Check for updates

NIST Internal Report
NIST IR 8621

NIST Workshop on SI Incident Management

M S Raunak
Sanjay (Jay) Rekhi
Alexander Nelson
Anita Rao
Peter Cihon

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8621

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Internal Report
NIST IR 8621

NIST Workshop on SI Incident Management

M S Raunak
Sanjay (Jay) Rekhi
Alexander Nelson
Computer Security Division
Information Technology Laboratory

Anita Rao
Peter Cihon
Center for Advancing Innovation and Standards for Super Intelligence
National Institute of Standards and Technology

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8621

November 2026



U.S. Department of Commerce
Howard Lutnick, Secretary of Commerce

National Institute of Standards and Technology
Arvind Raman, NIST Director and Under Secretary of Commerce for Standards and Technology

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. Many NIST cybersecurity publications, other than the ones noted above, are available at https://csrc.nist.gov/publications.

Authority

This publication has been developed by NIST in accordance with its statutory responsibilities under the Federal Information Security Modernization Act (FISMA) of 2014, 44 U.S.C. § 3551 et seq., Public Law (P.L.) 113-283. NIST is responsible for developing information security standards and guidelines, including minimum requirements for federal information systems, but such standards and guidelines shall not apply to national security systems without the express approval of appropriate federal officials exercising policy authority over such systems. This guideline is consistent with the requirements of the Office of Management and Budget (OMB) Circular A-130.

Nothing in this publication should be taken to contradict the standards and guidelines made mandatory and binding on federal agencies by the Secretary of Commerce under statutory authority. Nor should these guidelines be interpreted as altering or superseding the existing authorities of the Secretary of Commerce, Director of the OMB, or any other federal official. This publication may be used by nongovernmental organizations on a voluntary basis and is not subject to copyright in the United States. Attribution would, however, be appreciated by NIST.

NIST Technical Series Policies

Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History

Approved by the NIST Editorial Review Board on 2026-10-21

How to Cite this NIST Technical Series Publication

Raunak M S, Rekhi J, Nelson A, Rao A, Cihon P (2026) NIST Workshop on SI Incident Management. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Internal Report (IR) NIST IR 8621. https://doi.org/10.6028/NIST.IR.8621

Author ORCID iDs

M S Raunak: 0000-0003-4934-7331
Sanjay Rekhi: 0009-0008-8711-4030
Alexander Nelson: 0000-0002-3771-570X
Anita Rao: 0009-0009-6459-6425

Contact Information
si-incident-management@nist.gov

National Institute of Standards and Technology
Attn: Computer Security Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8930) Gaithersburg, MD 20899-8930

Additional Information
Additional information about this publication is available at https://csrc.nist.gov/pubs/ir/8621/final, including related content, potential updates, and document history.

All comments are subject to release under the Freedom of Information Act (FOIA).

Abstract

The NIST Workshop on SI Incident Management convened over 400 experts across government, frontier-model developers, cloud and cybersecurity providers, academia, standards organizations, and civil society to examine how organizations can prepare for, detect, handle, and recover from SI-related incidents. Motivated by federal directives, the workshop addressed the need for standards, response frameworks, and technical incident response capabilities as SI systems expand across critical infrastructure. To guide these efforts, participants framed SI incident management across two interrelated domains: cybersecurity attacks that target SI systems and incidents that stem from SI misuse, anomalous behavior, or severe operational malfunctions resulting in harm. Discussions and technical presentations highlighted the limitations of traditional, deterministic incident response life cycles when applied to probabilistic, autonomous, and deeply integrated SI systems. Participants emphasized critical operational needs, including reasoning-layer telemetry, non-human identity governance, authorization-scoped containment, clear evidentiary and reporting schemas, and coordinated responsibility across distributed SI value chains. This report synthesizes the core findings, practitioner insights, and actionable recommendations from the workshop to establish a strategic foundation for future NIST guidelines, operational playbooks, and updates to federal incident response frameworks.

Keywords

cybersecurity incident; incident management; incident response; SI attack; SI incidents.

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in federal information systems.

Note to Readers

Per the Sept. 29, 2026, Executive Order 14434: Inaugurating the Era of Super Intelligence, NIST is working to update its communications to incorporate the term "super intelligence" as directed. Section headers represent the titles of the talks used by the presenters.

