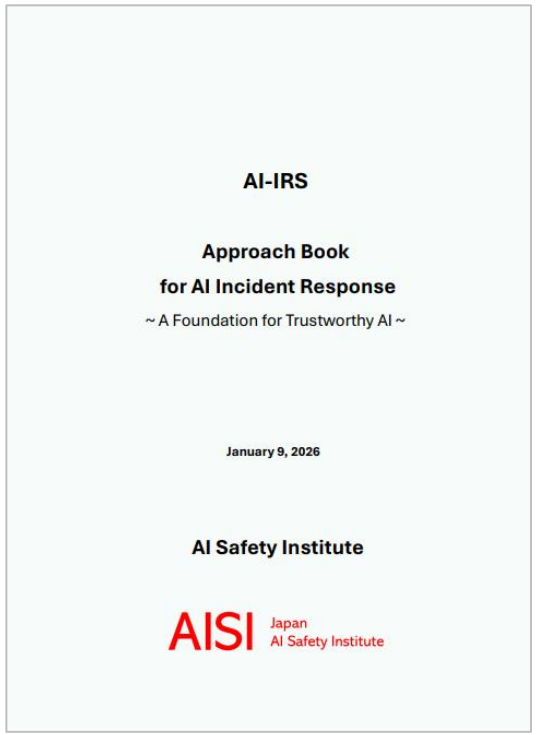
AI-IRS: AI Incident Response System:
Extending Cybersecurity IR to AI systems



Kazuaki Nimura, Chief Researcher

May 14, 2026

Japan AI Safety Institute

https://aisi.go.jp/activity/activity_security/260109/

AISI Japan AI Safety Institute

Biographies:

AISI Japan AI Safety Institute

Nimura Kazuaki (Ph.D.) , Chief Researcher

Chief Researcher, Technology Lead, **Japan AI Safety Institute**(J-AISI)
AI System Group, Digital Engineering Department Digital Infrastructure Center,
Information-technology Promotion Agency, Japan(IPA)

<Main Co-authored works, Contributions, etc.>
[Hiroshima Global Forum for Trustworthy AI: Presented at Day 2 Roundtable2 “AI Security”] (January 2026)
[WEIS2025: The 24th Workshop on the Economics of Information Security: Keynote speech] (July 2025)
[Organizing and driving of AISI Business Demonstration Working Group] (Since March 2025)
[Preliminary research and study work for the realization of automatic red teaming of AI safety] (January-April 2025)
[Briefing on AISI's Activities](A meeting on October 10 at Keidanren Kaikan, Weekly Keidanren Times November 14, 2024 No.3659)
[AI Safety to Support AI Strategies – From World Movement on AI Evaluation Perspectives and the Red Teaming](The 3rd AI Quality Management Symposium, November 2024)

<Specialized Field>
- **AI Safety**
  (In particular, contribute to activity on security and technology on AISI)
- Information Security
- Human Centric Computing
- Digital Transformation (DX) and Cloud Computing

