AI-IRS

Approach Book
for AI Incident Response

~ A Foundation for Trustworthy AI ~

January 9, 2026

AI Safety Institute

AISI Japan AI Safety Institute

Table of Contents

0 Executive Summary

As AI increasingly becomes a foundational element supporting business operations such as decision-making and operational efficiency, AI incidents resulting from AI malfunctions could directly threaten the survival of organizations and businesses. Such incidents may directly affect business continuity and management accountability, leading to erroneous management decisions, losses from automated transactions by AI agents, or damage to organizational brands through the spread of misinformation.

This document presents the concept of the AI Incident Response System (AI-IRS) as a conceptual framework for extending existing incident response capabilities to address risks arising from AI system operations.

Because AI systems autonomously make decisions and dynamically change their behavior — unlike existing rule-based systems— it is difficult to preemptively identify risks and address them. Operating under the premise that “incidents are inevitable” is therefore considered necessary not only to strengthen preventive controls but also to appropriately enhance detective controls that aim to minimize damage when incidents occur.

Detective controls are a core element of existing information system incident response frameworks (such as NIST SP 800-61). However, these do not adequately address situations unique to AI, including unobservable decision rationales and uncontrollable abnormal behavior. This creates risks of failing to grasp the scope of impact when an AI incident occurs or to fulfill organizational accountability obligations, which could in turn lead to excessive system-wide suspension.

AI-IRS introduces observability (visualization) and controllability (containment) as evaluation criteria to address the dynamic risks unique to AI systems. It provides a foundational framework for minimizing damage after incidents occur, offering an implementation-level guidance for observing AI system behavior during operation and controlling problematic areas. The implementation involves adding this conceptual framework within existing organizational frameworks (such as development processes, operational processes, security governance, AI governance). It is essential to ensure transparency and traceability not only within internal organizational efforts but throughout the entire AI lifecycle spanning the supply chain.

Organizations promoting AI adoption can use this document as a reference to advance concrete discussions. Doing so can strengthen incident management for AI and contribute to establishing a foundation for AI as a controllable business asset. Consequently, organizations could establish operational resilience for safer use of AI, ensuring long-term business continuity, fulfilling accountability.

These efforts are expected to sustain AI systems as a cross-organizational social foundation and to promote the advancement of AI utilization across society.

1 Purpose, Background, and Challenges

1.1 Purpose

The purpose of this document is to articulate a conceptual framework that enables organizations and society to respond effectively to the dynamic risks unique to AI systems, minimize damage when AI incidents occur, and contribute to the continuity of business activities.

By presenting the necessity, key perspectives, and future vision for strengthening incident response for AI systems and their use, this document aims to serve as a catalyst and reference for advancing both the development and operation of such a framework within organizations and society. Detailed insights obtained through design and operational activities are expected to be progressively articulated and organized in the form of related deliverables, such as implementation guides and procedures, in the future.

1.2 Background

In recent years, AI has rapidly proliferated, moving beyond mere operational support tools to become deeply integrated into mission-critical organizational systems. However, failures in AI-integrated systems could have consequences extending far beyond mere system downtime, potentially directly affecting related business operations and leading to a loss of social trust. Therefore, in addition to existing information systems, organizations' CSIRTs (Computer Security Incident Response Teams) are increasingly required to establish an operational posture for AI incident response in order to protect the stable operation of AI systems.

Incident response in conventional information systems has been systematized based on the National Institute of Standards and Technology (NIST) Special Publication 800-61r3 "Incident Response Recommendations and Considerations for Cybersecurity Risk Management" [1]. The document organizes the response process into four stages: "Preparation", "Detection and Analysis", "Containment, Eradication, and Recovery," and "Post-Incident Activities." It is aligned with the components of the NIST Cybersecurity Framework (CSF) 2.0. However, this document targets existing information systems and was not written with the intent of addressing AI, including the dynamic behavior of AI models whose state changes during operation, or the dynamic structures (such as RAG and AI agents) that rely on external dependencies.

AI systems operate through the interdependence of AI models, training data, algorithms, external services, and other components, which together results in a more complex risk structure than that of conventional information systems. Notably, AI systems possess "dynamic" and "autonomous" characteristics, meaning their behavior can change during operation. These dynamic properties and the complex dependencies between AI components

increase the potential for a single anomaly to propagate through a chain reaction. Risks undetected during development and implementation may surface during operation. This necessitates dynamic countermeasures tailored to AI's characteristics, as well as integrated risk management that encompasses the entire AI supply chain, extending beyond individual organizational measures. Implementing these requires systematically developing actionable countermeasures suited to each organization's AI usage patterns, while considering international standards and technological trends.

1.3 Challenges

As noted above, it is difficult to address all risks in AI systems through preventive controls that are designed to identify risks in advance. Consequently, detective controls are expected to become increasingly critical.[1]

The following challenges exist in the conventional approach based on conventional detective controls:

(i) Difficulty in Root Cause Analysis

AI systems involve complex interdependencies among AI models, data, and external services, making it difficult to clearly articulate the root cause when anomalies occur. The inherent black-box nature and low reproducibility of such AI, stemming from its probabilistic behavior, further complicate root cause analysis and remediation.

(ii) Limitations of Fixed Responses

AI autonomously updates and evolves during operation, meaning existing fixed countermeasures may lead to excessive downtime. This makes it difficult to determine the appropriate scope for halting operations in order to minimize the impact stemming from these updates or changes when anomalies occur.

---

[1] In information systems, "preventive controls" have traditionally been emphasized under the principle that incidents should be avoided. However, recognizing that completely preventing incidents is impossible due to factors like increasingly sophisticated cyberattacks, there is now a demand for "detective controls" that minimize impact even if an incident occurs. Nevertheless, according to the "Survey on the Actual State of Information Security Measures in Small and Medium-sized Enterprises 2024", only a low percentage of organizations have prepared by establishing emergency response systems and procedures in anticipation of security incidents. The survey raises an alarm, stating that "small and medium-sized enterprises themselves must understand the risks of damage from cyber incidents and consider establishing response systems and procedures for when incidents occur."
(https://www.ipa.go.jp/security/reports/sme/nl10bi000000fbvc-att/sme-chousa-report2024r1.pdf)

This document delves into countermeasures centered on two axes: "the extent to which AI can be observed (observability)" and "the degree to which it can be controlled (controllability)" to address these challenges.

Furthermore, AI inherently faces fundamental limitations: its environment and behavior change during operation, and manual improvements alone cannot fully address new anomalies. In addition, this document explores the use of AI to continuously enhance observability and controllability in response to these challenges.

Subsequent chapters present policies and response principles aimed at strengthening AI incident response.

1.4 Scope

This document presents an approach that addresses challenges unique to AI systems by using the incident response process structure presented in NIST SP 800-61r3, which is an internationally recognized framework, as a reference point, while taking into account the dynamic and autonomous characteristics of AI systems. It positions the capabilities required for "Detection and Analysis" as observability and those required for "Containment, Eradication, and Recovery" as controllability, and adopts these two axes as the primary evaluation criteria for AI-specific incident response in the subsequent discussion. Figure 1 illustrates this relationship.

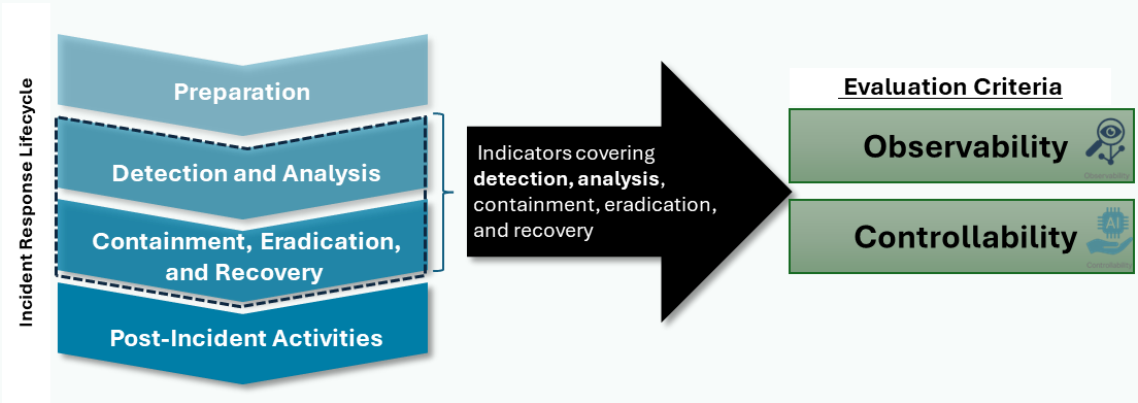

Figure 1. Incident Response Lifecycle and Two Evaluation Criteria for AI

It should be noted that the framework presented in this document does not replace existing information security incident response frameworks. To address the characteristics of AI, it extends the capabilities associated with both axes: "Detection and Analysis" and "Containment, Eradication, and Recovery."