Acknowledgments

The authors express their gratitude to Maia Hamin and Andrew Fasano from the Center for Advancing Innovations and Standards for Super Intelligence (CAISSI) who supported the planning and execution of the workshop. The NIST ITL and CAISSI team continues to collaborate actively to advance the strategic directions, technical roadmaps, and key outcomes resulting from the workshop.

Special thanks and appreciation go to Jon Boyens, Division Chief of the Computer Security Division, and Craig Schlenoff, Division Chief of the AI Research, Measurements, and Standards Division, for their full support in making the workshop a reality. The authors also gratefully acknowledge Kevin Stine, Director of ITL, and Jim St. Pierre, Deputy Director of ITL, for their steadfast support, strategic leadership, and direction throughout this initiative. The authors are also grateful to the reviewers of the document, including Harold Booth, Anna Sberegaeva, Jon Boyens, Yang Guo, Mark Pryzbocki, and Jim St. Pierre.

The authors and organizers of the workshop would also like to thank the conference program and audiovisual service group, especially Pauline Truong, for their dedicated and professional support in planning and executing this hybrid workshop. We are also thankful to Francis Durso and Mark Paes for helping by taking notes at the workshop.

Table of Contents

1. Introduction 1

2. Workshop Purpose, Scope, and Method 2

2.1. Purpose 2

2.2. Scope 2

2.3. Workshop Program and Participation 2

3. Keynote Address — "A Global Perspective on AI Incident Management: Risks, Responses, and Gaps," by Prof. Neil Thompson (MIT FutureTech) 5

4. Session — "AI Under Attack: Incident Response for AI Systems" 7

4.1. AI Under Attack: Where Your IR Life Cycle Cracks for Agentic Systems, by Kyriakos (Rock) Lambros (Zenity) 7

4.2. Investigating Artificial Intelligence Attacks: A Framework, by James Perry (CrowdStrike) 8

4.3. Incident Response Preparation for AI Systems, by Robert Saul 10

5. Session — Misuse and Malfunction: Understanding Emerging AI-Induced Incidents 12

5.1. AI Incident Database: The Next 6,173 AI Incident Reports, by Sean McGregor (AIID) 12

5.2. Information Sharing, Incident Reporting, and Incident Response for Frontier AI Risks, Chris Meserole (Frontier Model Forum) 13

5.3. Expect the Unexpected: Incident Response Strategies for GenAI, by Jolanda Kumakaw (Google) 14

6. Mini-Talks: Landscape of Existing Guidance 17

6.1. NIST SP 800-61 Revision 3, by Alex Nelson 18

6.2. Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile), by Katerina Megas (NIST) 18

6.3. IQT AI Incident Response Guidebook, by Andrea Brennen 19

6.4. MITRE ATLAS: Mapping the Adversary, by Keith Manville 20

6.5. Japan AI Safety Institute's AI Incident Response System, by Kazuaki Nimura (J-AISI) 20

6.6. ISO/IEC 25870: Data Elements for Reporting AI System Incidents, by Mike Garris 21

6.7. Exploring the AI Incident Documentation Practice, by Violet Turri (Carnegie Mellon University Software Engineering Institute) 22

7. Panel Discussion — Incident Response Roles and Responsibilities in the Era of AI Agents 23

8. Breakout Session Synthesis 25

8.1. Track 1: AI Systems as the Target of Cybersecurity Incidents 25

8.2. Track 2: Misuse, Malfunction, and Harm-Based Incidents 26

8.3. Cross-Breakout Themes 26

9. Conclusion 29

10. Next Steps 30

List of Tables

Table 1. Workshop Agenda.....3

Table 2. List of mini-talks covering some of the existing resources.....17

1. Introduction

Super intelligence (SI)[1] is rapidly transforming the foundation of modern infrastructure and driving unprecedented advancements across public and private sectors. As SI systems become increasingly integrated into critical operations, they introduce a complex landscape of potential vulnerabilities, both as a target and as a source or facilitator. Unlike traditional deterministic IT systems, SI models are probabilistic, operate at autonomous machine-speed that vastly outpace human reaction times, and rely on complex development and algorithmic process. This new construction model means that SI systems are not only targets for new types of cyber threats (e.g., data poisoning, prompt injection, model inversion) but can also act as a source of harm through adversarial and non-adversarial malfunctions, hallucinations, or malicious misuse. Consequently, traditional incident response life cycles built for conventional software systems may fracture or be incomplete when applied to SI, making the updating of incident response in the presence of SI a priority.

