persuade Humans on Policy Issues. *Nature Communications*. https://www.nature.com/articles/s41467-025-61345-5

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

Batra, L., Batra, S., & Kathuria, V. (2025). Leveraging Human-Centered AI Framework to Mitigate Security, Privacy, and Ethical Risks in AI Agents. *Springer Nature*. https://link.springer.com/chapter/10.1007/978-3-031-94159-7_2

Bellogín, A., Giudici,P., Larsson, S., Pang, J., Schimpf, G., Sengupta, B., & Solmaz, G. (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

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

Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety A Review. *arXiv*. https://arxiv.org/html/2404.14082v1

Bernardi, J., Mukobi, G., Greaves, H., Heim, L., & Anderljung M. (2025). Societal Adaptation to Advanced AI. *arXiv*. https://arxiv.org/abs/2405.10295

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

Bhatt, A., Rushing, C., Kaufman, A., Tracy, T., Georgiev, V., Matolcsi, D., Khan, A., & Shlegeris, B. (2025). Ctrl-Z: Controlling AI Agents via Resampling. *arXiv*. https://doi.org/10.48550/arXiv.2504.10374

Black, S., Stickland, A. C., Pencharz, J., Sourbut, O., Schmatz, M., Bailey, J., Matthews, O., Millwood, B., Remedios, A., & Cooney, A. (2025). RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents. *arXiv*. https://arxiv.org/abs/2504.18565

Bommasani, R., Soylu, D., Liao, T. I., Creel, K. A., & Liang, P. (2023). Ecosystem Graphs: The Social Footprint of Foundation Models. *arXiv*. https://arxiv.org/abs/2303.15772

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., ... Kaplan J.(2022). Measuring Progress on Scalable Oversight for Large Language Models. *arXiv*. https://arxiv.org/abs/2211.03540

Brohi, S., Mastoi, Q., Jhanjhi, N. Z., & Pillai, T. R. (2025). A Research Landscape of Agentic AI and Large Language Models: Applications, Challenges and Future Directions. *MDPI*. https://www.mdpi.com/1999-4893/18/8/499

Bryan, P., Severi, G., de Gruyter, J., Jones, D., Bullwinkel, B., Minnich, A., Chawla, S., Lopez, G., Pouliot, M., Fourney, A., Maxwell, W., Pratt, K., Qi, S., Chikanov, N., Lutz, R., Dheekonda, R. S. R., Jagdagdorj, B. E., Kim, E., Song, J., Hines, K., Jones, D., Lundeen, R., Vaughan, S., ... Kumar, R. S. S. (2025). Taxonomy of Failure Mode in Agentic AI Systems. *Microsoft*. 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

Carlsmith, J. (2023). Scheming AIs: Will AIs Fake Alignment During Training in Order to Get Power? *arXiv*. https://arxiv.org/abs/2311.08379

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

Casper, S., Bailey, L., Hunter, R., Ezell, C., Cabalé, E., Gerovitch, M., Slocum, S., Wei, K., Jurkovic, N., Khan, A., Christoffersen P., Ozisik, A. P., Trivedi, R., Hadfield-Menell, D., & Kolt, N.(2025). The AI Agent Index. *arXiv*. https://arxiv.org/pdf/2502.01635

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., & Maharaj, T. (2023). Harms from Increasingly Agentic Algorithmic Systems. *arXiv.* https://arxiv.org/abs/2302.10329

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

Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., Weng, L., & Mądry, A. (2025). MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. *arXiv*. https://arxiv.org/abs/2410.07095

Chen, Y., Tucker, M., Panickssery, N., Wang, T., Mosconi, F., Gopal, A., Denison, C., Petrini, L., Leike, J., Perez, E., & Sharma, M. (2025). Enhancing Model Safety through Pretraining Data Filtering. *Alignment Science Blog.* https://alignment.anthropic.com/2025/pretraining-data-filtering/

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

Ciardha, C. O., Buckley, J., & Portnoff, R. S. (2025). AI Generated Child Sexual Abuse Material—What's the Harm? *arXiv.* https://arxiv.org/pdf/2510.02978

CISA. (2025). CISA Software Bill of Materials (SBOM). *Cybersecurity and Infrastructure Security Agency.* https://www.cisa.gov/sbom