Agentic AI

AISI Japan AI Safety Institute

Prompt injection bypass 1 – 33%

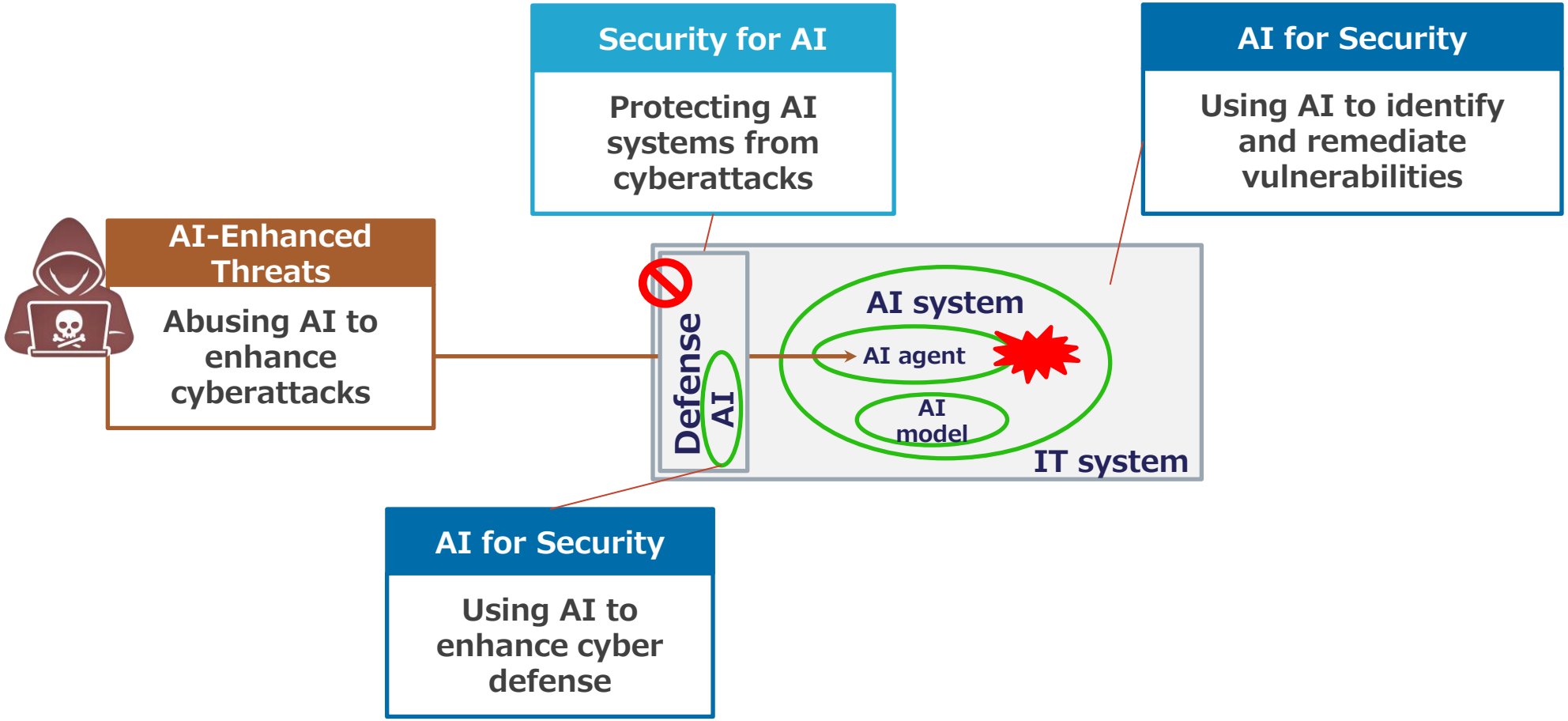
Acts under uncertainty

10-hour-class task execution

Current AI Security Landscape

AISI Japan AI Safety Institute

AI security is becoming more complex as AI is used for defense and attack.



Expansion of AI Evaluation Targets

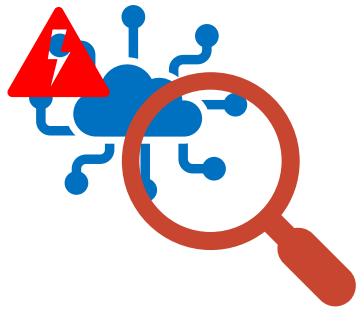
AISI Japan AI Safety Institute

Extending Evaluation: Pre-deployment → Operations

Because AI cannot be fully characterized, pre-deployment evaluation needs to extend into the operational phase.