[1] Per the Sept. 29, 2026, Executive Order 14434: Inaugurating the Era of Super Intelligence, NIST is working to update its communications to incorporate the term “super intelligence” or SI as directed.

2. Workshop Purpose, Scope, and Method

2.1. Purpose

The NIST Workshop on SI Incident Management was initiated following the federal AI Action Plan's directive for NIST and the Center for Advancing Innovation and Standards for Super Intelligence (CAISSI) to collaborate with SI and cybersecurity stakeholders on developing standards, response strategies, best practices, and technical tools for SI incident management. The organizers described the event as the start of an ongoing process to evaluate existing practices, identify gaps, create a common vision for effective responses, and inform future updates and new resources to be produced by NIST.

2.2. Scope

This workshop was organized around incidents grouped broadly into two overlapping categories: cybersecurity attacks that target SI systems and incidents stemming from anomalous or unexpected behavior resulting from misuse or malfunction. Both require distinct response strategies, governance, and technical guidance:

1. **Cybersecurity incidents that impact SI systems:** Incidents that jeopardize the confidentiality, integrity, or availability of SI systems (e.g., SI models, agents, data, or supporting infrastructure), such as model theft, data poisoning, prompt injection, model inversion, supply-chain attacks, compromised agents, identity misuse, attacks on RAG pipelines, and others.
2. **Incidents that can be caused by SI systems:** Cases of misuse, malfunction, or high-stakes accidents involving SI systems leading to harm, such as unsafe outputs, hallucinations with operational consequences, malicious use leading to cybersecurity attacks, overreliance, harmful autonomous actions, misinformation, and impacts on financial systems or public safety.

2.3. Workshop Program and Participation

The workshop combined formal presentations, a keynote, a session on attacks against SI systems, a session on misuse and malfunction, a survey of some existing guidelines, a panel on SI-agent roles and responsibilities, and concurrent breakout sessions with the attendees. The opening remarks reported participation by more than 400 in-person and online attendees, reflecting interest across government, industry, academia, civil society, and the cybersecurity and SI communities.

Table 1. Workshop Agenda

Program element	Speakers/leads
Opening and scene-setting	Kevin Stine Sanjay (Jay) Rekhi M S Raunak Peter Cihon
Keynote	Neil Thompson, MIT FutureTech
Session 1: AI Under Attack Session Chair: Craig Schlenoff (NIST)	Kyriakos Lambros (Zenity) James Perry (CrowdStrike) Robert Saul (AWS)
Session 2: Misuse and Malfunction Session Chair: Jon Boyens (NIST)	Sean McGregor (AI Incident Database) Chris Meserole (Frontier Model Forum) Jolanda Kumakaw (Google)
Mini Talks: Existing Guidance Session Chair: Harold Booth (NIST)	Alex Nelson (NIST) Katerina Megas (NIST) Andrea Brennen (IQT) Kazuaki Nimura (J-AISI) Michael Garris (GBS) Keith Manville (MITRE) Violet Turri (CMU)
Panel: Roles and Responsibilities Moderator: Vicky Pillitteri (NIST)	Madhu Srikumar (Partnership on AI) Mikel Rodreiquez (MITRE) Jim Reavis (Cloud Security Alliance) Apostol Vassilev (NIST CSD) Ian Reynolds (Hugging Face)
Breakouts and readouts Moderators: Alexander Nelson, Suzanne Lightman, Peter Cihon, Anita Rao	Facilitated breakouts in-person and online End-of-day readout

To capture the full spectrum of the SI life cycle, the workshop brought together a cross-sector coalition of leaders across government, industry, and academia:

- **Academia:** The event featured leading researchers from institutions like MIT, CMU, and the University of Maryland. Dr. Neil Thompson, Director of MIT FutureTech, delivered the keynote address, establishing the framework for prioritizing severe SI risks.

- **Industry Leaders:** Subject-matter experts and cybersecurity practitioners from major tech firms, including AWS, Google, Microsoft, CrowdStrike, and Zenity, shared front-line insights into defending SI assets. Core session speakers included technical leaders like Kyriakos Lambros, James Perry, Robert Saul, Chris Meserole, and Jolanda Kumakaw.