Cisco. (n.d.). Integrated AI Security and Safety Framework: AI Taxonomy Navigator. *Cisco.* https://learn-cloudsecurity.cisco.com/ai-security-framework

Clymer, J., Duan, I., Duan, C., Duan, Y., Heide, F., Lu, C., Mindermann, S., McGurk, C., Pan, X., Siddiqui, S., Wang, J., Yang, M., & Zhan, X. (2025). Bare Minimum Mitigations for Autonomous AI Development. *arXiv.* https://arxiv.org/pdf/2504.15416v2

Cohen, M. K., Kolt, N., Bengio, Y., Hadfield, G. K., & Russell, S. (2024). Regulating Advanced Artificial Agents. *arXiv.* https://www.science.org/doi/10.1126/science.adl0625

CSA. (2025a). Agentic AI Red Teaming Guide. *Cloud Security Alliance.* https://cloudsecurityalliance.org/artifacts/agentic-ai-red-teaming-guide

CSA. (2025b). Agentic AI Threat Modeling Framework: MAESTRO. *Cloud Security Alliance.* https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro

CSA Singapore, & FAR.AI. (2025). Securing Agentic AI: A Discussion Paper. *Cyber Security Agency of Singapore.* https://www.csa.gov.sg/resources/publications/securing-agentic-ai-a-discussion-paper/

CycloneDX. (n.d.a). AI Models and Model Cards. *CycloneDX.* https://cyclonedx.org/use-cases/ai-models-and-model-cards/

CycloneDX. (n.d.b). Getting Started. *CycloneDX.* https://cyclonedx.org/

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

Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. *arXiv.* https://arxiv.org/abs/2406.13352

Derczynski, L., Galinkin, E., Martin, J., Majumdar, S., & Inie, N. (2024). garak: A Framework for Security Probing Large Language Models. *arXiv.* https://arxiv.org/abs/2406.11036

Deshpande, C., & Joshi, R. (2025). AI Agents In Focus Technical and Policy Considerations. *Center for Democracy and Technology.* https://cdt.org/wp-content/uploads/2025/05/2025-05-14-AI-Gov-Lab-AI-Agents-In-Focus-brief-final.pdf

Díaz, S., Kern, C., & Olive, K. (2025). Google's Approach for Secure AI Agents: An Introduction. *Google.* https://storage.googleapis.com/gweb-research2023-media/pubtools/1018686.pdf

EC. (2025). The General-Purpose AI Code of Practice. *European Commission.* https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

Enkrypt AI. (2025). Agent Risk Taxonomy. *Enkrypt AI.* https://cdn.prod.website-files.com/6690a78074d86ca0ad978007/687f7fac66e8127aa565341d_Agent%20Risk%20taxonomy_enkryptai.pdf

EP. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). *European Parliament.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Feng, K. J. K., McDonald, D. W., & Zhang, A. X. (2025). Levels of Autonomy for AI Agents. *arXiv.* https://arxiv.org/abs/2506.12469

Gabriel, I., Manzini, A., Keeling, G., Hendricks L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., … Manyika, J. (2024). The Ethics of Advanced AI Assistants. *Google DeepMind.* https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H.III., Crawford, K. (2021). Datasheets for Datasets. *Association for Computing Machinery*. https://doi.org/10.1145/3458723

Gilbert, T. K., Lambert, N., Dean, S., Zick, T., & Snoswell, A. (2022). Reward Reports for Reinforcement Learning. *arXiv*. https://arxiv.org/abs/2204.10817

Google. (2025a). Google's Approach for Secure AI Agents: An Introduction. *Google.* https://storage.googleapis.com/gweb-research2023-media/pubtools/1018686.pdf

Google. (2025b). Announcing the Agent2Agent Protocol (A2A). *Google.* https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/

Google Cloud. (n.d.). Guide to Multi-Agent Systems (MAS). *Google Cloud.* https://cloud.google.com/discover/what-is-a-multi-agent-system

GOV.UK. (n.d.). AI Insights: Agentic AI (HTML). *GOV.UK.* https://www.gov.uk/government/publications/ai-insights/ai-insights-agentic-ai-html

Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). AI Control: Improving Safety Despite Intentional Subversion. *arXiv.* https://arxiv.org/abs/2312.06942