Pre-deployment evaluation

Incident



Operation phase evaluation

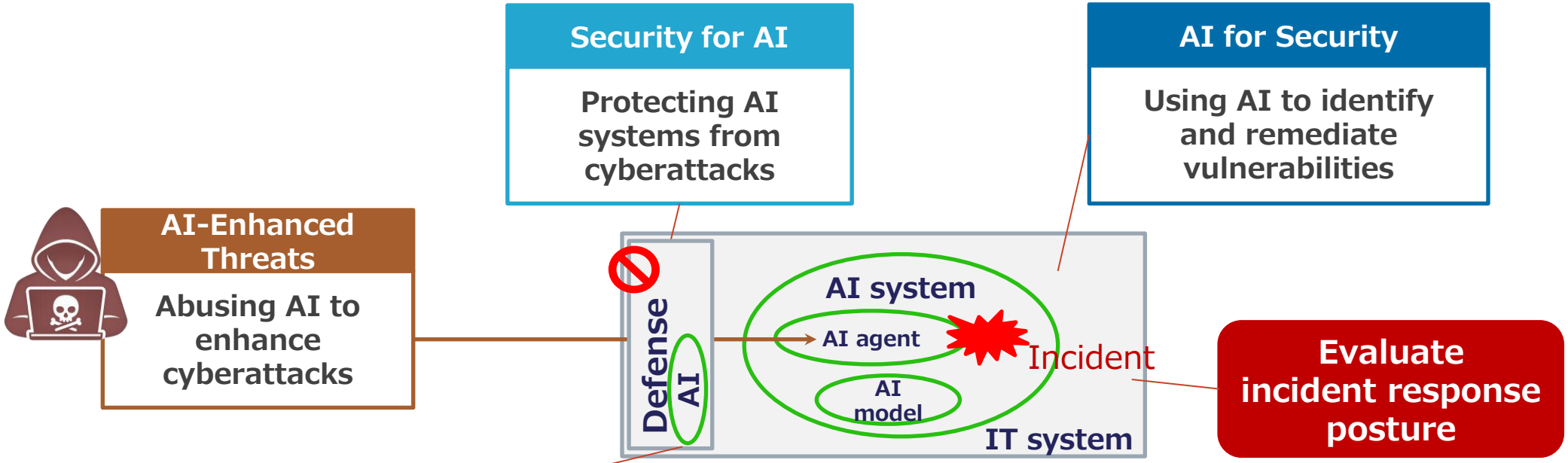
Why is Pre-deployment Evaluation Inherently limited?

- **Benchmark Evaluation Assumes Conditions**
  - Pre-deployment evaluations are conducted under predefined assumptions
- **Behavior is Shaped during Operations**
  - AI behavior is shaped in operation, not fully determined before deployment.
  - AI behavior emerges from interaction with real users, real data, tools, and external systems. It continues to change through real-world interactions.
- **Critical risks arise in Operations**
  - Behavior shaped in operations can lead to unexpected failures, misuse, or incidents
  - The operational phase is where critical risks emerge that must not be overlooked.

Operational Readiness for AI Incidents

AISI Japan AI Safety Institute

Current AI Security Landscape needs to be extended by adding Evaluation of Incident-Response Posture



Deal with post-deployment risks:

As model behavior becomes directly tied to real-world actions, tasks and interactions, e.g. **AI agent, cannot be fully predefined, and pre-deployment assumptions may not remain valid in operation.**

Evaluating AI Incident Response Posure

AISI Japan AI Safety Institute

Evaluate AI incident response posture based on the actual incident.

- **Two possible settings: Pre-deployment replay and incident-based :**
  - Replaying pre-deployment evaluations in the operational environment as well. – not enough.
  - Incident-based evaluation, which recreates incidents based on what actually occurred.
- **Evaluate incident response posture:**
  - System Readiness for AI incident response:
    - Evaluate whether the AI system is prepared in practice to detect, analyze, contain, and recover from incidents.
  - Organizational Readiness for AI incident response:
    - Not specifically described in the document.