- **Government:** Representatives from different government agencies participated in the workshop and provided valuable input.

Sections 3 through 8 include summaries of presenter materials, discussions, and terminology from the workshop on May 14, 2026.

3. Keynote Address — "A Global Perspective on AI Incident Management: Risks, Responses, and Gaps," by Prof. Neil Thompson (MIT FutureTech)

In his keynote address, Dr. Neil Thompson framed AI incident management as a problem of risk prioritization, evidence, organizational preparedness, and responsibility across the AI value chain. His keynote focused on three questions: which AI risks matter most, how companies are responding, and where major management gaps remain. He introduced the MIT AI Risk Initiative, including the AI Risk Repository, AI Risk Index, AI Mitigations Repository, AI Incident Tracking, and AI Governance Mapping. These resources are intended to help organizations identify AI risks, prioritize them, understand available mitigations, track real-world harms, and map relevant governance mechanisms.

Dr. Thompson discussed recent research from the project, which convened 272 professional and academic experts from 37 countries in a three-round survey to identify areas of agreement. These experts assessed AI risk by rating their concern, estimating probabilities for different levels of harm, identifying vulnerable actors and sectors, and assigning responsibility for mitigation. The results of the survey suggest that the most severe predicted AI risks over the next five years will be AI used for cyber attacks, weapons development, and mass harm; AI systems with dangerous capabilities; and competitive dynamics that may cause companies to compromise on safety and responsibility.

The expert assessment found substantial probabilities of severe or catastrophic outcomes, even when pragmatic mitigations were assumed. Based on current trajectories, experts assess that 18 of the 24 risks are more than 10 % likely to cause catastrophic outcomes (e.g., over 1 million deaths, more than $100 billion in financial losses) over the next five years. Even if pragmatic mitigations are adopted, experts still predict that five of the 24 risks are more than 10 % likely to cause catastrophic outcomes. These include dangerous AI capabilities (e.g., systems that can deceive, manipulate, or act autonomously in harmful ways), AI-enabled cyber attacks, environmental harm, inequality, unemployment, and power centralization (i.e., control over advanced AI becomes concentrated among a few firms or governments).

Experts also emphasized a mismatch between responsibility and vulnerability. Actors (e.g., general-purpose model developers, specialized developers, deployers, governance bodies) may have the greatest ability to reduce risk, while users and affected stakeholders may be the most vulnerable but have the least control. This gap creates a need for standards, contracts, reporting mechanisms, and coordinated governance across the AI value chain.

The presentation also reviewed how companies are responding to AI risks. MIT conducted a preliminary review of more than 550 public AI-related documents from 200 large market-cap global companies. The analysis identified apparent gaps between sectors' assessed vulnerability and public evidence of company responses. Dr. Thompson emphasized that companies may have relevant internal practices that are not publicly documented and invited companies to validate the findings.

Call to Action

- Define AI incidents, hazards, and near misses so that organizations can distinguish reportable incidents from routine model-performance issues.

- Improve AI incident reporting beyond voluntary disclosures and media reports while protecting sensitive information.
- Prioritize high-severity risks, especially AI-enabled cyber-attacks, weapons development, dangerous capabilities, and unsafe competitive dynamics.
- Clarify responsibility across the AI value chain, including infrastructure providers, model developers, deployers, users, affected stakeholders, and governance actors.
- Address the mismatch between responsibility and vulnerability through standards, contracts, governance mechanisms, and information-sharing structures.
- Develop risk-specific mitigation guidelines that help organizations select practical controls based on sector, risk type, cost, and feasibility.
- Prepare for serious failures and misuse with a special focus on agentic systems as well as multi-party incidents that require coordinated responses.
- Support joint investigations and information sharing when incidents span multiple actors across the AI ecosystem.

4. Session — "AI Under Attack: Incident Response for AI Systems"

This session was introduced by Craig Schlenoff and focused on the cybersecurity-centric aspects of AI incidents and the technical realities of defending AI assets as operational infrastructure, spanning training pipelines, models, inference endpoints, vector stores, agent toolchains, and downstream integrations. The session examined how traditional incident response assumptions are strained by AI-specific failure modes, including model inversion, data poisoning, adversarial prompt injection, model drift, and agentic systems whose containment depends on authorization scope rather than simple network segmentation.

