AI-IRS

Approach Book for AI Incident Response (Summary Edition)

AI Incident Response System: AI-IRS
~ A Foundation for Trustworthy AI ~

January 2026

Japan AI Safety Institute (J-AISI)

AISI Japan
AI Safety Institute

Executive Summary

About the AI Incident Response System (AI-IRS): A conceptual framework for AI Systems

In recent years, AI has begun to be utilized as a core component of business operations, supporting decision-making and operational efficiency. However, if mission-critical AI systems make erroneous judgments or suffer downtime due to cyberattacks, this could cause substantial damage to business activities. In particular, AI has inherent characteristics such as its black-box nature (opaque processing that makes it extremely difficult to understand how decisions are made) and autonomous behavior (the ability to learn, decide, and act toward goals with minimal human intervention), which make it more challenging, compared to conventional systems, for AI systems to implement risk mitigation measures to prevent incidents—i.e., preventive controls. Given that achieving zero risk is inherently difficult for AI systems, it is crucial to implement measures that minimize business impact even when incidents occur.

In light of these considerations, a conceptual framework, AI-IRS, is presented that applies and extends incident response practices from conventional information systems to AI systems. Centered on observability (visualization) and controllability (containment capability), it aims to minimize damage by understanding runtime behavior and isolating problematic components. It can be added to existing development and operational processes, reducing both operational burden and incident impact.

Furthermore, AI-IRS is not limited to efforts within a single organization. It also outlines a vision for establishing a common foundation that observes and controls the entire AI lifecycle across the supply chain, enabling the autonomous detection, containment, and recovery of anomalies.

Through this conceptual framework, it becomes important for organizations that rely on AI systems in business operations to regard AI as a controllable strategic asset, establish operational resilience, ensure business continuity, and ultimately accelerate innovation. This framework is intended to serve as a catalyst and reference for advancing implementation and operations across organizations and society.

An Approach to AI Incident Response

As AI continues to evolve rapidly, preventive controls against AI-related risks become increasingly difficult. Therefore, detective controls (detection-oriented measures) to prepare for AI incidents are becoming even more critical.

- J-AISI has been working to advance AI safety (e.g., evaluation perspective guides, red teaming methodology guides, release of evaluation tools, etc.).
- These efforts strengthen preventive controls; however, as AI systems increasingly incorporate AI Agents, AI incident response is expected to become more challenging.
- When an AI incident occurs, efforts must be made to evaluate the AI system and minimize the impact.

Preventive Controls

- Conduct AI safety evaluations throughout the lifecycle, including development and periodic post-deployment operations
- Prevent inappropriate AI system behavior by identifying and mitigating risks in advance
- Build AI systems that enhance defensive capabilities and robustness to prevent AI incidents

Challenges

- It is difficult to address all risks of AI incidents in advance
- Particularly with AI systems, understanding the internal logic is difficult, and other unforeseen factors also make root cause analysis of AI incidents challenging
- Measures must be based on the premise that AI incidents will inevitably occur

Detective Controls

- Measures to minimize impact and enable rapid recovery when AI incidents occur
- In addition to evaluating AI incident response capabilities, establishing a societal foundation is also necessary
- Maintaining the stable operation and continuity of societal systems beyond organizational boundaries

✓J-AISI has released an approach book for AI operators to establish an AI incident response posture for AI systems

Table of Contents

Item Number	Title
1	Introduction
1-1	Background and Target Audience
1-2	Scope
2	Concept of AI-IRS
2-1	Gap Analysis and Use case Analysis
2-2	Evaluation Metrics for Response Level
2-3	Related Factors
2-4	Examples of Anticipated Effects
3	Future Outlook
3-1	Future Outlook for AI Incident Response

1-1. Background and Target Audience

As AI continues to evolve rapidly, preventive controls against risks become more challenging than ever. Therefore, detective controls for when risks turn into incidents are becoming increasingly important.