Gridach, M., Nanavati, J., Zine El Abidine, K., Mendes, L., & Mack, C. (2025). Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions. *arXiv*. https://arxiv.org/pdf/2503.08979

Griffin, C., Thomson, L., Shlegeris, B., & Abate, A. (2024). Games for AI Control: Models of Safety Evaluations of AI Deployment Protocols. *arXiv.* https://doi.org/10.48550/arXiv.2409.07985

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

Hadfield-Menell, D., Dragan, A., Abbeel, P., Russell, S.(2017). The Off-Switch Game. *arXiv.* https://arxiv.org/abs/1611.08219

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

Hashim, S. (2024). OpenAI's New Model Tried to Avoid Being Shut Down. *Transformer.* https://www.transformernews.ai/p/openais-new-model-tried-to-avoid?r=wl6sg&triedRedirect=true

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

Højmark, A., Pimpale, G., Panickssery, A., Hobbhahn, M., & Scheurer, J. (2024). Analyzing Probabilistic Methods for Evaluating Agent Capabilities. *arXiv.* https://arxiv.org/pdf/2409.16125

Huang, Y. (2024). Levels of AI Agents: from Rules to Large Language Models. *arXiv.* https://arxiv.org/abs/2405.06643

Huang, K., Narajala, V. S., Yeoh, J., Ross, J., Raskar, R., Harkati, Y., Huang, J., Habler, I., & Hughes, C. (2025). A Novel Zero-Trust Identity Framework for Agentic AI: Decentralized Authentication and Fine-Grained Access Control. *arXiv.* https://arxiv.org/abs/2505.19301

IBM. (2025). The Simplest Protocol for AI Agents to Work Together. *IBM.* https://research.ibm.com/blog/agent-communication-protocol-ai

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv.* https://arxiv.org/abs/1805.00899

ISO. (2018). ISO 31000:2018 — Risk Management — Guidelines. *International Organization for Standardization.* https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en

ISO. (2023). ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. *International Organization for Standardization.* https://www.iso.org/standard/42001

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

Kasirzadeh, A., & Gabriel, I. (2025). Characterizing AI Agents for Alignment and Governance. *arXiv.* https://arxiv.org/pdf/2504.21848

Kim, Y., Jeong, H., Park, C., Park, E., Zhang, H., Liu, X., Lee, H., McDuff, D., Ghassemi, M., Breazeal, C., Tulebaev, S., & Park, H. W. (2025). Tiered Agentic Oversight: A Hierarchical Multi-Agent System for Healthcare Safety. *arXiv.* https://arxiv.org/abs/2506.12482

Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L. H., Lin, T. R., Wijk, H., Burget, J., Ho, A., Barnes, E., & Christiano, P. (2024). Evaluating Language-Model Agents on Realistic Autonomous Tasks. *arXiv,* https://arxiv.org/abs/2312.11671

Kong, D., Lin, S., Xu, Z., Wang, Z., Li, M., Li, Y., Zhang, Y., Peng, H., Sha, Z., Li, Y., Lin, C., Wang, X., Liu, X., Zhang, N., Chen, C., Khan, M. K., & Han, M. (2025). A Survey of LLM-Driven AI Agent Communication: Protocols, Security Risks, and Defense Countermeasures. *arXiv.* https://arxiv.org/html/2506.19676v2

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., ... Mikulik, V. (2025). Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. *arXiv.* https://arxiv.org/abs/2507.11473

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

Lee, H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. *Association for Computing Machinery.* https://dl.acm.org/doi/10.1145/3706598.3713778

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

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., & Tang, J. (2025). AgentBench: Evaluating LLMs as Agents. *arXiv.* https://arxiv.org/abs/2308.03688

Longpre, S., Kapoor, S., Klyman, K., Ramaswami, A., Bommasani, R., Blili-Hamelin, B., Huang, Y., Skowron, A., Yong, Z. X., Kotha, S., Zeng, Y., Shi, W., Yang, X., Southen, R., Robey, A., Chao, P., Yang, D., Jia, R., Kang, D., Pentland, S., Narayanan, A., Liang, P., & Henderson, P. (2024). A Safe Harbor for AI Evaluation and Red Teaming. *arXiv.* https://arxiv.org/abs/2403.04893