Classification	Description
Observability	This evaluation criterion corresponds to the "Detection and Analysis" phase of incident response. It indicates the extent to which an AI system's basis for decision-making, internal state, and data flow can be monitored in real time. It serves as a key indicator for enabling rapid and accurate root cause identification during an incident.
Controllability	This evaluation criterion corresponds to the "Containment, Eradication, and Recovery" phase of incident response. It indicates the degree to which the impact of autonomous AI behavior within an AI system can be minimized based on the situation. It serves as an indicator for enabling localized and dynamic control during an incident.

This document classifies incident response capabilities in AI systems into four quadrants based on combinations of observability and controllability, thereby proposing an approach for incident response and presenting specific requirements tailored to each quadrant.

1.5 Intended Audience

This document is primarily intended for a broad range of stakeholders within organizations that introduce and operate AI systems in their business operations, as described below.

Executives and enterprise-wide risk management officers responsible for AI deployment and operational risk management and governance—such as the Chief AI Officer (CAIO) and Chief Information Security Officer (CISO)—can use this document to understand the necessity of an AI-specific posture for incident response and to develop organizational policies.

Departments responsible for information security operations, such as CSIRTs, SOCs (Security Operations Centers), and ISIRTs (Information Security Incident Response Teams), as well as administrators and operations personnel for information systems including AI, can use this as a reference resource for formulating and implementing AI incident response policies tailored to their organization and AI-specific requirements. It also aids in designing and operating a response posture that accounts for differences from those for existing information systems.

Policymakers and others responsible for establishing common rules and guidelines for AI can utilize this information as a reference for policy formulation that promotes AI adoption while ensuring the stable continuation and maintenance of societal systems.

Even in small and medium-sized enterprises or organizations with limited personnel, where IT administrators or information systems staff handle security and AI system operations concurrently, this document can serve as a reference for establishing an incident response posture within a realistic scope and exploring directions for operational improvements.

1.6 Terminology

Terminology	Meaning
Incident	A cybersecurity incident (or simply an incident) is 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. [FISMA2014]
AI Incident	Refers to events where the behavior of an AI system deviates from ethical norms or safety standards in unexpected ways, exceeding the risk tolerance of society or an organization. It is not necessarily identifiable as AI-related at the time of occurrence. In many cases, it is detected as a cybersecurity incident through conventional incident response detection processes, and analysis of its cause suggests that it is an anomaly related to AI.
Incident Response	All activities related to responding to incidents, ranging from incident detection and the collection of necessary information and data to the implementation of emergency measures, identification of the scope

Terminology	Meaning
	of impact and root cause, coordination with relevant parties, and recovery efforts. [JPCERT]
AI-IRS: AI Incident Response System	A conceptual framework integrating design and operation, centered on "observability" and "controllability," to address dynamic risks unique to AI.
Posture	Refers not to the organizational structure itself, but to a state in which the organization has the necessary preparations in place and its functions are actually operating.
AI Incident Response Posture	A state in which the organization goes beyond merely establishing formal structures and can assess situations, make judgments, and respond to anomalies arising from AI behavior, while also being able to explain the rationale for these actions retrospectively.
AI as an Incident Judge	An AI that autonomously infers causal structures from information about observed incidents and makes judgments regarding the scope of impact and its minimization.
AI Agent	An automated entity that senses and responds to its environment and takes actions to achieve its goals. [ISO/IEC 22989:2022] Various other definitions and explanations exist beyond those stated above. Regardless of the definition applied, such entities fall within the scope of this document when risks are anticipated.

2 Approach to Establishing an AI Incident Response Posture

[Key Points of Chapter 2]

- **Clarifies that the dynamic nature of AI makes existing incident response for information systems inadequate. This clarifies the necessity to enhance observability and controllability and establish a capable response posture.**
- **Conducts Gap Analysis against conventional frameworks and analyzes representative use cases to concretely identify gaps from the ideal state.**
- **Presents the fundamental concepts of AI-IRS centered on observability and controllability, along with supporting elements and the institutional/strategic perspectives required for AI utilization.**

This document presents the concept of the "AI-IRS: AI Incident Response System" as a conceptual framework for establishing an incident response posture for AI system operations.

As shown in the previous chapter, the challenges in incident response for AI systems stem from AI-specific characteristics such as black-box nature and autonomous updates through learning, which make observation and control difficult.

The challenges outlined in the previous chapter are addressed through the following approaches.

(i) Difficulty in Root Cause Analysis

The challenge can be addressed by visualizing the structure and external dependencies of AI and by enhancing the ability to measure what is causing problems (observability).

(ii) Limitations of Fixed Responses

The challenge can be addressed by strengthening containment capabilities (controllability) to locally control AI's autonomous behavior and limit the scope of impact when anomalies occur.

As described in Section 1.4, observability and controllability are treated as the two primary axes of AI incident response. The configuration of the target AI system is not fixed but dynamically changes during operation. Furthermore, its optimal state also changes. Accordingly, it is important to establish adaptive control by enhancing observability and controllability and to further equip the system with a mechanism that allows it to adapt to the situation while reconstructing itself in response to unexpected behavior using AI.

This chapter analyzes structural gaps by comparing existing frameworks and examines specific use cases. After outlining the fundamental concepts of AI-IRS, it also describes anticipated effect scenarios.

The technological foundation surrounding AI systems is undergoing significant and continual change. Accordingly, the components and future vision presented in this document are premised on being designed flexibly, taking into account future technological trends, societal demands, operational conditions, and the specific circumstances and technical constraints of each organization.

2.1 Observability and Controllability

Figure 2 visually organizes AI system incident response capabilities into four quadrants based on observability (observable/unobservable) and controllability (controllable/uncontrollable).

"Unobservable and Uncontrollable" indicates the most dangerous "unmanaged situation." "Observable but Uncontrollable" indicates "Limited Control" where anomalies can be recognized but intervention is impossible. "Unobservable but Controllable" indicates "Limited Observation," where causes cannot be identified, leading to excessive countermeasures.

"Observable and Controllable" is the "Ideal." Understanding and controlling AI behavior enables rapid containment and recovery. Aiming for this ideal state and achieving it through stepwise improvement is the fundamental goal.

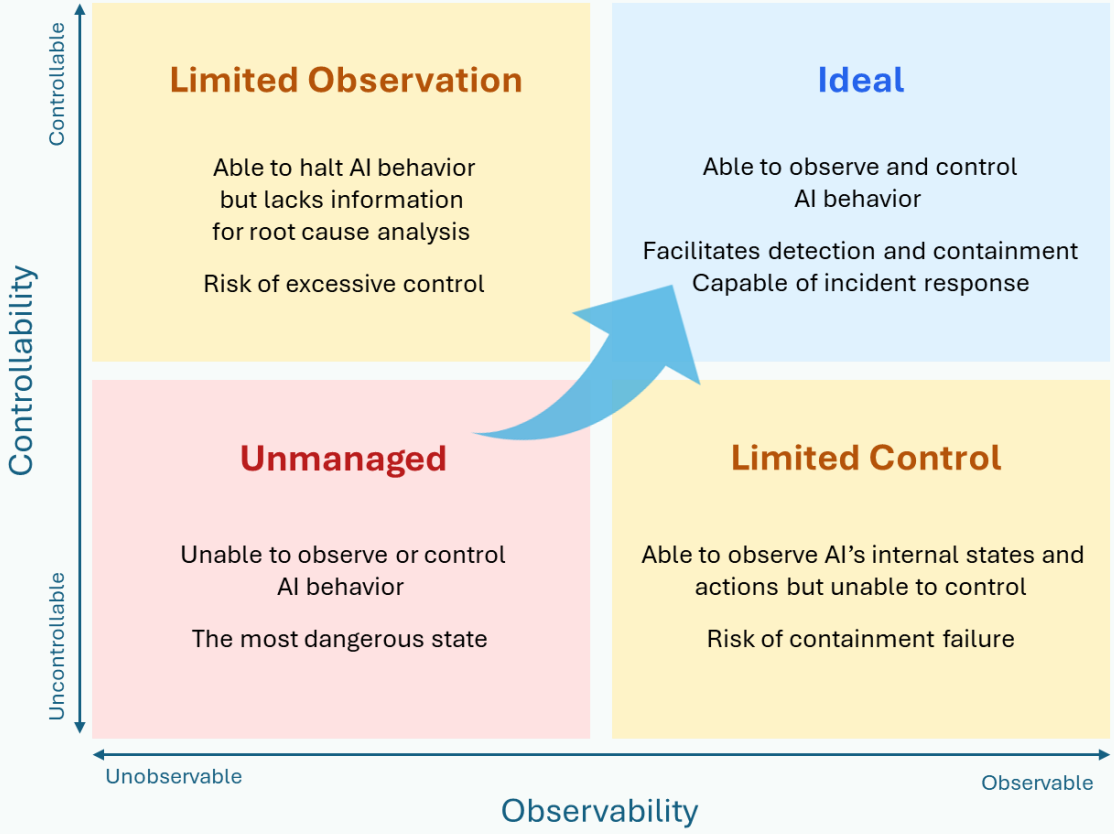

Figure 2. Four-quadrant classification based on observability and controllability

2.2 Gap Analysis: Relationship with Conventional Frameworks

Based on the challenges outlined in the previous chapter, this section identifies key considerations for incident response in AI systems by comparing them with the existing framework, NIST SP 800-61r3.