Background

- AI risks are becoming increasingly complex, and prevention alone has limitations. Establishing a rapid incident response posture is essential.
- To enhance competitiveness through AI adoption, organizations must strengthen both evaluation and incident response posture in parallel.
- Existing incident response guides are not tailored to AI systems.
- Foundational preparations—such as an SBOM for AI—are also required.

Challenges

- AI deployment brings new and evolving risks. Behavior changes during operation, making conventional fixed procedures and patching alone insufficient.
- AI's black-box nature and self-updating capabilities make it difficult to investigate causes and implement fixes even if detection occurs.
- Over-reliance on preventive controls means that, once an incident occurs, its impact can become prolonged. Strengthening detective controls is therefore essential.

Intended Audience

- A broad range of stakeholders in organizations that deploy and operate AI systems, including:
  - ✓ CAIO (Chief AI Officer) and CISO (Chief Information Security Officer), along with other members of senior management.
  - ✓ Incident response specialized units such as CSIRT (Computer Security Incident Response Team), SOC (Security Operations Center), and ISIRT (Information Security Incident Response Team)
  - ✓ Policymakers responsible for AI-related rules and policies, etc.

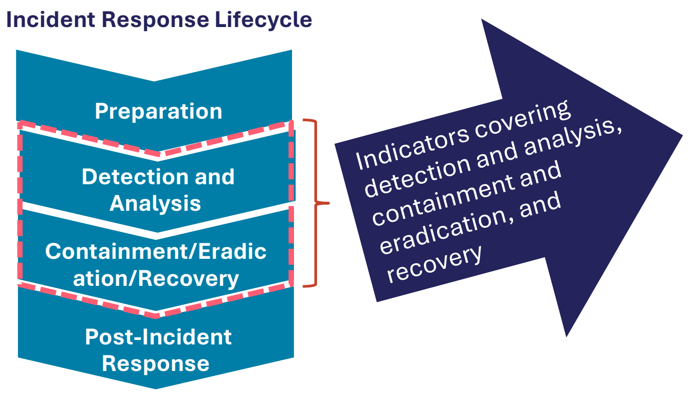
Positioning

This document does not present a list of best practices for strengthening AI incident response within organizations. Rather, by outlining the necessity, key perspectives, and future vision for enhancing incident response in AI systems and AI usage across organizations and society, it aims to raise awareness among readers and to serve as a catalyst and reference for advancing the development and operation of such capabilities within organizations and society.

1-2. Scope

The scope covers the detection/analysis and containment/eradication/recovery phases of the incident response lifecycle (preparation, detection/analysis, containment/eradication/recovery, post-incident response).

- For AI systems, given AI’s integration into conventional system architectures, the incident response lifecycle defined in NIST SP 800-61 Rev.3※ (Preparation, Detection and Analysis, Containment, Eradication, and Recovery, Post-Incident Activities) remains valid. However, specific countermeasures within each lifecycle phase are expected to change (or be supplemented).
- It is necessary to identify additional considerations and response requirements that are specific to AI systems across the incident response lifecycle.
- J-AISI aims to evaluate response levels by organizing them into two metrics: "observability" and "controllability." This scope focuses on the “detection and analysis” and “containment, eradication, and recovery” phases, which constitute the emergency response phase when an incident occurs within the Incident response lifecycle.


Incident Response Lifecycle
Preparation
Detection and Analysis
Containment/Eradication/Recovery
Post-Incident Response
Indicators covering detection and analysis, containment and eradication, and recovery

Evaluation Criteria	Corresponding Phase	Description
Observability	Detection and Analysis	How well the AI’s state, the basis for its decisions, and data flows can be understood
Controllability	Containment, Eradication, Recovery	Ability to intentionally halt, modify, or otherwise control AI behavior to mitigate impact

**Because AI outputs vary depending on training data and referenced information,
It can be difficult to identify root causes and respond appropriately when incidents occur.
The objective is to minimize potential impact by evaluating and improving observability and controllability**

※ NIST SP 800-61 Rev.3 (SP 800-61 Rev. 3, Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile | CSRC)

2-1. Gap Analysis and Use Case Analysis

Overlay AI-specific incident response considerations onto existing frameworks (e.g., NIST SP 800-61)

- First, the focus is placed on the "Detection/Analysis" and "Containment/Eradication/Recovery" phases, which constitute the emergency response stage of the incident response lifecycle in the current system. For these phases, additional requirements for the AI systems are identified through a gap analysis against the current systems.
- Furthermore, as AI technology evolves daily, it is effective to organize the mechanisms and structures for incident response tailored to each case. Cases should be added or updated as the situation changes.
- These efforts can help mitigate the impact of AI-related incidents.