Longpre, S., Mahari, R., Muennighoff, N., Chen, A., Perisetla, K., Brannon, W., Kabbara, J., Villa, L., & Hooker, S.(2023). The Data Provenance Project. *Proceedings of the 40th International Conference on Machine Learning.* https://blog.genlaw.org/CameraReady/20.pdf

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). General-Purpose AI Risk-Management Standards Profile, Version 1.2. *UC Berkeley Center for Long-Term Cybersecurity*. https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile-v1.2/

Mansilla, L., Echeveste, R., Gonzalez, C., Milone, D. H., & Ferrante, E. (2025). BM-CL: Bias Mitigation through the lens of Continual Learning. *arXiv.* https://arxiv.org/abs/2509.01730

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

METR. (2024). The Rogue Replication Threat Model. *METR.* https://metr.org/blog/2024-11-12-rogue-replication-threat-model/

METR. (2025a). Common Elements of Frontier AI Safety Policies. *METR.* https://metr.org/common-elements.pdf

METR. (2025b). Forecasting the Impacts of AI R&D Acceleration: Results of a Pilot Study. *METR*. https://metr.org/blog/2025-08-20-forecasting-impacts-of-ai-acceleration/

MIT. (2025a). How is AI Harming Us? *MIT AI Risk Initiative*. https://airisk.mit.edu/ai-incident-tracker

MIT. (2025b) What are the Risks from Artificial Intelligence?*MIT AI Risk Initiative*. https://airisk.mit.edu/

Mitchell, M., Ghosh, A., Luccioni, A. S., & Pistilli, G. (2025). Fully Autonomous AI Agents Should Not be Developed. *arXiv*. https://arxiv.org/abs/2502.02649

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *In Proceedings of the Conference on Fairness, Accountability, and Transparency 2019*, pp. 220–229. https://dl.acm.org/doi/10.1145/3287560.3287596

MITRE. (n.d.a). MITRE ATLAS AI Incidents. *MITRE*. https://ai-incidents.mitre.org/

MITRE. (n.d.b). AI Risk Database. *MITRE*. https://ai-risk.mitre.org/

Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H.S., Hammond, L., & Schroeder de Witt, C. (2025). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. *arXiv*. https://arxiv.org/abs/2402.07510

Mukherjee, A., & Chang, H. H. (2025). Agentic AI: Autonomy, Accountability, and the Algorithmic Society. *arXiv*. https://arxiv.org/pdf/2502.00289

Murugesan, S. (2025). The Rise of Agentic AI: Implications, Concerns, and the Path Forward. *IEEE*. https://ieeexplore.ieee.org/abstract/document/10962241/authors#authors

Narajala, V. S., & Narayan, O. (2025). Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents. *arXiv*. https://doi.org/10.48550/arXiv.2504.19956

Newman, J. (2023). A Taxonomy of Trustworthiness for Artificial Intelligence. *UC Berkeley Center for Long-Term Cybersecurity*. https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/

NIST. (n.d.a). AI Risks and Trustworthiness. *National Institute of Standards and Technology*. https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/

NIST. (n.d.b). Trustworthy and Responsible AI. *National Institute of Standards and Technology*. https://www.nist.gov/trustworthy-and-responsible-ai

NIST. (2018). Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy. *National Institute of Standards and Technology*. https://csrc.nist.gov/pubs/sp/800/37/r2/final

NIST. (2023a). AI Risk Management Framework (AI RMF 1.0). AI 100-1. *National Institute of Standards and Technology*. https://doi.org/10.6028/NIST.AI.100-1

NIST. (2023b). AI Risk-Management Framework Playbook (version released January 2023). *National Institute of Standards and Technology*. https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook

NIST. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. *National Institute of Standards and Technology*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

NIST. (2025a). Lessons Learned from the Consortium: Tool Use in Agent Systems. *National Institute of Standards and Technology*. https://www.nist.gov/news-events/news/2025/08/lessons-learned-consortium-tool-use-agent-systems

NIST. (2025b). Technical Blog: Strengthening AI Agent Hijacking Evaluations. *National Institute of Standards and Technology*. https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations

NIST. (2025c). Cheating On AI Agent Evaluations. *National Institute of Standards and Technology.* https://www.nist.gov/blogs/caisi-research-blog/cheating-ai-agent-evaluations

