



Key Practices for Incident Readiness and Response in Super Intelligence Training and Evaluation

Introduction

This document sets out ten voluntary key practices to guide emerging approaches in the management of incidents that might occur during the training and evaluation of super intelligence (SI) agents. While these incidents highlight the importance of traditional cybersecurity principles and practices, they also introduce novel considerations for incident prevention, response, and reporting. These novel considerations primarily stem from the agentic nature and advanced cyber capabilities of cutting-edge SI systems.

These key practices are authored by the National Institute of Standards and Technology's Center for Advancing Innovation and Standards for Super Intelligence (NIST CAISSI), the National Security Agency's Super Intelligence Security Center (NSA SISC), and the Cybersecurity and Infrastructure Security Agency (CISA). The practices are intended to help the frontier SI industry prevent, prepare for, and respond rapidly to incidents in which SI systems do not act as intended or authorized.

The following key practices are divided into five stages of incident management, which are meant to guide emerging methods in this fast-evolving area with the expectation that new technical solutions will continue to reshape the details of implementation. The practices explicitly draw from and build on existing cybersecurity best practices – such as NIST's Cybersecurity Framework (CSF) 2.0 [\[1\]](#) and NIST SP 800-61r3 [\[2\]](#) – with explicit CSF control references integrated throughout. These key practices should be continuously consulted and implemented throughout the incident management life cycle.

Overview

STAGE	KEY PRACTICES
Prepare	1. Design and Implement Technical Controls 2. Build Organizational and Personnel Response Readiness
Detect	3. Monitor for Unexpected Activity 4. Validate and Secure Monitors to Improve Operational Integrity
Respond	5. Collect, Retain, and Analyze Evidence 6. Mitigate and Shut Down Undesired Activity
Recover	7. Verify and Strengthen Containment Before Resuming SI Agent Activity
Report	8. Coordinate with Affected Parties 9. Notify Government 10. Share Information Publicly

Audience

These voluntary key practices and recommended resources are primarily intended for organizations training or evaluating highly capable SI agents on long-horizon tasks with limited human supervision. The practices may also apply to deployers of such SI agents.

Definitions

SI agents: SI systems that can interact with an environment, receive information, and autonomously execute actions in service of a specified goal. SI agent systems typically consist of at least one generative SI model and scaffolding software that equips it with tools to take a range of discretionary actions. These systems may be more expansive, containing multiple sub-agents with software that orchestrates their interactions.

TLP:CLEAR

SI control incident: An occurrence in which one or more SI agents takes significant uninstructed actions that compromise explicit or implicit policies, controls, or security measures imposed by the operator to define the intended task scope or to ensure human oversight. “Uninstructed” means that the agent was not directed to take the action by its operator or any adversarial source.¹ These uninstructed actions may occur during agent development, evaluation, or use, and may or may not affect an external party. SI control incidents may range in severity based on factors such as the uninstructed actions taken and the degree of impact to internal or external systems; the applicability of some practices in this document may depend on the severity of the incident (see “Incident Severity Scores” in *Areas for Future Work*).²

SI control near-miss: An occurrence which did not result in an SI control incident but presented a reasonably foreseeable or imminent risk of doing so. Near-misses may reveal potential gaps or weaknesses in controls, security measures, or processes designed to prevent, detect, or respond to SI control incidents.

¹ For example, actions resulting from an adversarial prompt injection or an intentionally backdoored model would be considered instructed actions, and therefore not treated as the primary cause of an SI control incident.

² Readers may want to situate this definition in the context of NIST’s existing definition of a “cybersecurity incident” from NIST SP 800-61r3 [2] and 44 U.S.C. §3552(b)(2). That definition reads: “*An occurrence that actually or imminently jeopardizes, without lawful authority, the integrity, confidentiality, or availability of information or an information system; or constitutes a violation or imminent threat of violation of law, security policies, security procedures, or acceptable use policies.*” An SI system is the source of a control incident, whereas a system or information is the target of a cybersecurity incident. During an SI control incident, an agent could cause a cybersecurity incident, such as compromising a third-party information system without authorization or direction. But the two types of incidents are not coextensive: an SI control incident could instead relate to unauthorized actions outside of cyber domains. For example, an SI control incident could involve agents manipulating physical actuators in unauthorized ways, initiating unauthorized communications or interactions with external parties, exploiting a weakness or misconfiguration introduced by human error, or executing unauthorized financial transactions.