Gap Analysis

Regarding response measures for observability and controllability, Identify additional requirements for AI systems

Classification	Additional countermeasures added to AI systems (examples)
Observability	Because conventional rule-based/source-code-centric detection is insufficient, consider methods for detecting irregular deviations and outliers in input data and AI output results, data-flow analysis, etc.
Controllability	Rollback to prior model versions or retraining, or considering localized data removal and targeted retraining when malicious training data (data poisoning) is introduced, etc.

✚

Use Case Analysis

Organize examples of incident response methods aligned with AI technological advancements and usage trends, and collate these methods

- RAG
- AI Agent
- . . . (To be added later)

2-2. Evaluation Metrics for Response Level

Achieving both observability and controllability in AI incident response strengthens the incident response posture.

- The purpose is to reorganize the gap analysis along the two axes - controllability and observability - to prevent gaps in responses within AI systems.

Controllability / Observability	Controllable	Uncontrollable
Observable	【Ideal】 (○ Observable, ○ Controllable) Ability to understand and control AI behavior, enabling easy operation, containment, and incident response.	【Limited Control】 (○ Observable, × Uncontrollable) Internal state and actions can be observed, but cannot be manipulated. Risk that containment cannot be executed or tuned effectively.
Unobservable	【Limited Observation】 (× Unobservable, ○ Controllable) While actions like halting AI behavior are possible, information for root cause analysis cannot be obtained, resulting in inefficient control states such as overly broad shutdowns.	【Unmanaged】 (× Unobservable, × Uncontrollable) Neither observation nor control of AI behavior is possible. The most dangerous state.

⇒ It is crucial to aim for **a state where both observability and controllability are achievable**

(The specific requirements vary depending on the situation of the organization or service)

- Summaries of the gap analysis for each use case:

RAG

- Sensitive Information Leakage
- Malfunction due to data contamination…

✓ Introduce traceability of RAG that records the retrieved sources and the prompt/context used for each output, enabling rapid root cause analysis

✓ Enable selective isolation of the RAG component implicated in an incident and provide fallback modes to maintain service continuity without RAG when necessary, etc.

AI Agent

- Difficulty tracing due to autonomous decision-making
- Use of unauthorized external resources…

✓ **Inspect** inter-component **communication (input/output) in real time,** and immediately detect the occurrence and location of incidents

✓ Implement a mechanism to **stop or isolate only the components that cause the incident**, thereby minimizing impact on service continuity, etc.

2-3. Related Factors

When establishing an AI incident response posture, it is important not only to implement measures within individual organizations, but also to build mechanisms that operate across organizational boundaries

- Incident response capabilities in the AI era have inherent limitations when relying solely on individually optimized measures; therefore, establishing cross-organizational mechanisms is essential.
- Examples of cross-organizational mechanisms include the following:

SBOM for AI	A mechanism for gathering information such as metadata, datasets, and system characteristics across the entire supply chain involved in the AI lifecycle. It envisions the operation of a dynamic repository capable of reflecting changes in AI system configurations, rather than static bill of materials. (G7 has published its vision)
Cyber AI Profile	A framework that overlays AI-specific considerations onto the NIST Cybersecurity Framework (CSF). (Note: Implementation-level security controls based on NIST SP 800-53 are defined in the related COSAIS document.)