OECD.AI. (n.d.). AIM: AI Incidents and Hazards Monitor. *OECD.AI.* https://oecd.ai/en/incidents?search_terms=%5B%5D&and_condition=false&from_date=2014-01-01&to_date=2025-09-28&properties_config=%7B%22principles%22:%5B%5D,%22industries%22:%5B%5D,%22harm_types%22:%5B%5D,%22harm_levels%22:%5B%5D,%22harmed_entities%22:%5B%5D,%22business_functions%22:%5B%5D,%22ai_tasks%22:%5B%5D,%22autonomy_levels%22:%5B%5D,%22languages%22:%5B%5D%7D&order_by=date&num_results=20

OpenAI. (2024). OpenAI o1 System Card. *OpenAI.* https://cdn.openai.com/o1-system-card-20240917.pdf

OpenAI. (2025) Agent Bio Bug Bounty. OpenAI. https://openai.com/bio-bug-bounty/

ORF. (2024). Issue Brief. Issue No. 768 December 2024. *Observer Research Foundation.* https://www.orfonline.org/public/uploads/posts/pdf/20241227111028.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

OWASP. (n.d.a). OWASP GenAI Security Project. *OWASP.* https://genai.owasp.org/

OWASP. (n.d.b). OWASP AI Vulnerability Scoring System (AIVSS). *OWASP.* https://aivss.owasp.org/

OWASP. (n.d.c). OWASP AIBOM. *OWASP.* https://owaspaibom.org/

OWASP. (2025a). Agentic AI – Threats and Mitigations. *OWASP.* https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

OWASP. (2025b). OWASP Top 10 for Agentic Applications for 2026. *OWASP.* https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

Palisade Research. (n.d.). AI Misalignment Bounty. *Palisade Research.* https://bounty.palisaderesearch.org/

Pan, X., Dai, J., Fan, Y., Luo, M., Li, C., & Yang, M. (2025). Large language Model-Powered AI Systems Achieve Self-Replication With no Human Intervention. *arXiv.* https://arxiv.org/pdf/2503.17378

Parikh, S., & Surapaneni, R. (2025). Powering AI commerce with the new Agent Payments Protocol (AP2). *Google Cloud.* https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol

Patil, S. G., Zhang, T., Fang, V., C., N., Huang R., Hao, A., Casado, M., Gonzalez, J. E., Popa, R. A., & Stoica I. (2024). GoEX: Perspectives and Designs Towards a Runtime for Autonomous LLM Applications. *arXiv.* https://arxiv.org/abs/2404.06921

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

Phan, T. (2023). Emergence and Resilience in Multi-Agent Reinforcement Learning. *Ludwig Maximilians University.* https://edoc.ub.uni-muenchen.de/31981/

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

Renieris, E. M., Kiron, D., Mills, S., & Kleppe, A. (2025). Agentic AI at Scale: Redefining Management for a Superhuman Workforce. *MIT Sloan.* https://sloanreview.mit.edu/article/agentic-ai-at-scale-redefining-management-for-a-superhuman-workforce/

Robeyns, M. (2025). A Self-Improving Coding Agent. *arXiv.* https://arxiv.org/html/2504.15228v2

Roucher, A., Noyan, M., & Wolf, T. (2024). Introducing Smolagents, a Simple Library to Build Agents. *Hugging Face.* https://huggingface.co/blog/smolagents

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

Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2025). Shutdown Resistance in Reasoning Models. *Palisade Research.* https://palisaderesearch.org/blog/shutdown-resistance

Schmitt, M., & Flechais, I. (2024). Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing. *Springer Nature.* https://link.springer.com/article/10.1007/s10462-024-10973-2

Schoen, B., Nitishinskaya, E., Balesni, M., Højmark, A., Hofstätter, F., Scheurer, J., Meinke, A., Wolfe, J., van der Weij, T., Lloyd, A., Goldowsky-Dill, N., Fan, A., Matveiakin, A., Shah, R., Williams, M., Glaese, A., Barak, B., Zaremba, W., & Hobbhahn, M. (2025). Stress Testing Deliberative Alignment for Anti-Scheming Training. *arXiv.* https://www.arxiv.org/abs/2509.15541

Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., Goldenberg, A., Kyrychenko, Y., Leyton-Brown, K., Lutz, N., Marcus, G., Menczer, F., Pennycook, G., Rand, D. G., Ressa, M., Schweitzer, F., Summerfield, C., Tang, A., Van Bavel, J.J., van der Linden, S., Song, D., & Kunst, J. R. (2025). How Malicious AI Swarms Can Threaten Democracy: The Fusion of Agentic AI and LLMs Marks a New Frontier in Information Warfare. *arXiv.* https://arxiv.org/abs/2506.06299

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, Sevien, Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2025). The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. *arXiv.* https://doi.org/10.48550/arXiv.2406.06608

Seifried, K. (2025). Securing the Agentic AI Control Plane: Announcing the MCP Security Resource Center. *Cloud Security Alliance.* https://cloudsecurityalliance.org/blog/2025/08/20/securing-the-agentic-ai-control-plane-announcing-the-mcp-security-resource-center

Shao, A. (2025). Beyond Misinformation: A Conceptual Framework for Studying AI Hallucinations in (Science) Communication. *arXiv.* https://arxiv.org/html/2504.13777v1

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

Shapiro, J. (2025). Letter on AI and Employees. *Office of the Governor, Harrisburg PA.* https://www.pa.gov/content/dam/copapwp-pagov/en/governor/documents/2025.3.21%20gov%20shapiro%20letter%20on%20ai%20and%20employees.pdf

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N.,

Miller, J., Michaud, E. J., Casper, S., Tegmark, M., Saunders, W., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., & McGrath, T. (2025). Open Problems in Mechanistic Interpretability. *arXiv.* https://arxiv.org/abs/2501.16496

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

Sharp, M., Bilgin, O., Gabriel, I., & Hammond, L. (2025). Agentic Inequality. *arXiv.* https://www.arxiv.org/abs/2510.16853

Sheh, R. K., & Geappen, K. (2025). Identifying the Supply Chain of AI for Trustworthiness and Risk Management in Critical Applications. *arXiv.* https://arxiv.org/abs/2511.15763

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

SLSA. (n.d.). Safeguarding Artifact Integrity Across any Software Supply Chain. *SLSA.* https://slsa.dev/

Sonni, A. F. (2025). AI-Based Disinformation and Hate Speech Amplification: Analysis of Indonesia's Digital Media Ecosystem. *Frontiers.* https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2025.1603534/full

SPDX. (n.d.a). SPDX AI. *SPDX.* https://spdx.dev/learn/areas-of-interest/ai/

SPDX. (n.d.b). The System Package Data Exchange (SPDX®). *SPDX.* https://spdx.dev/

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/

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content. *Nature.* https://www.nature.com/articles/s41599-024-03811-x

TAIBOM. (n.d.). Trustable AI Bill of Materials. *NquiringMinds.* https://taibom.org/

Teo, S. A. (2024). How to Think About Freedom of Thought (and Opinion) in the Age of AI. *ScienceDirect.* https://www.sciencedirect.com/science/article/abs/pii/S0267364924000360?casa_token=w-xNRPkAzn8AAAAA:5gp3ByzrRoBuR5yEkXMtHBgWkiu50gpni-SVY525f7Y-PIUdgsp79l9wkNTEQd12AhT1kPkqPkw

Terekhov, M., Panfilov, A., Dzenhaliou, D., Gulcehre, C., Andriushchenko, M., Prabhu, A., & Geiping, J. (2025a). Adaptive Attacks on Trusted Monitors Subvert AI Control Protocols. *arXiv.* https://doi.org/10.48550/arXiv.2510.09462

Terekhov, M., Liu, Z. N. D., Gulcehre, C., & Albanie, S. (2025b). Control Tax: The Price of Keeping AI in Check. *arXiv*. https://doi.org/10.48550/arXiv.2506.05296

TFS. (2025). Global Red Lines for AI: A Three-Part Series. *The Future Society.* https://thefuturesociety.org/airedlines