4.1. AI Under Attack: Where Your IR Life Cycle Cracks for Agentic Systems, by Kyriakos (Rock) Lambros (Zenity)

Kyriakos (Rock) Lambros from Zenity spoke about severe structural "cracks" in the foundational incident response (IR) life cycle that could be exacerbated by the unique pressures of agentic systems. He warned the audience against two dangerous extremes: treating these systems as "just IT," which ignores these fundamental cracks and leads to playbook failures, and assuming that AI requires a brand new model, which unnecessarily discards working methodologies. Instead, he urged practitioners to find the specific pressure points in their existing blueprints and patch them.

Lambros organized his fracture points around the four phases of incident response: *Detection*, *Containment*, *Recovery*, and *Post-Incident Analysis*. The first major fracture point Lambros identified is in the *Detection* phase, caused by the fact that the malicious "signal has moved up the stack." Traditional security telemetry (e.g., network logs, endpoint agents, identity tracking) is largely blind to attacks like prompt injections, training-time data poisoning, and slow goal drift because they often masquerade as perfectly valid system interactions. Using the 2025 "EchoLeak" case as an anchor, he explained that harm now happens within the model's reasoning layer long before any bad output triggers classical alerts. To mend this crack, he stressed the absolute necessity of implementing reasoning-layer telemetry.

The second and third fracture points disrupt the *Containment* and *Recovery* phases. Lambros challenged the traditional IR assumption that compromised assets can be safely hard-isolated or unplugged. Because AI agents are often deeply integrated into critical workflows, their potential effects are defined by their authorization scope rather than a simple network segment. Referencing a 2025 Replit Agent incident, he explained that containment must now rely on "time-to-revoke" primitives and "soft isolation." Furthermore, recovery is complicated by the fact that surgically removing poisoned data (i.e., machine unlearning) is incredibly difficult, and some AI vulnerabilities simply cannot be patched in the traditional sense, a reality highlighted by the 2018 Amazon Recruiting and 2025 Anthropic GTG-1002 cases.

Finally, Lambros addressed the *Post-Incident Analysis* phase, which is heavily fractured by what he termed the "reproducibility envelope." Unlike traditional malware, AI models are non-deterministic; a prompt that triggers a data leak today might yield a completely benign response tomorrow. This unpinned state makes investigating an attack impossible unless specific preparation work is done beforehand. To operationalize these fixes, Lambros left the

audience with a strict 90-day action plan: within the first 30 days, organizations must add reasoning-layer logging to every AI deployment runbook; by day 60, they must measure and aggressively cut their time-to-revoke for agent authorizations; and by day 90, they must build comprehensive "soft isolation" playbooks to contain threats without breaking deeply integrated business systems.

Call to Action

- **Define reasoning-layer telemetry for AI incident response.** Traditional network, endpoint, and identity telemetry is insufficient when the signal moves into the model or agent reasoning layer. NIST guidelines should define AI-specific forensic telemetry categories, including prompts, completions, intermediate reasoning or trace data, tool calls, authorization decisions, model/version information, and relevant agent states.
- **Shift containment from host isolation to authorization-scoped revocation.** For agentic systems, the area of effect (sometimes called the "blast radius" by practitioners) is determined less by network location and more by the agent's inherited authorization scope and downstream tools. NIST guidelines should recommend containment practices based on per-action authorization binding, capability scoping, tested kill-switch APIs, and measurable time-to-revoke for agent credentials and permissions.
- **Require the recovered system to produce both technical and governance evidence.** AI recovery may require more than patching or rollbacks. Guidance should be developed to describe recovery outputs as both a remediated technical artifact and an updated governance package that demonstrates why the recovered system should be trusted after the incident.
- **Define a reproducibility envelope.** AI systems can be stochastic, vendor-updated, tool-connected, and difficult to reproduce after the fact. A reproducibility-envelope specification must be defined using model versions, prompt sets, tool states, dependent service versions, seeds, retrieved context, logs, and relevant environment variables. This preparation work needs to be done before an incident, not only during post-incident analysis.
- **Practical readiness benchmark.** Organizations should add reasoning-layer logging to AI deployment runbooks, measure and cut time-to-revoke for agent authorizations, and capture reproducibility-envelope specifications for production AI systems before they are needed.