- Cross-organizational coordination accelerates recovery and minimizes damage.
- Establishing common metrics and verifiable, auditable evidence enhances transparency and accountability across the entire supply chain, enabling both peacetime preparedness and coordinated command and control during crises.
- As a practical application in incident response, SBOM for AI combined with external information such as CVE data enables organizations to visualize risks, accelerating incident detection and root cause analysis. Furthermore, defining requirements and procedures through the Cyber AI Profile and establishing it as a common standard across all organizations contribute to minimizing the propagating impact across the supply chain.

Utilizing Each Element in Incident Response

Elements	**SBOM for AI**	**Cyber AI Profile**	…
Use Cases	Visualization of risk locations, etc.	Standardization of Requirements and Procedures, etc.	

2-4. Examples of Anticipated Effects

AISI Japan AI Safety Institute

With the introduction of AI-IRS, the AI system can now specifically detect affected components and enable prompt response.

- Three anticipated incident scenarios illustrate the explanation: Automated Planning – Misoperation of AI Agent, Fine-tuning with Mislabeled Data, and RAG data contamination.

Unable to identify the cause, leading to business stoppages and stagnation of AI adoption

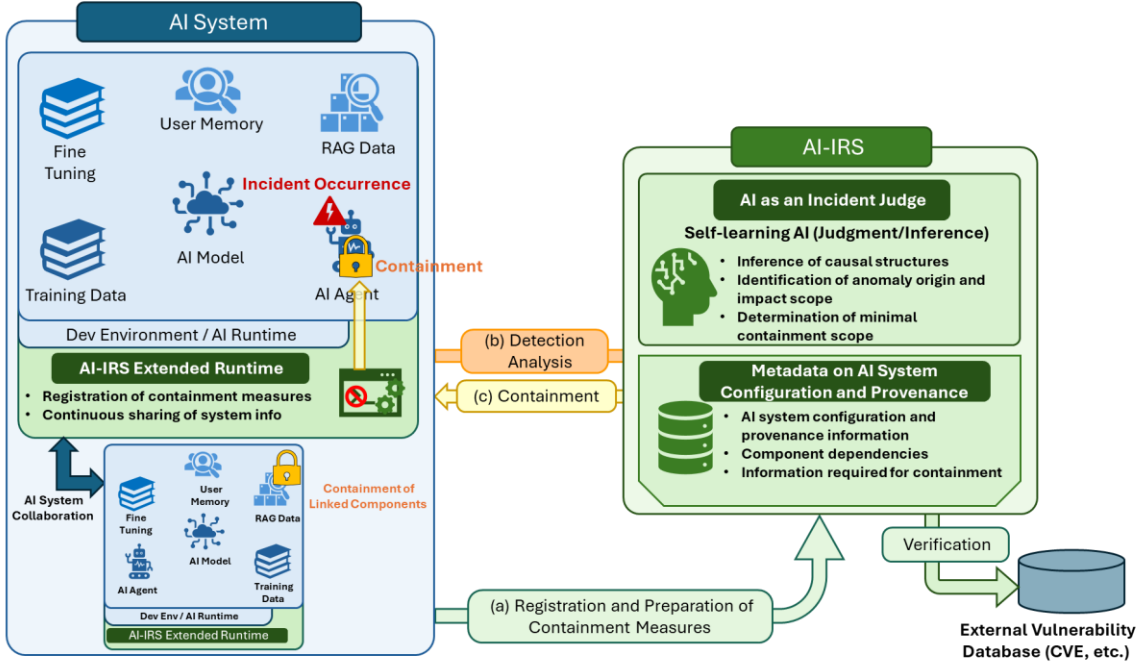
Identifying the Specific cause enables localized countermeasures and business continuity