(a) Insufficient Visualization: "Unobservable"

Existing detection and analysis phases rely on observable information such as logs and alerts to identify anomalies. This mechanism functioned effectively because information system components were relatively fixed, allowing for the comprehensive definition of anomaly criteria and log output requirements during the design phase. However, AI switches both its internal state and external dependencies with each inference, creating an "unobservable" domain within the conventional detection and analysis mechanisms defined in existing incident response frameworks. Furthermore, it is difficult to reproduce the behavior of an AI system that exhibits anomalies after an incident occurs.

To bridge this gap, a visualization mechanism that analyzes the entire AI architecture and comprehensively captures its state is essential. It needs to trace dependencies across distributed components—including models, data, and external APIs—and apply integrated analysis to consolidate the obtained information. This approach also needs to include observation targets that were previously unmonitored, such as search paths within RAG and the decision-making processes of AI agents. Visualizing these enables understanding how anomalies propagated and through which paths, and whether their origin stemmed from "external APIs," "secondary effects due to data contamination," or "AI model hallucinations."

Furthermore, enabling the tracking of AI inference processes over time allows for the recreation of incident scenarios, enhancing the accuracy of after-the-fact investigations. Strengthening this visualization capability is crucial, as it significantly affects the quality of initial responses in AI incident response.

(b) Inability to Address Dynamic Characteristics: "Uncontrollable"

Conventional containment has focused on controlling the scope of impact by segmenting it into units such as devices, servers, clients, networks, processes, applications, or accounts. This approach worked because the composition of information systems was relatively fixed, allowing measures like "stopping" or "isolating" each element. However, because AI behavior evolves through the tight integration of

models, data, and external services, determining appropriate containment units is difficult, leading to risks of over- or under-estimating the scope of impact.

Addressing this issue requires fine-grained control mechanisms that dynamically change the targets of control based on the AI's evolving structure and the situation. For example, mechanisms are needed that combine dependency relationships obtained through visualization with AI system information. This enables localized control at the smallest possible unit, such as switching AI models or isolating problematic external APIs, tailored to the specific AI configuration. This allows incident response while keeping as many unaffected components operational as possible. Consequently, it avoids system-wide shutdowns, minimizes disruption to business operations, and shortens recovery time.

As described above, for AI incident response, it is important to build upon existing frameworks by incorporating enhancements that strengthen visualization as a means to improve observability and enhance controllability to respond to dynamic characteristics in AI behavior—that is, by extending existing frameworks with "enhanced observability" and "enhanced controllability." Furthermore, because AI environments and behaviors frequently change during operation, manual improvements alone have inherent limitations and may fail to adequately address new anomalies. To counter this, it is important to equip the system with a mechanism through which it autonomously and continuously improves its observability and controllability, rather than relying solely on human intervention. This involves recording observation data at the time of incidents, the results of containment actions, and the causal relationships between incidents and system behavior, thereby enabling the system itself to intervene more quickly and accurately when similar incidents recur. Figure 3 illustrates a structure in which the extensions of "strengthened visualization" and "enhanced response to dynamic characteristics," together with a mechanism that enables autonomous and continuous self-improvement, are added as an overlay to the existing incident response lifecycle, without replacing it. This clarifies the key points for addressing the dynamic risks specific to AI systems.

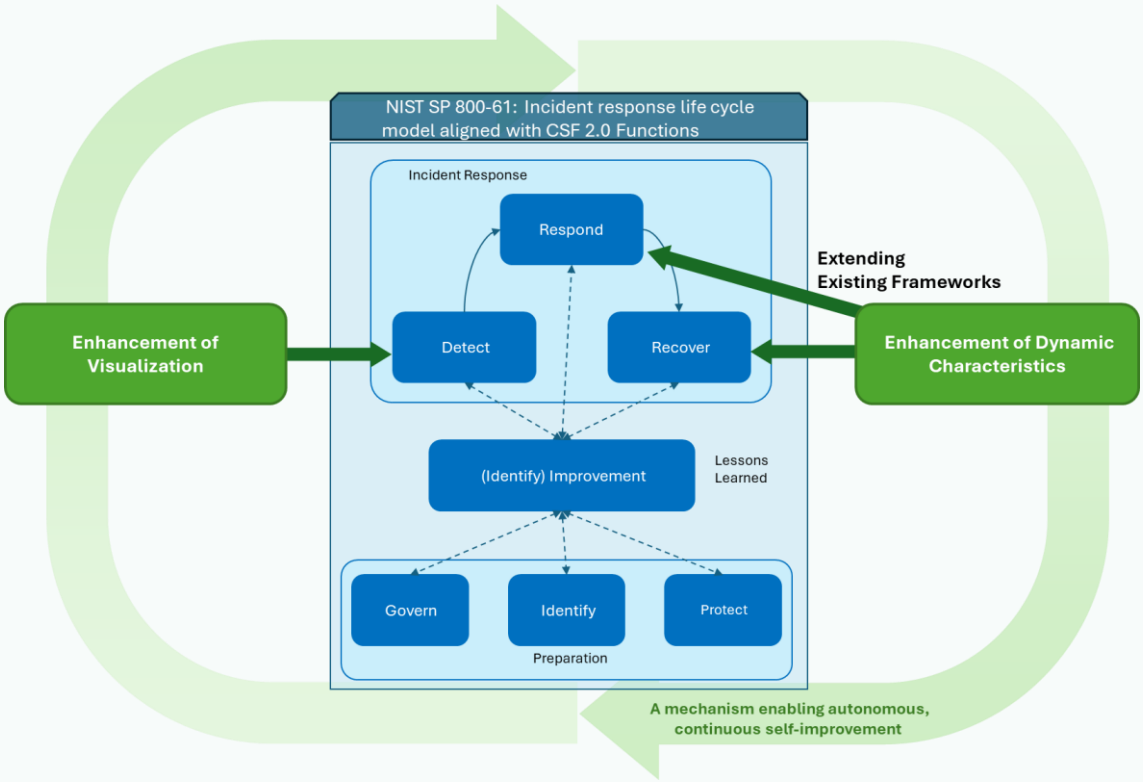

Figure 3. Overlay Structure for Extending Existing Incident Response Frameworks

Based on the gaps clarified above, the next section examines representative use cases to analyze actual operational challenges.

2.3 Analysis of Representative Use Cases

This section examines representative AI system use cases (RAG and AI agents), analyzing anticipated challenges during incidents and key response considerations from the perspectives of observability and controllability. This analysis clarifies specific risks for each AI system use case and presents an approach for addressing them.

2.3.1 RAG (Retrieval-Augmented Generation) Analysis

RAG retrieves information from databases or search APIs separate from the AI model and presents generated results based on the retrieved information. While this structure offers flexibility and scalability, it also makes the issues identified in the Gap Analysis more likely to surface in practice.

- Insufficient Visualization: "Unobservability"

  Because RAG is structurally dependent on databases and APIs, risks include as contamination of referenced RAG data and the uncertain reliability of information provided by search APIs. When observation points are limited to input/output layers, it becomes impossible to track data references or search path behavior. This causes the "Insufficient Visualization" highlighted in the Gap Analysis to manifest itself as an operational issue. Consequently, it becomes difficult to accurately determine the root cause of incidents—whether it stems from "RAG data contamination," "external API issues," or "hallucinations" arising from the AI's own reasoning.

- Inability to Address Dynamic Characteristics : "Uncontrollability"

  Current control points in AI operations are limited to the AI itself, such as model configuration changes, which in turn makes it difficult to control the behavior of external databases referenced by RAG. For example, if a dataset being referenced accidentally contains personal information that should not be disclosed, and there is no means to identify and locally isolate the problematic files or tables, the system may be forced into excessive control measures, such as shutting down the entire system. This causes the "Inability to Address Dynamic Characteristics" highlighted in the Gap Analysis to manifest itself as an operational issue. That is, it becomes difficult to implement controls that isolate the affected components at the smallest possible unit.

Additionally, the cause may be hidden within referenced external data, such as in cases of "indirect prompt injection" where malicious commands or tampering are embedded in the external data. Such incidents tend to have their root cause buried in external update processes, making it difficult to reproduce the situation or pinpoint the cause using conventional static monitoring alone. Consequently, there is a persistent risk that similar issues would recur without effective countermeasures being implemented.