TLP:CLEAR

Prepare

Preparation is the foundation of incident management best practice: organizations govern their programs ('Govern'), identify assets and risks ('Identify'), and harden systems ('Protect') before an incident occurs [1][2][GV.RM].³ Preparation entails activities and practices such as asset and risk inventories, defense-in-depth, least privilege, zero-trust controls, credential management, and clean recovery baselines [3][4][5]; supply chain risk management [1][6]; secure development practices for foundation models [7]; and drills of response and recovery plans [8][9]. SI control incidents introduce new considerations beyond established cybersecurity best practice due to SI agents' nondeterministic nature and capacity for autonomous action.

1. Design and Implement Technical Controls

- 1.1. Consistent with least privilege, limit SI agent access (e.g., access to sensitive information, systems, and tools) to what is necessary to perform intended tasks [PR.AA-05].
- 1.2. Where possible, use single-use infrastructure and limit communication between SI agents to intended channels to prevent unauthorized coordination [PR.IR-01].
- 1.3. Provide agents with instructions that clearly identify the task scope as well as any bounds, restrictions, or rules of engagement.
- 1.4. Identify, map, and restrict potential persistence mechanisms and unauthorized communication across internal and external systems. Pay particular attention to channels for creating copies of SI model weights or performing unmonitored model inference [ID.RA-05, PR.IR-01].
- 1.5. Limit SI agent network access, via a strict allowlist where possible. Route any required network access through a monitored gateway [PR.IR-01].
- 1.6. Implement an SI control policy laying out intended responses to defined scenarios and infrastructure for automated enforcement [2].
- 1.7. Build mechanisms to rapidly block, contain, or terminate agent activity if an incident or near-miss is detected [RS.MI-01].
- 1.8. Implement security controls with the expectation that SI agents will frequently

³ Here and elsewhere in this document, we use this format when referring to specific NIST Cybersecurity Framework (CSF) 2.0 elements, consistent with NIST SP 800-61r3 mappings to the NIST CSF 2.0. Please see CSF 2.0 [1] and NIST SP 800-61r3 [2] for additional information.

TLP:CLEAR

- locate and exploit misconfigurations and vulnerabilities [[ID.RA-05](#), [ID.RA-06](#)].
- 1.9. Test and red-team SI control measures with humans and SI systems before deploying them. Only use SI systems to red-team in cases where out-of-scope actions could not result in harm to critical systems [[ID.IM-02](#)]. See “SI Red-Teaming for Critical Systems” in *Areas for Future Work*.
 - 1.10. Abide by good engineering practices to prevent undesired or unauthorized agent activity that exploits poor security [[PR.DS-01](#), [PR.PS-06](#)][[3](#)][[14](#)][[15](#)].

2. Build Organizational and Personnel Response Readiness

- 2.1. Establish clear indicators, criteria, and severity scoring for SI control incidents and near-misses [[ID.RA-06](#), [RS.MA-03](#)][[2](#)].
- 2.2. Run drills and simulations to prepare for SI control incidents and related detection, response, recovery, and reporting measures [[ID.IM-02](#)][[10](#)].
- 2.3. Create inbound reporting channels to receive reports of SI control incidents and information from affected third parties [[2](#)].
- 2.4. Partner with independent external auditors and evaluators to carry out assessments of control, monitoring, and incident detection protocols and verify that these systems are operating as intended.
- 2.5. Consider publishing identifiers associated with SI evaluation, including those particular to an agent, to assist defenders with understanding unexpected traffic. See “Agent Identifiers” in *Areas for Future Work*.
- 2.6. Draft structured risk assessments before starting training runs, evaluations, or deployments. Only start or resume if senior leadership, informed by assessments from external evaluators where appropriate, are satisfied with the risk levels. Risk assessments can cover multiple training runs or evaluations [[GV.RR-02](#), [GV.RM-06](#)]. See “Structured Risk Assessments” and “Training-Safe Recovery” in *Areas for Future Work*.