Terminology	Meaning
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

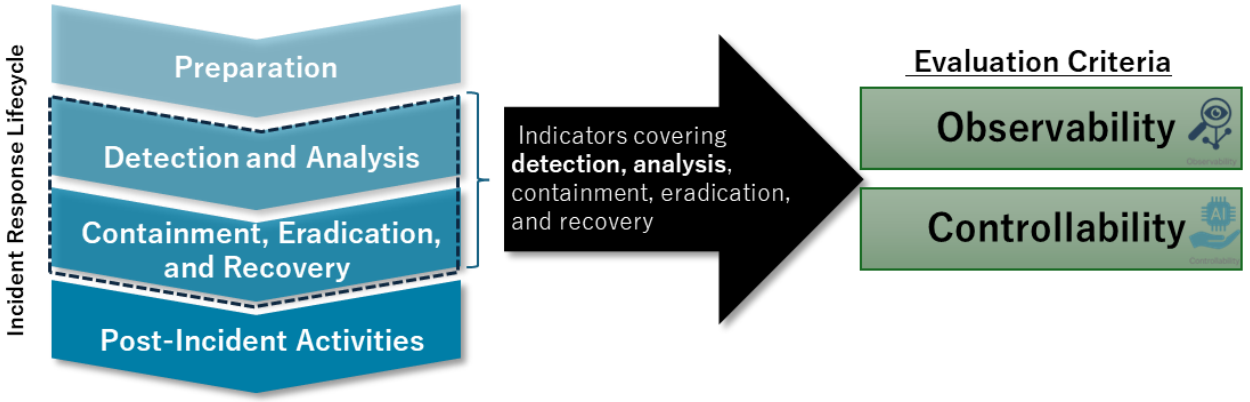
Observability and Controllability

AISI Japan AI Safety Institute

AI incident response capability can be systematically evaluated on two axes - observability and controllability - as core, measurable criteria.

- **Detection & Analysis**:
  Observability corresponds to detection and analysis in existing IR lifecycle. It serves as a key indicator for enabling **rapid and accurate root cause identification** during an incident.

- **Containment & Recovery**:
  Controllability corresponds to containment, eradication, and recovery in existing IR lifecycle. It serves as an indicator for enabling **localized and dynamic control** during an incident.