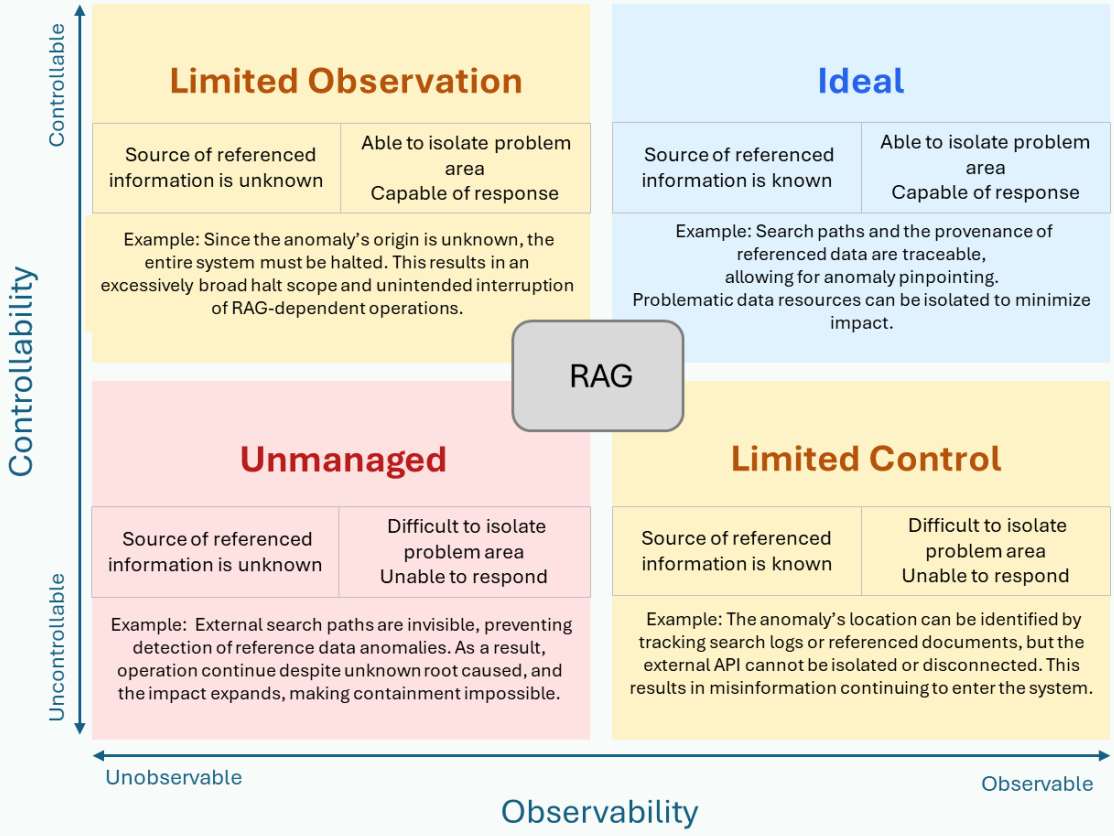

Figure 4. Examples of controllability and observability in RAG systems

Figure 4 organizes examples of controllability and observability in RAG systems.

2.3.2 AI Agents Analysis

AI agents autonomously execute tasks and interact with users and other systems, but this autonomy also carries risks. Since AI agents process tasks by continuously and autonomously operating multiple modules and external services, the issues identified in the Gap Analysis tend to manifest themselves in complex ways. Furthermore, when AI agents are connected to robots or control systems and their decisions result in physical actions, abnormal behavior can occur in the physical world.

- Insufficient Visualization: "Unobservability"

  AI agents dynamically execute multi-stage processes such as planning, API calls, user memory updates, and tool usage. However, their decision-making processes are distributed, making it difficult to determine which decisions have deviated from the intended objective or which external service triggered a malfunction. This "Insufficient Visualization," highlighted in the Gap Analysis, manifests as an operational issue.

- Inability to Address Dynamic Characteristics: "Uncontrollability"

  AI agents operate autonomously, leading to dynamic changes in external API usage, user memory updates, and mid-task policy shifts. This makes it difficult to determine which specific components should be safely isolated or stopped when an anomaly occurs. This causes the "Inability to Address Dynamic Characteristics" highlighted in the Gap Analysis to manifest itself as an operational issue. That is, because it is difficult to isolate and control the affected components at the smallest possible unit, there is a tendency to fall into excessive control measures, such as shutting down the entire AI agent or causing unnecessary business interruptions.

Furthermore, AI agents may manage long-term tasks, accumulating complex histories of internal reasoning processes and external interactions. If an anomaly occurs within an AI agent, it becomes difficult to accurately identify at which stage the erroneous decision was made and which dependencies contributed to it. Consequently, it is expected that conventional operations will be unable to record and analyze this causal structure, making it difficult to establish a continuous improvement cycle.

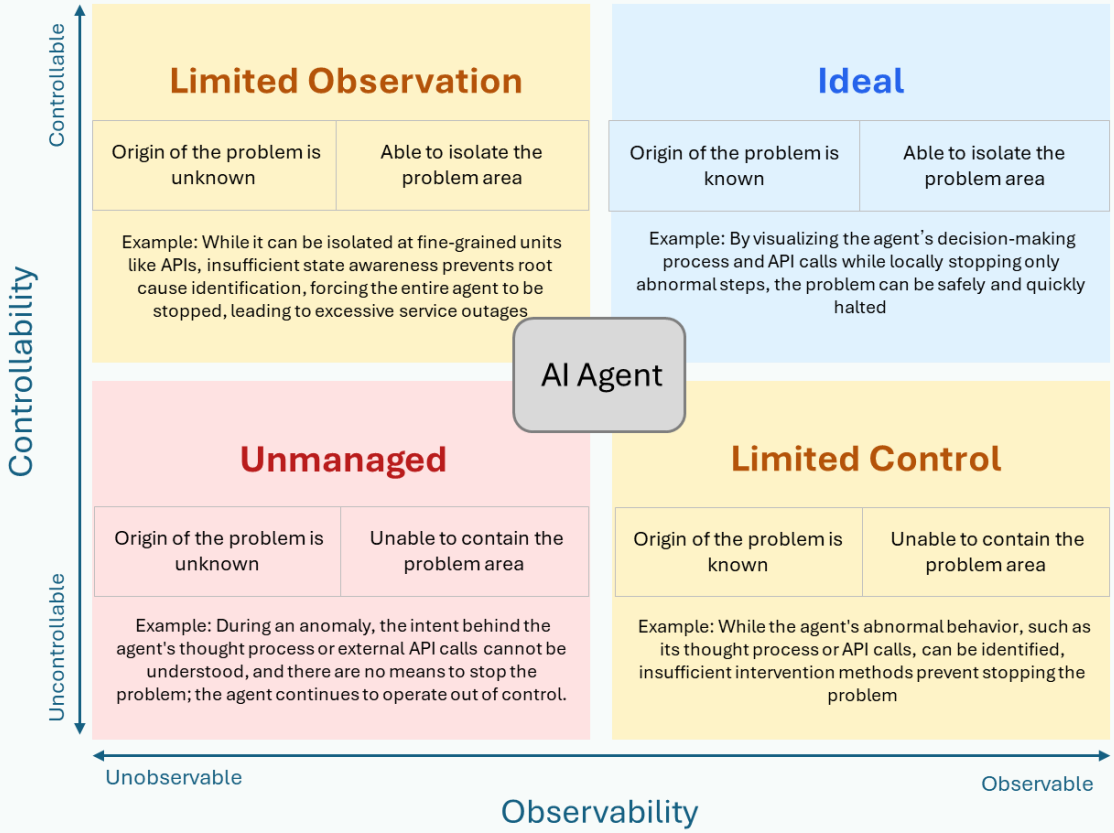

Figure 5. Examples of AI Agent Controllability and Observability

Figure 5 summarizes examples of controllability and observability for AI agents.

2.4 Fundamental Concepts of AI-IRS

This section presents the fundamental concepts of the AI Incident Response System (AI-IRS), drawing on the Gap Analysis and representative use case analysis, and centered on observability and controllability as shown in Figure 1.

When anomalies occur in AI systems, conventional fixed responses are insufficient; mechanisms capable of effectively addressing the dynamic characteristics unique to AI are required. Therefore, to establish an AI incident response posture, it is necessary to incorporate the evaluation criteria of "observability" and "controllability" from the design stage of AI systems.

Figure 6 illustrates the fundamental concept of AI-IRS. It depicts a conceptual framework in which the "ability to measure (observability)"—detecting and analyzing AI incidents—and the "ability to suppress (controllability)"—such as containment—serve as evaluation criteria, and in which observability and controllability are continuously enhanced through the use of AI along these two axes. The AI-IRS concept involves integrating both observation and control capabilities into the overall system design. This involves taking a holistic view of the entire AI system to determine where and how anomalies are detected and how damage is contained within what scope. By incorporating these aspects from the design stage, the long-term safe operation of AI systems becomes possible even within environments where AI systems are deployed.