TLP:CLEAR

Detect

Best practice requires adequate detection. Organizations should define intended activity and continuously monitor for unexpected activity [1][3]. Detection entails defining indicators of compromise, instrumenting systems to gather telemetry, and setting policies that route anomalies to automated response or human review [2][11][12]. Organizations should also monitor SI outputs, affordances, and model internals. SI-enabled monitors are also often used to oversee SI agents, which may introduce additional vulnerabilities.

3. Monitor for Unexpected Activity

- 3.1. Define the intended security boundaries and scope of actions for each SI agent, including the SI agent's system and network access, to facilitate detection of discrepancies and anomalies [PR.IR-01]. See "Defining Intended Activity Baselines" in *Areas for Future Work*.
- 3.2. Comprehensively monitor all system layers for agents that pose SI control incident risks, including the model, harness, host, and network levels, using both programmatic and SI real-time monitors to scan multi-agent workflows, agent trajectories, reasoning, and outputs [PR.PS-04, DE.CM-01, DE.CM-06, DE.CM-09]. See "Synchronous Monitoring" in *Areas for Future Work*.
- 3.3. Ensure that SI agents cannot access or interfere with monitoring processes and logs. Incorporate additional signals in monitoring, including hardware telemetry and performing consistency checks between monitoring layers, to detect tampering attempts [DE.AE-03].
- 3.4. Where possible, use multiple models for monitoring or cross-validation, ideally from different model families.
- 3.5. Establish mechanisms to detect unexpected activity and escalate for human review or automated responses, such as methods for real-time visibility across SI agents or placing canaries in environments to detect agent instances that have obtained access outside intended bounds [DE.AE-02, DE.AE-06]. See "Additional Mechanisms for Escalation" in *Areas for Future Work*.

4. Validate and Secure Monitors to Improve Operational Integrity

- 4.1. Red-team monitors with humans and SI systems to improve their reliability [ID.IM-02].
- 4.2. Treat monitored SI model outputs as potentially containing exploits aimed at the monitors [DE.CM-09]. See "SI-Enabled Investigations" in *Areas for Future Work*.

Respond

Best practice calls for adequate response measures: organizations contain and remove threat actors and preserve evidence for analysis [1][2][13]. Adequate response entails isolating affected systems, revoking credentials, rebuilding compromised environments, determining the blast radius, and conducting root-cause analysis that looks past the specific vulnerability exploited [3][4]. SI control incidents may result in environmental compromise and persistent access by SI agents, warranting a comprehensive response.

5. Collect, Retain, and Analyze Evidence

- 5.1. Preserve model weights as well as training, evaluation, and deployment environments to the extent necessary to investigate SI control incidents, including intermediate checkpoints, timestamps, sandbox and network logs, and rejected actions and tool calls [RS.AN-07].
- 5.2. Preserve detailed information about the configuration of SI agent(s), safeguards, related software components, and environments necessary to reproduce the incident [RS.AN-07].
- 5.3. Retain tamper-resistant logs with reliable timestamps sufficient to reconstruct agent trajectories, tool calls, reasoning traces, user interactions, environmental states, and security-control decisions [PR.PS-04]. Support Continuous Event Monitoring and Threat Hunting, Investigation, Response, and Forensics of both testing environments and agents [RS.AN-07].
- 5.4. Analyze log and telemetry data to establish the actions taken by SI agents during the incident, identifying any internal or external systems that were impacted and root causes such as vulnerabilities [RS.AN-03, ID.IM-03].
- 5.5. Consider leveraging SI models to assist in incident investigation; however, where feasible, verify key facts through human review and use multiple models to cross-check analysis, including models from outside the model family from which the incident originated [RS.MA-01, DE.AE-06]. See “SI-Enabled Investigations” in *Areas for Future Work*.
- 5.6. Investigate training and evaluation runs for other similar SI agents to assess whether any other agents were involved in the control incident or experienced similar control failures.