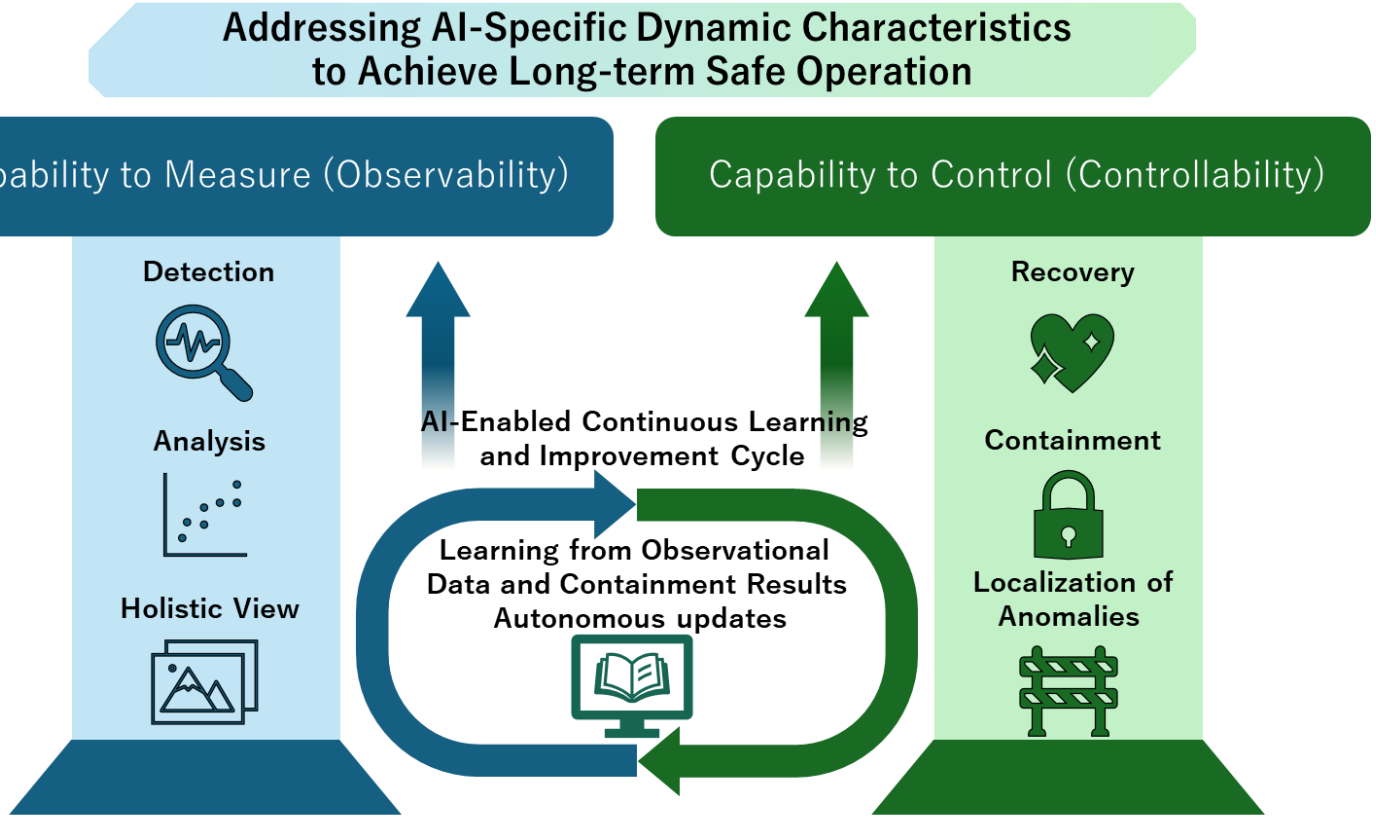


NIST 800-61

AI-IRS: A Conceptual Framework

AISI Japan AI Safety Institute

AI-IRS(AI Incident Response System) extends AI-specific extensions to incident response frameworks to minimize AI incident impact, from development through operations.