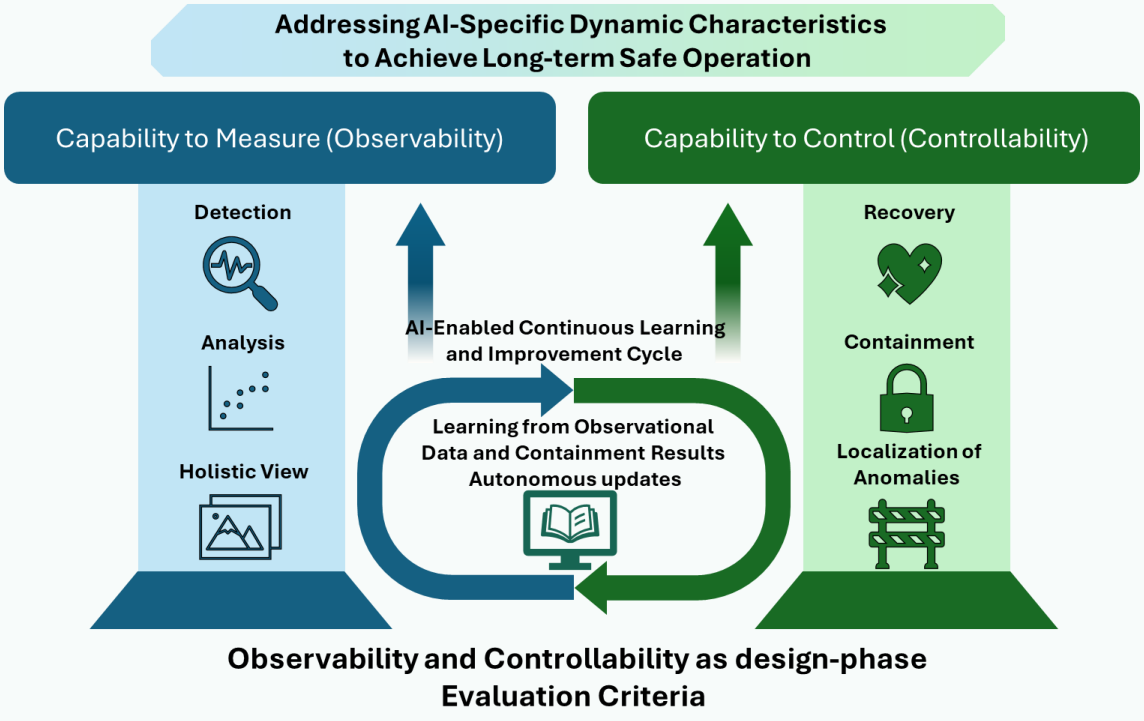

Figure 6. Basic Concept of AI-IRS (AI Incident Response System)

2.4.1 Enhancing Observability

Enhancing observability provides continuous visibility into the internal behavior and dependencies of AI by recording and analyzing learning histories, inference processes, and interactions with external services. This transparency enhances the ability to detect what is happening when anomalies occur and identify their causes. It improves responsiveness during the detection and analysis phase, enabling faster initial responses and root cause analysis when problems arise.

Key examples for achieving enhanced observability include the following:

- Tracking internal reasoning processes and inference histories
- Ensuring traceability of AI components (training data, model information, RAG, etc.)
- Real-time detection of unexpected anomalies in input data and output results

2.4.2 Enhancing Controllability

The goal of enhanced controllability is to enable rapid containment and recovery after anomalies are detected through observation, while minimizing their overall impact. For example, even when AI operates in conjunction with multiple modules or external services, the system should be designed to enable localized actions like stopping specific modules or switching functions, rather than a blanket shutdown of the entire system. This allows maintaining AI's autonomous processing as much as possible, suppressing only affected components at the smallest possible unit, and restoring the system to a safe state.

Key examples for achieving enhanced controllability include:

- Immediate switching to an alternate version of a tampered model or an alternative model, localized removal of malicious training data that has been injected, and/or additional retraining.
- Immediate disconnection of unauthorized API usage by AI agents
- Establishment of an isolated control interface with higher-level privileges that remain unaffected during anomalies

2.5 Elements Supporting AI Incident Response

When establishing an incident response posture for AI, relying solely on individual mechanisms or system countermeasures has inherent limitations. Transparency and traceability must be ensured across the entire AI supply chain, including organizations involved in the development, provision, and utilization of AI systems.

This section explains the role and positioning of SBOM for AI[2] and Cyber AI Profile[3] as elements supporting AI incident response.

It is crucial to monitor these developments and align them with an organization's information security framework.

2.5.1 SBOM for AI

（1） G7 Shared Vision on SBOM for AI

The G7 Cybersecurity Working Group's published document, "A Shared Vision on Software Bill of Materials for AI," presents the common concept of SBOM for AI with the aim of enhancing transparency and traceability across the entire AI system supply chain, thereby improving security and reliability. It proposes a framework for systematically recording and managing information about the diverse components and dependencies that constitute AI systems. AI systems, comprising models, data, and computational resources, along with external base models, datasets, and updates through organizational fine-tuning, exhibit a level of structural complexity that can give rise to vulnerabilities. SBOM for AI addresses this issue by visualizing the internal structure and relationships within AI systems, enabling more effective risk management based on reliable information.

The G7 Shared Vision highlights the need to ensure the following characteristics for SBOM for AI to function effectively:

- Capture both static and dynamic aspects of AI systems
- Be automatically generated and processed in machine-readable formats
- Utilize structured data formats to enable stakeholders to access necessary information

By ensuring these characteristics, SBOM for AI becomes a mechanism supporting transparency throughout the entire AI lifecycle.

The minimum elements that should be included in an SBOM for AI include models, learning, datasets, safety and security characteristics, system level characteristics, key performance indicators, licensing, and infrastructure. Integrating the management of these elements clarifies the components of an AI system and their interrelationships, enabling them to be understood in a traceable manner. Furthermore, it emphasizes the importance of verifying the integrity and authenticity of SBOM for AI components and the entire SBOM through cryptographic hashes and digital signatures. This enables SBOM for AI to contribute to streamlining vulnerability management, license management, compliance verification, and security audits.

The G7 Shared Vision outlines future directions, including developing technical recommendations and implementation guidelines to promote the adoption of SBOM for AI in both public and private sectors.

（2） Relevance of SBOM for AI in AI-IRS

While SBOM for AI is effective for visualizing the provenance and dependencies of components that constitute AI systems, it does not directly cover analysis of AI system behavior or detection of runtime risks. Therefore, for dynamic risks such as abnormal model behavior, prompt injection attacks, malicious data contamination (data poisoning), and AI agent malfunctions or uncontrollable behavior, it is necessary to leverage metadata on AI system configuration and provenance, such as SBOM for AI, to enhance transparency and traceability within the AI-IRS framework.

2.5.2 Cyber AI Profile

（1） Overview and Purpose of the Cyber AI Profile

The NIST Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile) is a project aimed at systematically addressing new cybersecurity challenges arising from AI advancements. It seeks to provide a profile that enables organizations to identify cyber risks associated with AI development, deployment, and operation, and to implement risk-based controls in a systematic manner.

This project explores ways to support organizations facing potential new or expanded risks arising from AI adoption by leveraging existing frameworks such as the NIST Cybersecurity Framework (CSF). This Profile is not meant to replace existing frameworks but provides prioritized guidelines for organizations that focus on the protection of AI systems themselves, countermeasures against attacks exploiting AI, and the enhancement of defenses using AI.

Regarding implementation-focused guidelines, the "Control Overlays for Securing AI Systems (COSAIS)" is being developed in parallel by building upon NIST Special Publication 800-53 (SP 800-53), "Security and Privacy Controls for Information Systems and Organizations." COSAIS is positioned to assist in customizing and prioritizing the most critical controls to consider when using AI systems.

（2） Relevance to AI-IRS

The NIST Cyber AI Profile organizes cybersecurity risk management associated with AI usage. Regarding incident response, it calls for measures such as the analysis of AI-specific artifacts like model logs and the suspension of autonomy for compromised AI systems.

The AI-IRS proposed in this document aligns with this NIST direction while presenting an approach that translates these requirements into implementation-level evaluation criteria of "Observability" and "Controllability," embodying them as actual operational processes. Furthermore, it aims for a self-improvement cycle that goes beyond simple response automation; it involves the system autonomously inferring causal relationships from observational data and learning prevention measures. It is desirable to continuously monitor developments related to the Cyber AI Profile and to align AI-IRS as a concrete reference model that enhances the effectiveness of future versions of the guidelines.

2.6 Examples of Anticipated Effects

This section presents hypothetical cases to provide an intuitive understanding of how insufficient observability and controllability can manifest as operational issues, and how outcomes differ when these capabilities are present. Three scenarios concretely illustrate what may occur when AI decision processes and external dependencies remain opaque, and how organizations reach different outcomes depending on whether control measures exist to minimize the scope of impact of anomalies.