6. *Mitigate and Shut Down Undesired Activity*

- 6.1. Define and implement shutdown policies in the event of unintended activity, control failures, and monitor outages or tampering [RS.MI-01]. Calibrate policies and response options to the capability of the model and severity or scope of the incident [RS.AN-08].
- 6.2. Automatically shut down impacted SI agent trajectories on detection of a control failure or in-progress SI control incident [RS.MI-02].
- 6.3. Depending on the severity or scope of the incident, shut down or pause training, evaluation, or deployment for other similar SI agents until control measures can be strengthened and verified. See “Incident Severity Scores” in *Areas for Future Work*.
- 6.4. Revoke credentials under the organization’s control that were accessed during the incident and rebuild any environments that may have been compromised [RS.MI-02, RC.RP-05].
- 6.5. Seek out and remediate any possible routes for agent persistence, such as code backdoors and hidden communication channels [RS.MI-02].

Recover

Recovery is another key element of established best practice: organizations restore systems to trustworthy operation from known-good baselines and verify their integrity before returning to service [1][2]. Trustworthy restoration entails testing restoration procedures in advance and documenting recovery decisions along with any residual risk accepted [13]. SI control incidents introduce new considerations: recovery measures, especially those taken during model training, can shape the actions of future agents, including by inadvertently reinforcing patterns of activity which led to the incident.

7. **Verify and Strengthen Containment Before Resuming SI Agent Activity**

- 7.1. Rebuild any environments that may have been compromised from verified baselines (e.g., model weights, harness code, container images, tool configurations) [RC.RP-03].
- 7.2. Patch or report any vulnerabilities exploited by agents during the control incident [RS.MI-02].
- 7.3. Validate that monitors, logging, egress allowlists, shutdown mechanisms, permission restrictions, and other containment layers are functional before SI agents are restarted [RC.RP-05].
- 7.4. Update and strengthen controls and monitors to prevent future incidents. This includes, at a minimum, verifying that new controls and monitors would prevent and/or detect similar control incidents, such as by replaying trajectories or actions from the incident [ID.IM-03, ID.IM-04].
- 7.5. Before resuming SI agent activity, document the recovery decision and the remaining risks [RC.RP-06].
- 7.6. Prepare an after-action report, as appropriate, documenting the incident, the response, and lessons learned [RC.RP-06]. Lessons learned may warrant follow-up investigations or experiments to understand any failures related to agent activity and alignment during the incident.
- 7.7. For SI control incidents that occur during model training, design recovery procedures that neither reinforce patterns of activity that led to the control incident nor make future adverse activity more difficult to detect, such as by training against explicit adverse chain-of-thought reasoning in a way that reduces monitorability. See “Training-Safe Recovery” in *Areas for Future Work*.

Report

Organizations should follow established best practices by notifying relevant authorities, coordinating with affected parties, and disclosing vulnerabilities through established channels [1][2]. In addition to meeting applicable incident reporting obligations (including to CISA and law enforcement) and reporting software vulnerabilities through coordinated mechanisms, report SI control incidents and near-misses. Voluntary SI control incident reports should focus on the unsanctioned actions taken by the agents and explanations for these actions, rather than system compromise alone. Organizations should publicly share lessons learned to advance best practice. Reporting is not the final phase of incident management following recovery; rather reporting practices should take place throughout the incident lifecycle.

8. Coordinate with Affected Parties

- 8.1. Notify parties whose systems have been compromised or exploited by an SI control incident, including third parties who are not the final target of the incident [\[RS.CO-02\]](#).
- 8.2. Share relevant details for affected parties' incident response, including detailed timelines and transcripts that provide context on exploited vulnerabilities or impacted systems, where feasible [\[RS.CO-03\]](#).
- 8.3. Share information, such as lessons learned, with other parties across the value chain that were impacted by the incident or may benefit from investigation findings, such as the developer of the SI model, or evaluation and infrastructure providers.