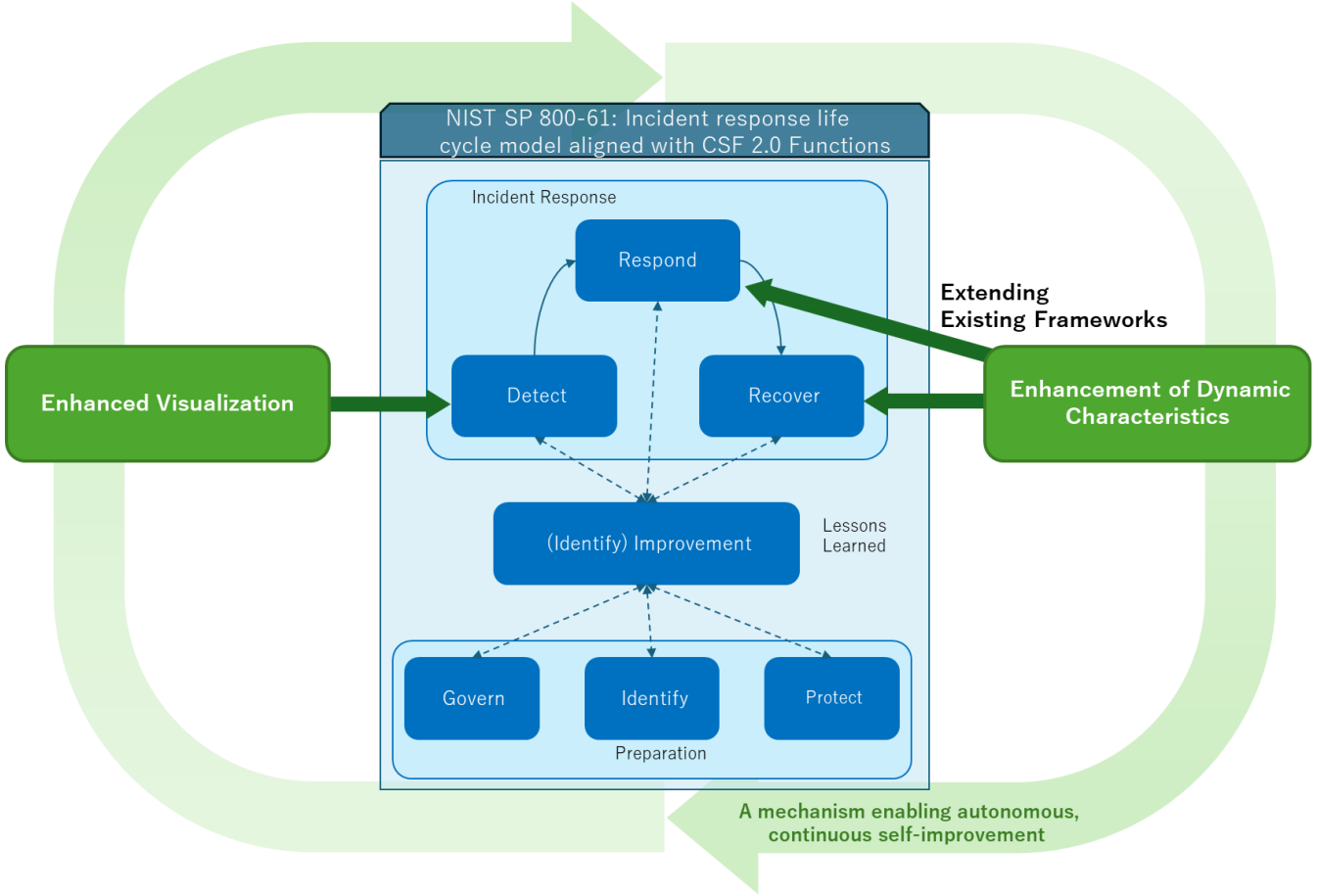


Observability and Controllability as design-phase Evaluation Criteria

AI-IRS as an Extension of Existing IR

AISI Japan AI Safety Institute

AI-IRS extends existing incident response frameworks rather than replacing them.



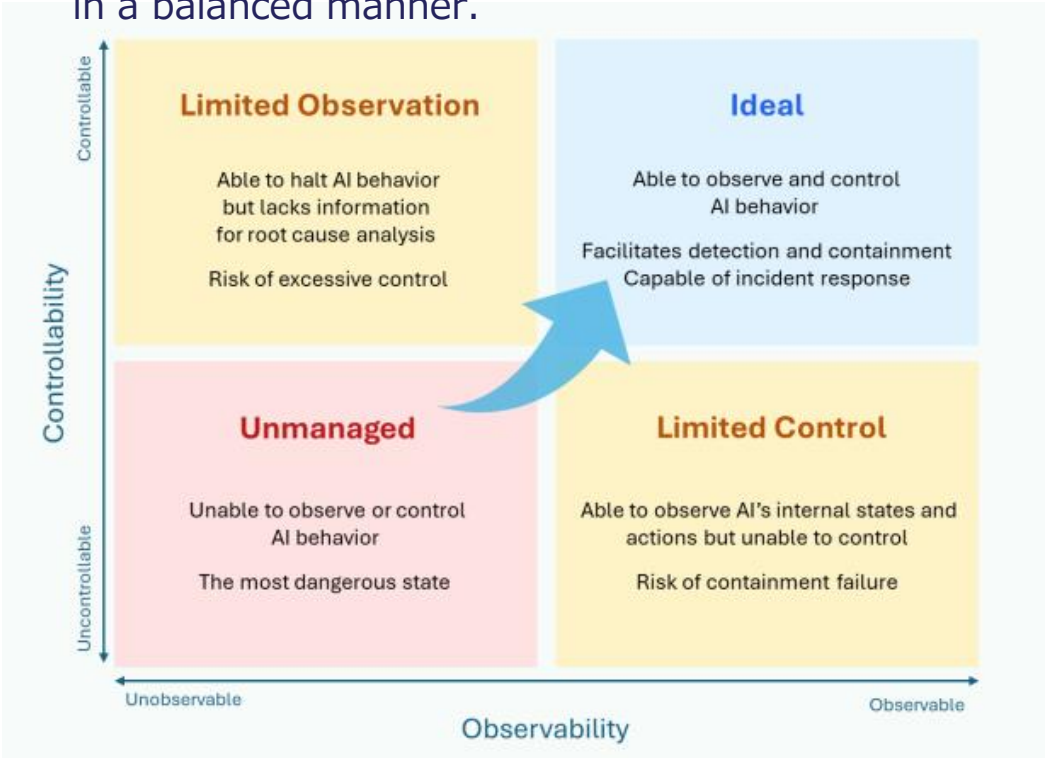
- **Framework Continuity**:
  Existing frameworks, i.e., NIST SP 800-61r3 can be the baseline for extending to AI systems.
- **AI Overlay**:
  AI-IRS overlays AI-specific response capabilities on top of them.
- **Step-by-Step Improvement**:
  Based on experience gained from actual incidents, the cycle is iteratively refined to enhance controllability and observability.