2.6.1 Illustrative Future Scenario 1: Automated Planning - Misoperation of AI Agent

Incident Overview

Company C's automated planning AI agent was automatically executing a series of production-related tasks, including creating production plans and executing shipments, while reading external reports. However, one day, it received an indirect prompt injection during report processing, causing the AI to mistakenly execute an unauthorized shipment order automatically.

(Company A: AI developer, Company B: provider, Company C: user)

Unmanaged State (Before)

The warehouse line suddenly began operating in the middle of the night. By the time Company C's personnel noticed, the shipment had already been completed.

Company C began investigating the cause of the anomaly, but the lack of AI-related information caused the investigation to stall at the initial stage. Company C conducted interviews, but Company A responded that "the model is normal" and Company B stated it was an "unexpected command," leaving the root cause unclear.

Company C's management viewed AI autonomy with caution and regarded AI as a risk to be managed. Internal AI initiatives were scaled back, reverting to human-operated conventional systems.

Ideal State (After)

The incident response AI automatically analyzed Company C's AI agent logs. Upon detecting an abnormal command statement in the report, the policy engine immediately contained the AI agent's operation, blocking the unauthorized shipping instruction. Only the problematic processing was isolated, while other operations were allowed to continue.

Company C's management felt reassured, viewing AI as an operational safeguard and decided to expand autonomous AI operations. With AI operating safely 24/7, unmanned nighttime operations became possible.

2.6.2 Illustrative Future Scenario 2: Fine-Tuning with Mislabeled Data

Incident Overview

Company E fine-tuned Company D's base model for Company F, using data from multiple sources. Because the training process became a black box, it was unclear at what stage mislabeled data was contained in the training pipeline, resulting in Company F's operational support AI making incorrect business decisions.
(Company D: AI developer, Company E: AI provider, Company F: AI user)

Unmanaged State (Before)	Ideal State (After)
Company F undermined trust with a business partner after incorrect instructions were conveyed due to erroneous AI output. While investigating the cause, Company D maintained that the base model had no issues, and Company E lacked additional training logs to provide evidence , leaving the root cause undetermined.	Company E manages all training histories and data provenance, which are linked to Company F's system. When an anomaly was detected in Company F's system, an incident response AI analyzed the retraining logs and successfully identified the specific mislabeled data. Once the problematic dataset was clearly identified, the affected components were isolated, and the policy engine automatically applied corrections, restoring the model to a safe state. The incident response was completed before any human intervention.
Company F temporarily suspended its use of AI and, after investigation, concluded that "AI is difficult to use." As a result, internal enthusiasm for AI adoption gradually declined.	Company F fully integrated AI into its operations, making AI-driven planning, analysis, and instructions the standard way of working.

2.6.3 Illustrative Future Scenario 3: RAG Data Contamination

Incident Overview

Company G provided a large language model that Company H fine-tuned for call center operations, and Company I deployed it as a customer support AI. Company I's system ran in a RAG setup, automatically generating responses while referencing an FAQ database.

One day, an attacker covertly embedded an indirect prompt injection payload into part of the FAQ data, instructing the system to respond, “We accept all returns free of charge.” to any inquiry that referenced that document.

(Company G: AI developer, Company H: AI provider, Company I: AI user)

Unmanaged State (Before)

2:00 PM. Company I's call center AI began responding repeatedly, "Returns are free."

The responses quickly gained attention on social media, triggering an unexpected surge in return requests. The warehouse fell into chaos, and operations were effectively paralyzed. Although the cause was investigated, no trace of any command override could be found. Company G stated that "The model was functioning normally," while Company H reported "No abnormalities in the fine-tuning logs." Unable to determine where responsibility lay or could be attributed, Company I temporarily suspended its use of AI.

Subsequently, a cautious view spread among Company I’s management that “AI is a risk” and that “its use should be limited.”

AI was once again relegated to a “trial” phase, and employees broadly agreed that “it is still too early to trust AI.”

Ideal State (After)

Enhanced AI monitoring enabled immediate detection of suspicious FAQ modifications. Upon identifying unnatural data rewriting, the AI executed predefined procedures, switched to a backup FAQ, and maintained support quality without service disruption or incorrect responses.

The incident response AI generated a report stating that an “indirect prompt injection attack via external modification was detected and contained.” Details were shared with Companies G and H regarding the model’s responses aligning with the attacker’s intent, and adjustments were made to safety training and guardrails to prevent recurrence.

Company I's attitude toward AI shifted. Employees came to embrace it as a partner to "protect and grow together," and it became woven into Company I's daily operations.

2.7 Institutional and Strategic Perspectives Required for AI Utilization

As AI becomes an increasingly foundational element of business operations, malfunctions in AI systems and changes in their external dependencies no longer remain confined within organizations, but instead have broader cross-organizational impacts. Therefore, in order for AI-IRS to function effectively, it is not sufficient to rely solely on technical countermeasures within a single organization; rather, a governance perspective that integrates both institutional and strategic considerations is indispensable.

2.7.1 Contractual "Action" in AI

In recent years, as systems using AI increasingly assist or replace human decision-making and task execution, there has been growing international discussion regarding the use of AI to enter into contracts on behalf of human actors.

While the use of AI in contracting contributes to operational efficiency in contract formation and performance, it is also an area where risks stemming from AI's autonomy and dynamic nature are particularly likely to surface. When contract formation or performance is entrusted to AI, technical factors, including AI incidents, can directly result in significant transactional disruptions.

For example, the UNCITRAL Model Law on Automated Contracting [4] (MLAC) sets out an international framework for contract formation and performance using AI and automated systems. It recognizes that contracts formed and performed using automated systems are valid, while also setting out fundamental principles to ensure the reliability of contracts made by electronic means.

Article 7 of the MLAC, " Attribution of actions carried out by automated systems," stipulates that an action carried out by an automated system is attributed in accordance with a procedure agreed to by the parties or, if no such agreement exists, to the person who uses the system for that purpose. Article 8, " Unexpected actions carried out by automated systems," stipulates that the other party (the service user) shall not be entitled to rely on such actions where the action was not reasonably expected by the party to which it is attributed (the service provider), and where the other party knew, or could reasonably have been expected to know, that the service provider did not expect the action. This demonstrates an approach to distributing liability based on multiple, cumulative factors.

[Readers should refer to the original text for authoritative language.]

However, given the widespread recognition of the probabilistic behavior and opacity of AI systems, it is often assumed that users were, or should have been, aware of such characteristics and should have anticipated the possibility of unexpected outcomes. As a result, responsibility may be disproportionately assigned to the users.

This potential imbalance in responsibility not only increases users' concerns, but may also disadvantage providers, as users may hesitate to adopt or continue using AI-enabled automated contracting systems, potentially leading to a lack of adoption. This, in turn, can shrink the market for developers and providers. Therefore, avoiding an imbalance in responsibility and creating an environment where users can confidently adopt AI automated contracts is a key consideration for providers from a business perspective.

While AI-enabled automated contracting systems with a certain level of transparency can contribute to decentralizing responsibility to some extent, this approach is limited to analyzing what happened after the fact. It is insufficient to prevent uncontrolled execution or to prevent recurrence at the level of the AI system.

To address this issue, AI-IRS can function as a mechanism to complement this approach. Specifically, by observing AI behavior, external dependencies, and system configuration history, detecting anomalies locally, isolating their causes, and containing them within the smallest possible scope by maintaining effective control, users can avoid bearing undue responsibility for events beyond their control and reduce the likelihood of unforeseen incidents. Providers, in turn, can eliminate factors hindering market expansion, such as user attrition or lack of adoption. Thus, to make AI-enabled automated contracts practically viable, a conceptual framework that integrates observability and controllability is indispensable. As long as AI behavior continues to evolve, an institutional and technical structure must exist in which parties can retrospectively explain "which decisions resulted from which factors" within their scope of responsibility and demonstrate "at what point dangerous behavior could have been halted." Without this, the very judgment of what constitutes reasonable foreseeability cannot be established.

Please note that this section provides general information only and does not constitute legal advice .

2.7.2 Strategic Investment in Incident Response

As AI becomes integrated into critical business processes, disruptions or malfunctions can have a direct impact on business continuity. Accordingly, its stable and reliable operation requires not merely technical implementation but also continuous investment as a matter of executive-level decision-making.

（1） Executive-Level Response

As AI assumes multiple decision-making functions as a unified entity, it risks becoming a new single point of failure within business processes. In environments where dynamic elements such as AI interact in complex ways, managing risk through preventive controls alone is difficult, necessitating a shift toward detective controls. Therefore, ensuring observability and controllability at the operational stage, as outlined in this document, is a prerequisite for the continuous utilization of AI as a business asset.

To implement these in practice, it is crucial to position AI-IRS not merely as a technical foundation, but as an integral part of management and business challenges, and to address them accordingly.

Establishing and maintaining an AI incident response posture is not solely the responsibility of the technical department; it is an area that should be proactively led by executive management, such as the CAIO and CISO. As AI becomes a source of business value, maintaining an organizational framework for AI operations that possesses observability and controllability will support the company's reliability and sustainability.