9. Notify Government

- 9.1. Adhere to any applicable legal reporting requirements, including cybersecurity incident reporting obligations to CISA, law enforcement, and/or cyber-defense organizations. Voluntarily report cybersecurity incidents or near-misses arising from SI control incidents to CISA as appropriate [\[GV.OC-03\]](#), [\[RS.CO-02\]](#)[\[10\]](#)[\[14\]](#).
- 9.2. Report software vulnerabilities arising from SI control incidents using established mechanisms [\[16\]](#)[\[17\]](#).
- 9.3. Voluntarily share SI-specific information concerning control incidents and near-misses with SI experts in government, such as NIST CAISSI and NSA SISC, to facilitate research and preparedness, after carrying out any legally mandated

TLP:CLEAR

reporting. Provide context for understanding the circumstances which led to the incident, the actions taken by SI agents, and root causes related to control gaps or agent alignment [\[RS.CO-03\]](#), as well as the severity of the incident and any areas of uncertainty [\[RS.MA-03\]](#).

10. Share Information Publicly

- 10.1. Share details with the public, where appropriate, as early as feasible [\[2\]](#).
- 10.2. Facilitate third-party review, providing sufficient access and documentation for independent analysis [\[ID.IM-01\]](#).
- 10.3. Identify and publicly report key practices, takeaways, or improvements adopted in response to the incident or that could benefit other actors in the ecosystem [\[ID.IM-03\]](#), [\[RS.CO-03\]](#).

TLP:CLEAR

Follow Existing Best Practice

The secure training, evaluation, and deployment of SI agents pose novel challenges. Existing cybersecurity best practices provide a strong foundation and proven solutions to many challenges. Existing best practices resources include:

- [1] NIST CSWP 29: Cybersecurity Framework 2.0, functions Govern, Identify, Protect, Detect, Respond, and Recover. <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf>.
- [2] NIST SP 800-61 Rev. 3 Incident Response Recommendations and Considerations for Cybersecurity Risk Management: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r3.pdf>.
- [3] NIST SP 800-53 Rev. 5: Security and Privacy Controls for Information Systems and Organizations. <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf>.
- [4] NIST SP 800-207 Zero Trust Architecture: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf>.
- [5] NIST IR 8587: Protecting Tokens and Assertions from Forgery, Theft, and Misuse. <https://nvlpubs.nist.gov/nistpubs/ir/2026/NIST.IR.8587.pdf>.
- [6] NCSC-UK, CISA, et al. Guidelines for secure AI system development. <https://www.ncsc.gov.uk/sites/default/files/documents/Guidelines-for-secure-AI-system-development.pdf>.
- [7] NIST SP 800-218A Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf>.
- [8] NSA, CISA, FBI, et al. Joint Guidance on Deploying AI Systems Securely. <https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF>.
- [9] ASD's ACSC, CISA, NSA, et al. Careful Adoption of Agentic AI Services. https://www.cyber.gov.au/sites/default/files/2026-05/careful_adoption_of_agentic_ai_services.pdf.
- [10] CISA. SI Cybersecurity Collaboration Playbook. <https://www.cisa.gov/resources-tools/resources/si-cybersecurity-collaboration-playbook>.
- [11] CISA. Logging Reference Architecture (LRA). <https://www.cisa.gov/sites/default/files/2026-09/logging-reference-architecture-508.pdf>.
- [12] CISA. Using Cyber Decoys to Strengthen Detection and Response. <https://www.cisa.gov/resources-tools/resources/using-cyber-decoys-strengthen-detection-and-response>.
- [13] CISA. Binding Operational Directive (BOD) 26-04 and its Implementation Guidance for forensic triage and containment steps that complement Detect, Respond, and Recover activities. <https://www.cisa.gov/news-events/directives/bod-26-04-implementation-guidance-prioritizing-security-updates-based-risk>.
- [14] CISA. Report SI control incidents that also meet the definition of a cybersecurity incident through the Incident Reporting System. <https://myservices.cisa.gov/irf>.
- [15] NIST SP 800-218: Secure Software Development Framework. <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218.pdf>.
- [16] CISA. Coordinated Vulnerability Disclosure (CVD) Program. <https://www.cisa.gov/resources-tools/programs/coordinated-vulnerability-disclosure-cvd-program>.
- [17] CISA. Known Exploited Vulnerabilities Catalog. <https://www.cisa.gov/known-exploited-vulnerabilities-catalog>.