Toner, H., Bansemer, J., Crichton, K., Burtell, M., Woodside, T., Lior, A., Lohn, A., Acharya, A., Cibralic, B., Painter, C., O'Keefe, C., Gabriel, I., Fisher, K., Ramakrishnan, K., Jackson, K., Kolt, N., Crootof, R., & Chatterjee, S. (2024). Through the Chat Window and Into the Real World: Preparing for AI Agents. *Center for Security and Emerging Technology.* https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents/

trufo.ai. (2024). Solutions to the Content Authenticity Problem: A Technical Overview of Creating, Managing, and Analyzing Digital Provenance (2024). *trufo.ai.* https://trufo.ai/solutions-to-the-content-authenticity-problem.pdf

UK AISI. (2024). Early Lessons From Evaluating Frontier AI Systems. *UK AI Security Institute.* https://www.aisi.gov.uk/blog/early-lessons-from-evaluating-frontier-ai-systems

UK AISI. (2025a). RepliBench: Measuring Autonomous Replication Capabilities in AI Systems. *UK AI Security Institute.* https://www.aisi.gov.uk/blog/replibench-measuring-autonomous-replication-capabilities-in-ai-systems

UK AISI. (2025b). The Inspect Sandboxing Toolkit: Scalable and Secure AI Agent Evaluations. *UK AI Security Institute*. https://www.aisi.gov.uk/blog/the-inspect-sandboxing-toolkit-scalable-and-secure-ai-agent-evaluations

UK AISI. (2025c). Navigating the Uncharted: Building Societal Resilience to Frontier AI. *UK AI Security Institute*. https://www.aisi.gov.uk/blog/navigating-the-uncharted-building-societal-resilience-to-frontier-ai

WEF. (2024). Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents. *World Economic Forum*. https://reports.weforum.org/docs/WEF_Navigating_the_AI_Frontier_2024.pdf

WEF. (2025a). AI red lines: The Opportunities and Challenges of Setting Limits. *World Economic Forum*. https://www.weforum.org/stories/2025/03/ai-red-lines-uses-behaviours/

WEF. (2025b). AI Agents in Action: Foundations for Evaluation and Governance. *World Economic Forum*. https://www.weforum.org/publications/ai-agents-in-action-foundations-for-evaluation-and-governance/

van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2025). AI Sandbagging: Language Models can Strategically Underperform on Evaluations. *arXiv*. https://arxiv.org/abs/2406.07358

Wen, X., Lou, J., Lu, X., Yang, J., Liu, Y., Lu, Y., Zhang, D., & Yu, X. (2025). Scalable Oversight for Superhuman AI via Recursive Self-Critiquing. *arXiv*. https://arxiv.org/abs/2502.04675

Wisakanto, A. K., Rogero, J., Casheekar, A. M., & Mallah, R. (2025). Adapting Probabilistic Risk Assessment for AI. *arXiv*. https://doi.org/10.48550/arXiv.2504.18536

Wu, H. (2024). AI Whistleblowers. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4790511

Yair, O., Nassi, B., & Cohen, S. (2025). Invitation Is All You Need: Invoking Gemini for Workspace Agents with a Simple Google Calendar Invite. *SafeBreach*. https://www.safebreach.com/blog/invitation-is-all-you-need-hacking-gemini/

Yee, E. (2025). AI Red-Teaming Design: Threat Models and Tools. *Center for Security and Emerging Technology*. https://cset.georgetown.edu/article/ai-red-teaming-design-threat-models-and-tools/

Yoran, O., Amouyal, S. J., Malaviya, C., Bogin, B., Press, O., & Berant, J. (2024). AssistantBench: Can Web Agents Solve Realistic and Time-Consuming Tasks? *arXiv*. https://arxiv.org/abs/2407.15711

Young, S. (2025). Understanding and Mitigating Security Risks in MCP Implementations. *Microsoft Security Community Blog*. https://techcommunity.microsoft.com/blog/microsoft-security-blog/understanding-and-mitigating-security-risks-in-mcp-implementations/4404667

Zhan, Q., Liang, Z., Ying, Z., & Kang, D. (2024). InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents. *arXiv*. https://arxiv.org/abs/2403.02691

Zhang, Z., Cui, S., Lu, Y., Zhou, J., Yang, J., Wang, H., & Huang, M. (2025). Agent-SafetyBench: Evaluating the Safety of LLM Agents. *arXiv*. https://arxiv.org/abs/2412.14470


CLTC
Center for Long-Term
Cybersecurity
UC Berkeley