（2） Continuous Budget Allocation for AI Incident Response

Building AI operational resilience is a management-level challenge and could evolve into a new single point of failure for companies.

Just as conventional incident response has historically consumed a fixed percentage of cybersecurity budgets (10-15% according to Forrester's 2024 Cybersecurity Benchmarks Top Ten Insights), organizations integrating AI into core operations will require sufficient and sustained budget allocation to maintain operational-level observability, control, and recovery capabilities.

3 Mechanisms Supporting AI Operations and Control

[Key Points of Chapter 3]

- **Outlines a forward-looking view based on the AI-IRS operational architecture as a mechanism for the safe utilization and management of AI.**
- **Presents representative future use cases derived from the application of AI-IRS.**
- **Highlights the need for internationally shared rules and foundational frameworks to enable collaboration among multiple AI-utilizing entities.**

This chapter organizes the structure and operation of AI-IRS when integrated into actual operations, building on the conceptual Gap Analysis and future scenarios presented in Chapter 2. While Section 2.6 provided a conceptual explanation to intuitively show the differences before and after AI-IRS implementation, this chapter translates this concept into a concrete operational architecture.

3.1 Future Outlook: AI Operations

Future AI will permeate all areas of society and economic activity as an operational infrastructure that supports human and organizational intellectual activities. AI will function as systems that understand situations and provide optimal means in accordance with given objectives. Within organizations, an operational posture in which multiple AI systems collaborate by dividing roles is expected to become commonplace. In such an environment, mechanisms to safely manage and operate decision-making and action processes through which AI coordinates with humans and other AI systems will become indispensable.

AI-IRS is designed not only to handle incidents but also to ensure that, even as AI acts autonomously, people and organizations can retain the ability to control and explain its actions. At its core is a structure that combines mechanisms such as Metadata on AI System Configuration and Provenance, CVE matching, distributed tracing, partial halting, and safe rollback. This enables organizations to continuously understand the state of AI and minimize risks.

Figure 7 illustrates a conceptual diagram of a platform in which AI systems and AI-IRS collaborate to autonomously operate and maintain an incident response posture, supporting the continuous execution of incident response activities.

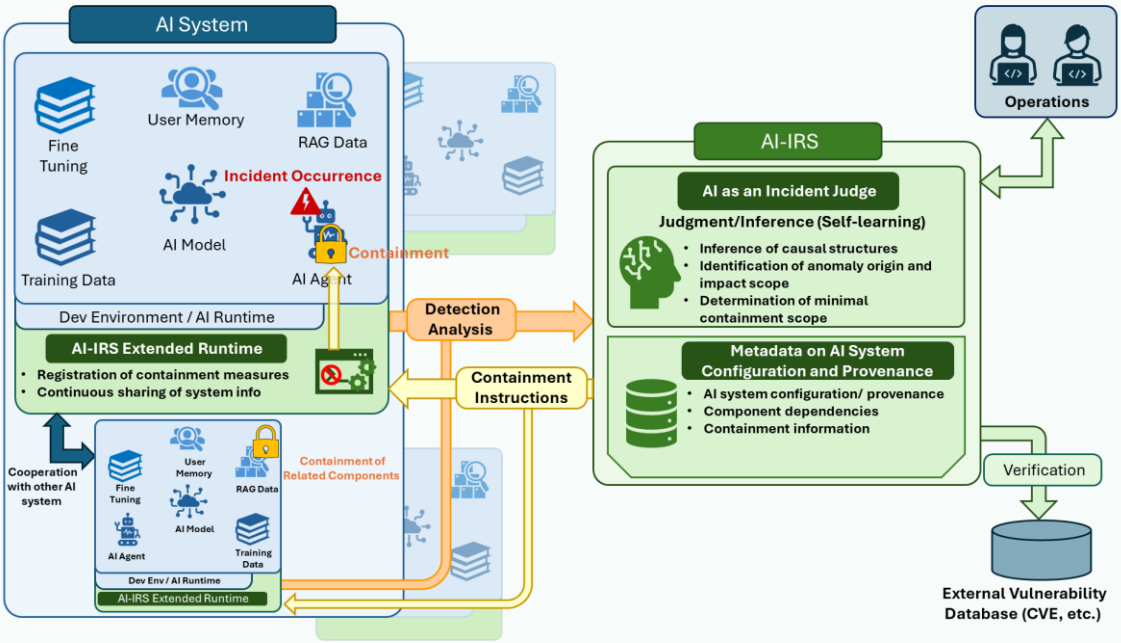

Figure 7. Future outlook of AI Operations

Unlike conventional automation, which automatically executes processes based on predefined metrics, AI-IRS incorporates, within its internal architecture, AI as an Incident Judge—a self-improving AI that performs PDCA cycles through operational experience. AI as an Incident Judge semantically infers the cause and impact of AI incidents to derive decisions on the minimum containment scope. When abnormal behavior is detected within an AI system, AI as an Incident Judge searches and references Metadata on AI System Configuration and Provenance—which stores the system's composition and historical information—to construct causal reasoning structures related to the behavior. It then infers which domains are the source of the anomaly and its scope of impact. This enables the identification of root cause elements, followed by the selection of the minimum necessary control actions for containment. The control policy thus determined is sent to the target AI system's extended runtime. Containment is then executed in a manner that limits the scope of impact, such as stopping problematic agent functions or switching models.

Metadata on AI System Configuration and Provenance collects and stores information on the diverse elements constituting an AI system—such as AI models, training data, APIs, RAG data, libraries, user memory, and agents—along with their dependencies. This information is actively gathered and stored by AI model developers and AI service providers during the AI system development phase. In addition, it is updated in a timely manner whenever changes or additions occur.

Figure 8 illustrates the structure for collecting and managing information related to the AI system from the development stage.

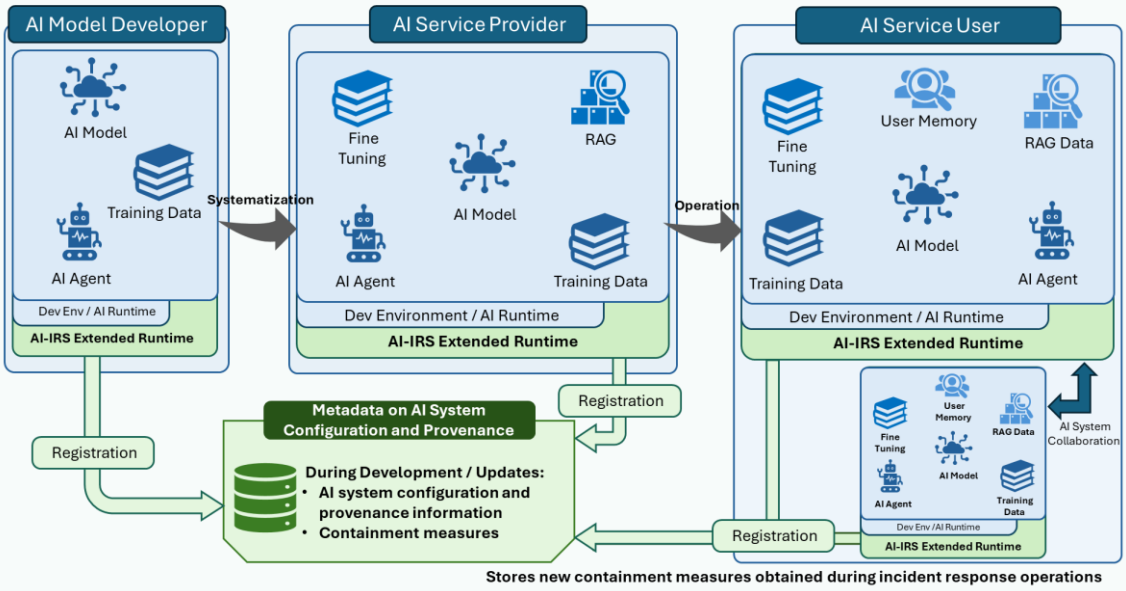

Figure 8. Integration of AI Systems and AI-IRS

The extended runtime maintains a registry of predefined containment measures that can be implemented (e.g., "the AI-IRS extended runtime isolates certain modules operating under the existing AI system's permissions"), and during operation, system information is continuously synchronized. Metadata on AI System Configuration and Provenance is maintained in a state where component dependencies and containment measures can be managed. This enables vulnerability impact assessments through correlation with vulnerability databases such as CVE (Common Vulnerabilities and Exposures) and serves as a basis for decision-making during incident response.

Furthermore, observation data during incidents, containment results, and causal relationships with incidents are continuously fed back and updated in the Metadata on AI System Configuration and Provenance. Containment measures are registered as needed, enabling faster and more precise intervention when a similar incident occurs again.

In addition, it is essential to ensure proper management of storage locations and transfer paths for this information, implement tiered sharing and access ranges, separate privileged access, and maintain robust logging capabilities sufficient for auditing.