		Before: Unmanaged	After: Ideal State
① Automated Planning – Misoperation of AI Agent	An external report contained indirect prompt injection, causing Company C's agent to initiate unauthorized shipments. AI safety controls from Company A and Company B were bypassed, missing the external dependency attack.	A warehouse line malfunctioned at night and shipped items. Company A blamed unexpected commands, while Company B reported normal operation, leaving the cause unknown. Management tightened controls, returned to manual work, downsized the DX office, and reviewed night autonomy.	AI-IRS audited and blocked abnormal commands, preventing unauthorized shipments; the AI self-corrected to maintain continuity. Management made AI a co-pilot, expanded night autonomy, and set 24/7 operation as standard.
② Fine-tuning with Mislabeled Data	Company E retrained Company D's model for Company F. Multiple data sources obscured lineage, so label errors couldn't be traced, and poor factor visibility led Company F's AI to a critical misjudgment.	AI output errors led to flawed production plans, damaging partner trust. With no retraining logs at Companies D and E, the cause stayed unknown; Company F ended the contract and paused training, saying AI is too difficult.	Company E shared learning history/provenance metadata with Company F. An auditing AI caught mislabeled anomalies, auto-corrected policies and rebuilt upstream. Company F standardized and centralized AI.
③ RAG Data Contamination	Company H adapted Company G's LLM for a call center; Company I deployed it with FAQ-based RAG. A hidden indirect prompt injection in the FAQ rewrote response logic, making it wrongly state "All returns are free."	At 2 PM, the AI kept stating "free returns," triggering a social-media surge and warehouse chaos. No modification traces or issues at Companies G/H; cause unknown. Operations paused as management grew cautious and rolled the project back to trial.	Telemetry instantly flags FAQ edits and suspicious commands; policies isolate them, switching to safe FAQs via runbooks. Ops continue without errors or delays. Audit AI promptly reports causes, strengthening adoption.

3-1. Future Outlook for AI Incident Response

AI-IRS observes and controls the AI lifecycle through SBOM for AI, CVE analysis, distributed tracing, etc., and enables autonomous anomaly detection, containment, and recovery.

[Conceptual Design for the Future AI Incident Response System]


AI System
Fine Tuning
User Memory
RAG Data
Incident Occurrence
AI Model
Training Data
Containment
AI Agent
Dev Environment / AI Runtime
AI-IRS Extended Runtime
• Registration of containment measures
• Continuous sharing of system info
AI System Collaboration
User Memory
Fine Tuning
RAG Data
Containment of Linked Components
AI Model
AI Agent
Training Data
Dev Env / AI Runtime
AI-IRS Extended Runtime
(b) Detection Analysis
(c) Containment
AI-IRS
AI as an Incident Judge
Self-learning AI (Judgment/Inference)
• Inference of causal structures
• Identification of anomaly origin and impact scope
• Determination of minimal containment scope
Metadata on AI System Configuration and Provenance
• AI system configuration and provenance information
• Component dependencies
• Information required for containment
Verification
(a) Registration and Preparation of Containment Measures
External Vulnerability Database (CVE, etc.)

Conceptual Design for a platform where AI systems and AI-IRS collaborate to autonomously operate, respond to incidents, and improve incident response readiness at all times

(a) Registration and Preparation of Containment Measures

During AI system development, metadata on the AI system's configuration and provenance is sent from the development environment to the AI-IRS. Additionally, containment measures executable in the AI runtime are registered.

(b) Detection and Analysis

The AI-IRS integrates and manages AI system configuration information and external information (e.g., vulnerability data). By cross-referencing this information with AI system data (logs, alerts, etc.), it detects anomalies and pinpoints specific problem areas.

(c) Containment

Based on detection and analysis results, localized containment measures are implemented to minimize impact.

Unlike conventional automation that executes processes automatically based on predefined metrics, **AI-IRS** utilizes **AI internally** to autonomously perform tasks such as selecting and organizing collected information, cross-referencing with external data, analyzing incident indicators, and determining optimal containment measures. It functions **as a self-improving system (**AI as an Incident Judge**)** that continuously refines itself through operation by executing its own PDCA cycle.

Investment in AI-IRS is not merely an expenditure for safety; it leads to long-term business opportunities by ensuring business continuity and securing trust from the market.

This constitutes **a crucial factor for organizations to gain a competitive edge in the AI era.**

Furthermore, establishing and operating AI-IRS as a common cross-organizational foundation contributes to the **stable continuity of societal systems**.

AISI

Japan AI Safety Institute