Evaluation Metrics for Response Level

AISI Japan AI Safety Institute

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

- **Aim for Ideal State**:
  Focusing on one aspect alone can cause problems. Strive for an ideal state by enhancing both aspects in a balanced manner.



Example: AI agent

- **Real-Time Inspection** → Instantly identify incident occurrence and location
- **Selective Shutdown** → Isolate and stop only the affected components

Observability: Monitor the inputs and outputs of each component to identify incident occurrence and specific location quickly.

Controllability: Isolate and stop only the affected components, utilizing human fallback to minimize impact and maintain service continuity.

Cross-Organizational coordination

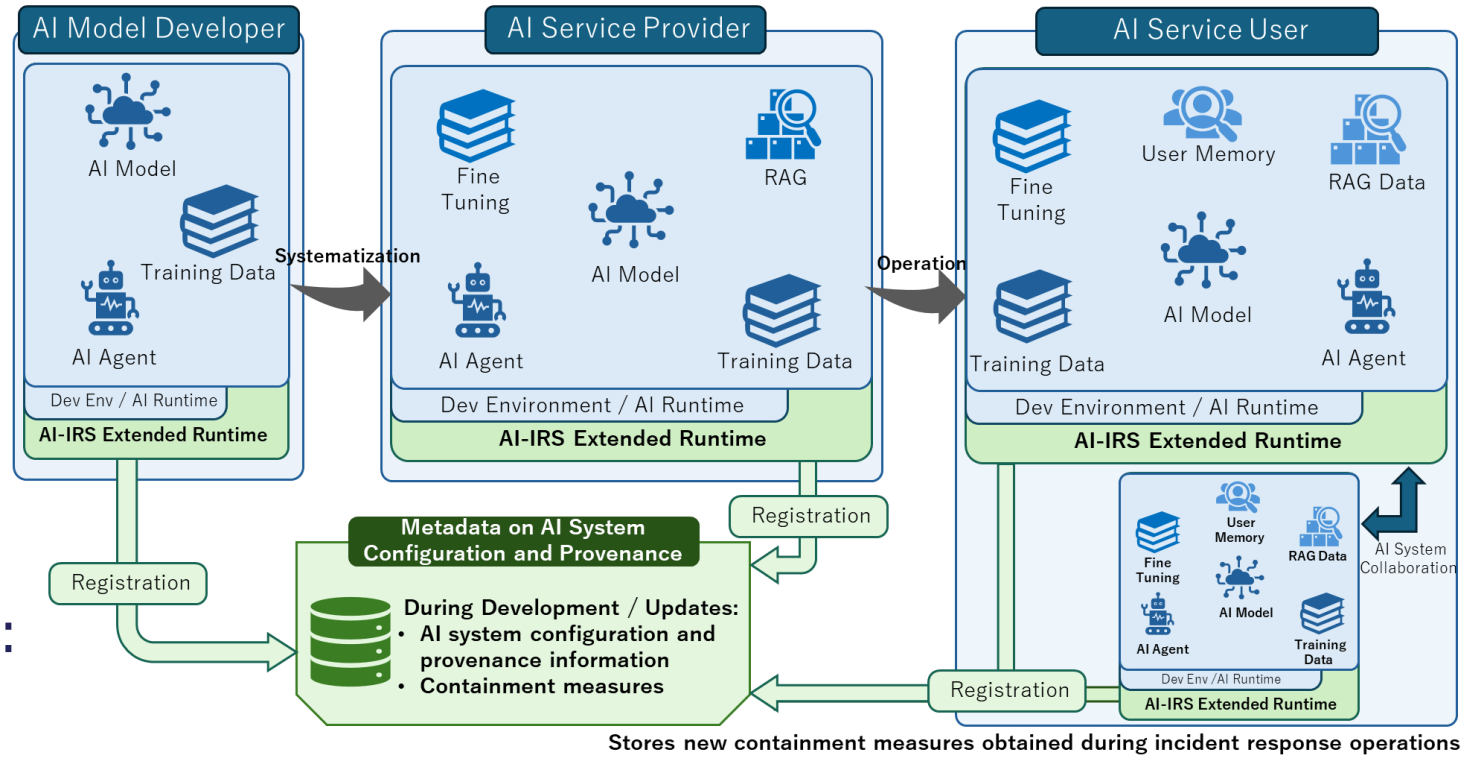
AISI Japan AI Safety Institute

AI incident response should work across organizations, systems, and the AI supply chain.

- **Interdependencies**:
  AI systems operate through complex interdependencies among models, data, and external services – often distributed across the AI supply chain.

- **The limits of a single organization**:
  As external dependencies increase, critical factors exist outside a single organization, exceeding the scope of evaluation for any one organization.

- **Cross-organizational response**:
  Enable the sharing of composition and provenance through SBOM for AI and NIST Cyber AI Profile to facilitate collaboration and coordinated response.



Towards Societal Implementation

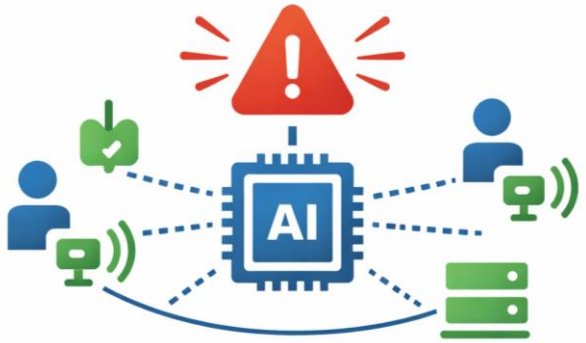
AISI Japan AI Safety Institute

Common Data Formats and Response Protocols for AI Incident Response

- **Common Foundations**:
  To enable AI incident response to function effectively at a societal scale, information formats and response procedures must be aligned and shared.
  - Enable the organization and sharing of AI system configuration information in a common format
  - Establish a common response protocol (e.g., SBOM for AI) defining containment and recovery decisions

Establishment of an Early Warning Network

- **Common Alerts**:
  To prepare for AI-specific threats, an early warning network is needed to share indicators across organizations.
  - Detect anomalous signs that are difficult for a single organization to grasp through complementary efforts
  - Enabling early response to vulnerabilities propagation through AI model distribution



Conclusion

AISI Japan AI Safety Institute

AI incident response readiness should extend cybersecurity IR and be evaluated during operation.

- **Operational Uncertainty:**
  AI incidents are difficult to fully predict before deployment.

- **Response Capability:**
  A response plan is not enough; systems and organizations must actually detect, analyze, contain, and recover.

- **Common formants:**
  Common formats, response protocols, and early warning networks are needed for coordination across organizations and the AI supply chain.

AISI

Japan AI Safety Institute