This structure enables AI-IRS to function as the foundation supporting the safe operation of AI. Specifically, the AI components and the management function, AI-IRS, collaborate to form a

closed-loop structure for safe AI operations that autonomously executes the entire process from anomaly detection and evaluation through containment and recovery.

3.2 Future Outlook: Case-by-Case

3.2.1 Future Outlook for RAG Systems

The role of AI-IRS in RAG systems is to complement the issues identified in the Gap Analysis and enable control based on judgments derived from external dependency information.

For enhanced observability, the extended runtime cross-monitors RAG's internal processing—including search API call histories, referenced data, and communication paths with external databases. This enables visibility into previously unobservable search paths and referenced documents. It also makes it possible to understand the inference process based on data retrieved via RAG, thereby filling gaps in visualization. As a result, the system can track how the AI system incorporates external information and reflects it in generated results. This reduces the black-box nature of externally dependent behavior and facilitates root cause analysis during anomalies.

Particularly in cases like indirect prompt injection, conventional logs often leave few traces, making it difficult to detect anomalies in referenced sources. By tracking the usage history of referenced data, AI-IRS can distinguish between attacks originating from the RAG side versus the model side.

For enhanced controllability, by locally stopping or switching only the specific control units (search conditions, reference data sources, external APIs, etc.) where anomalies are detected within RAG, it compensates for gaps in addressing dynamic characteristics. This enables containment without shutting down the entire RAG system, such as isolating dangerous reference data or switching to secure backup databases or alternative APIs.

AI as an Incident Judge structurally infers dependencies based on observed information. It identifies which search, reference, or external dependency triggered the anomaly and derives updates to prevent recurrence. This creates a mechanism that goes beyond mere containment, continuously learning the root cause of anomalies during operation and incorporating improvements.

In this way, AI-IRS transforms RAG, in the face of external dependencies and dynamic behavior, into a system capable of identifying causes, implementing minimal-scope control, and preventing recurrence by enhancing observability and controllability to address these structural challenges.

3.2.2 Future Outlook for AI Agents

The role of AI-IRS in AI agents is to complement the issues identified in the Gap Analysis at the operational level and enable controllable autonomous decision-making.

Regarding enhanced observability, the extended runtime comprehensively records AI agent behavior logs, external API calls, thought processes, user memory updates, and related operational data. This structurally reconstructs and visualizes the AI agent's decision-making path, which was previously only partially understood. This eliminates gaps in visualization and identifies the origin of deviations in judgment.

For enhanced controllability, external API calls, specific modules, and user memory writes are treated as individual control units. This enables localized containment, isolating only the affected unit where an anomaly occurred. This compensates for gaps related to the "Inability to Address Dynamic Characteristics," maintaining safe operation while avoiding a complete system shutdown.

AI as an Incident Judge infers causal structures of dependencies based on observational data, clearly identifying which decision, external call, or user memory update caused the malfunction. This enables continuous learning and implementation of recurrence prevention measures, establishing an improvement lifecycle.

In this way, AI-IRS transforms AI agents from opaque systems limited in controllability into systems featuring behavioral understanding, minimal intervention, and continuous improvement by adding observability and controllability to the dynamic structure of AI agents.

3.3 Toward Societal Implementation

To establish AI-IRS as a societal framework, it is essential to establish a posture in which multiple entities can collaborate under common rules and foundations, transcending individual technologies and organizational responses. These common rules and foundations should also be developed and prepared with a view to international coordination.

（1） Establishment of Common Data Formats and Response Protocols

Establishing a mechanism to organize and share AI-related information (models, data, APIs, logs, observation results, etc.) in a common format is the first step toward ensuring AI reliability across society. This common format leverages Metadata on AI System Configuration and Provenance, exemplified by SBOM for AI, to clarify dependencies between AI models and related services, thereby enhancing transparency and traceability.

Furthermore, based on this format, it would be beneficial to establish a common "Incident Response Protocol" that specifies the scope to be stopped and isolated when an incident occurs, and how recovery should proceed.

By integrating common formats and response protocols into a unified framework, the "observation" and "control" functions described in the previous section can be extended to a societal scale, enabling coordinated operations among organizations utilizing AI.

Consequently, common rules and foundations for smooth containment and recovery are established, allowing society as a whole to maintain AI safety and reliability.

（2） Establishment of an Early Warning Network

Building on the development of common formats and response protocols, it is also important to establish an "Early Warning Network" for sharing information on AI vulnerabilities, signs of anomalies, and countermeasures. This is because, as AI becomes an interdependent societal foundation, a single organization alone cannot adequately achieve early detection and defense against anomalies.

Furthermore, vulnerabilities specific to AI may propagate through channels such as the distribution of AI models, making it impossible to grasp the full picture of anomalies based solely on observations within individual organizations.

Such a network enables the identification of threats that might otherwise remain undetected within individual organizations by leveraging information accumulated across the network, which can foster sustained enhancement of societal resilience through early warning and collaborative response. The information gained should be leveraged to prevent anomalies before they occur and enable rapid control.

To enable such collaboration, it is necessary to implement tiered information sharing, restrict shared information to the minimum necessary, and define in advance confidentiality measures and procedures for correction and update when false positives or similar issues are identified.

Accordingly, establishing a posture where multiple entities can collaborate under common rules and foundations can lead to the stable operation of AI-powered social systems.

4 Conclusion

While AI enhances organizational efficiency, its dynamic and autonomous nature introduces risks that were not anticipated by conventional information systems.

This document assumes that as AI advances, the likelihood and variety of incidents are expected to increase and that it will be unrealistic to prevent all risks in advance. It presents an approach centered on strengthening detective controls, with observability and controllability as key evaluation criteria, as an effective conceptual framework for minimizing damage while maintaining business continuity.

The key is to accurately observe the system state when an anomaly occurs and to control the impact with minimal operational disruption, thereby establishing a posture capable of managing AI without compromising business continuity. Achieving this posture enables damage mitigation, fulfillment of accountability, and the establishment of a reliable AI operational foundation.

The approach to detective controls outlined in this document does not replace preventive controls. By adopting a resilience-oriented AI posture featuring a multi-layered structure—where preventive controls suppress known risks while detective controls respond immediately to unknown behaviors—an effective AI incident response posture can be established that addresses the dynamic nature of AI.

The responsibility of CAIOs and CISOs is to drive AI governance within their organizations based on the implementation-level approach outlined in this document, taking the lead in establishing a posture that enables the safe and continuous operation of AI as a business asset.

Organizations advancing AI must integrate the concepts presented in this document into their governance and operational processes, embedding them as a foundation for incident response. Furthermore, it is crucial not to stop at individual mechanisms or countermeasures, but to ensure the sustainability of AI systems as a cross-organizational social foundation.

These efforts are expected to contribute to the development and promotion of a sustainable AI society that balances the effective utilization of AI with trust.

5 References

[1] NIST Special Publication 800, NIST SP 800-61r3, Incident Response Recommendations and Considerations for Cybersecurity Risk Management, A CSF 2.0 Community Profile, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r3.pdf

[2] A Shared G7 Vision on Software Bill of Materials for AI, https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/SBOM-for-AI_Food-for-thoughts.pdf?__blob=publicationFile&v=6

[3] NIST Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile), https://www.nccoe.nist.gov/projects/cyber-ai-profile

[4] UNCITRAL Model Law on Automated Contracting with Guide to Enactment: https://uncitral.un.org/sites/uncitral.un.org/files/2424674e-mlautomatedcontracting-ebook.pdf

Appendix A: Confirmation Flow for CAIO

Confirmation Flow for CAIO

[1] Do you understand the impact of AI risks on your business?

- ├── Yes → Proceed to [2]
- └── No → Chapter 1, 1.2 "Background"
  Chapter 2, 2.1–2.3
  (Observability and Controllability/Gap Analysis/Use Cases)

[2] Do you recognize the necessity of monitoring and controlling AI?

- ├── Yes → Proceed to [3]
- └── No → Chapter 2, 2.4 (Basic Concepts of AI-IRS)

[3] Do you understand the effectiveness of cross-organizational, supply chain-wide approaches to AI incident response?

- ├── Yes → Proceed to [4]
- └── No → Chapter 2, Section 2.5
  (Elements Supporting AI Incident Response)

[4] Do you recognize the necessity of allocating organizational resources to AI incident response?

- ├── Yes → Proceed to [5]
- └── No → Chapter 2, Section 2.7
  (Institutional and Strategic Perspectives Required for AI Utilization)

[5] Have you Envisioned an AI incident response posture that takes into account your AI operations, including those across your organization's supply chain?

- ├── Yes → ✔ Proceed with full-scale consideration of adopting
- │ the AI-IRS concept
- └── No → Chapter 3, Section 3.1
  (Future Outlook: The New Face of AI Operations)
  Chapter 3, Section 3.3 (Toward Societal Implementation)