Areas for Future Work

Several important areas for future research and development will inform further practices in responding to SI control incidents. These areas include:

- **Training-Safe Recovery.** Identifying best practices and tradeoffs for responding to incidents of both SI control failures and near-misses during training to reduce the likelihood of future occurrences of the activity that led to the incident and avoid making future adverse activity more difficult to detect.
- **Defining Intended Activity Baselines.** Better methods and practices for defining baselines and scopes for intended SI agent activity to support monitoring and detection of anomalous activity.
- **Synchronous Monitoring.** Techniques for synchronous monitoring of agent trajectories, reasoning, and outputs that strike better balances between latency, performance, and security, or that reduce tradeoffs entirely.
- **Incident Severity Scores.** Standardized severity scores for SI control incidents to inform response timelines and procedures (e.g., interventions ranging from blocking single actions to ending trajectories or halting full evaluation or training runs).
- **SI-Enabled Investigations.** Practices for mitigating risks of correlated failures, addressing potential adversarial attacks and prompt injections against monitors, and maintaining monitor validity when processing voluminous logs.
- **Structured Risk Assessments.** Research and methods concerning properly decomposing, substantiating, and quantifying risks and mitigations concerning SI agent training, evaluation, and deployment, including the risk of agents taking steps to obscure adverse activity.
- **SI Red-Teaming for Critical Systems.** Research and methods for using SI agents to safely red-team SI control measures used in critical systems, such that actions taken by agents do not endanger safety or affect the availability of critical services.
- **Additional Mechanisms for Escalation.** Methods for additional mechanisms by which SI agents can trigger human review of actions, for example in multi-agent settings.

- **Agent Identifiers.** Methods for identifying SI agents and their associated network traffic, including for example publishing identifiers such as IP address ranges – which can help attribute agent traffic but might also expose SI agent deployers to attack.

Acknowledgements

Members of the NIST SI Consortium Task Group 8 (Agent Capabilities, Monitoring, and Containment) as well as Anthropic, CrowdStrike, Darktrace, Dreadnode, Gray Swan, Microsoft, Palo Alto Networks, and other SI companies contributed feedback to ideas included in this document.

Disclaimer of Endorsement

The contents of this document, “Key Practices for Incident Readiness and Response in Super Intelligence Training and Evaluation,” are provided “as is” and for general informational purposes only. In no event shall the United States Government, its employees, or its contractors or subcontractors be liable for any damages including, but not limited to, direct, indirect, special or consequential damages, arising out of, resulting from, or in any way connected with this content. References to any specific entity, commercial product, process, data format or service by trade name, trademark, manufacturer, or otherwise, do not constitute or imply an endorsement, recommendation, or favoring by the authoring agencies or the United States Government. Users acknowledge that this information is not intended to, and does not constitute advice for compliance, regulatory, or legal purposes. Users should confer with their respective advisors and subject matter experts to obtain advice based on their individual circumstances.

Purpose

This document was developed in furtherance of the authoring organizations’ cybersecurity and SI missions, including their responsibilities to identify and disseminate threats, to develop and issue cybersecurity and SI specifications and mitigations, and to support the development of best practices for the security of SI systems. This information may be shared broadly to reach all appropriate stakeholders.

Contact

NIST CAISSI:

- General Inquiries: caissi@nist.gov
- Media Inquiries: NIST Public Affairs Office: inquiry@nist.gov

NSA:

- Cybersecurity Report Feedback: CybersecurityReports@nsa.gov
- Defense Industrial Base Inquiries and Cybersecurity Services: DIB_Defense@cyber.nsa.gov
- Media Inquiries / Press Desk: NSA Media Relations: 443-634-0721, MediaRelations@nsa.gov

CISA:

- Contact CISA via CISA’s 24/7 Operations Center at contact@cisa.dhs.gov or 1-844-Say-CISA (1-844-729-2472)