4.2. Investigating Artificial Intelligence Attacks: A Framework, by James Perry (CrowdStrike)

James Perry, Global Vice President and Head of Incident Response at CrowdStrike, focused on the forensic challenge of investigating attacks against AI systems. His central message was that AI incident response will still need to answer the basic questions of who, what, when, where, and why. However, AI systems complicate each of these questions because the evidence needed to reconstruct actor behavior, data access, identity use, model behavior, and

downstream system interactions is often missing, incomplete, ephemeral, or too voluminous to process in time.

Perry emphasized that traditional incident responders know how to investigate conventional endpoints, servers, cloud systems, and SaaS environments because they understand where the relevant logs, artifacts, and forensic evidence reside. By contrast, AI systems often do not preserve comparable evidence. When models are tampered with or AI systems are compromised, responders cannot simply image a disk, preserve the state, and reproduce the incident later. Hosted LLMs, self-hosted inference systems, agentic frameworks, Model Context Protocol (MCP) servers, orchestration layers, and downstream APIs all create different forensic visibility problems.

A major theme of Perry's talk was the lack of reliable AI-native logging. He noted that many LLM deployments do not log prompts, responses, agent actions, the identities used to access data stores, or the specific data accessed. Without those records, investigators may be forced to operate using educated guesses rather than facts. Perry also highlighted the practical challenge of scale: even when logs exist, they may be so voluminous that organizations cannot store, retrieve, or analyze them quickly enough to meet legal, regulatory, and operational requirements.

Perry urged organizations and NIST to think in terms of a common investigative data model for AI. The required evidence should span hosted LLM platforms, self-hosted inference servers, frontier-model services, agentic frameworks, MCP servers, harnesses, and the connective tissue between them. He stressed that this data needs to be available in a common and machine-understandable format because investigations now need to operate at machine speed.

He closed by returning to the forensic questions that an AI investigation must answer: which identity, API key, or system component performed the action; what data was accessed; when the data was collected, trained on, retrieved, or returned; where the data went; and why the actor or system behaved as it did. He also identified several immediate practices, such as logging and retaining prompt and completion content by default where permissible, preserving the forensic traceability of collected data, avoiding the automatic destruction of ephemeral systems before evidence is captured, and standardizing timestamps so that investigative data can be correlated reliably across systems. Perry specifically identified timestamp standardization as a major practical issue and a concrete area where NIST could help.

Call to Action

- **Define AI forensic telemetry requirements.** NIST guidelines should identify the minimum evidence needed to investigate AI attacks, including prompts and completions where permissible, agent actions, the identities and API keys used, the data stores accessed, retrieved content metadata, model and data version information, tool calls, MCP/harness activity, invocation logs, IP addresses, activity logs, and relevant compliance APIs.
- **Standardize a machine-readable AI investigation record.** Perry's strongest NIST-facing recommendation was to establish a common or machine-understandable format for AI

investigative data across hosted LLMs, self-hosted inference servers, agentic frameworks, and AI orchestration layers.

- **Preserve prompt and completion content where legally and operationally feasible.** In the Q&A, Perry recommended 30 days as a baseline retention period for prompt/completion content with longer retention when cost, complexity, and regulatory constraints allow.
- **Require forensic traceability of collected AI evidence.** Organizations should know where evidence came from, who or what collected it, and whether it was collected by a human, agent, system process, or other mechanism. This is necessary for later legal, regulatory, or litigation-driven reviews.
- **Standardize timestamps across AI incident evidence.** Perry made this one of his clearest practical asks for NIST since inconsistent time zones and timestamp formats consume excessive investigative effort, and AI-scale investigations require consistent, machine-readable timestamping.
- **Account for ephemeral and distributed AI infrastructure.** AI incident guidelines should warn organizations not to destroy or auto-delete ephemeral systems, containers, agents, or cloud resources before evidence is captured. The framework should address how to preserve state and investigative artifacts from systems that may spin up, act, and disappear within seconds.

4.3. Incident Response Preparation for AI Systems, by Robert Saul

Robert Saul, General Manager of the Amazon Web Services (AWS) Customer Incident Response Team, focused on incident response preparation for AI systems. His message was pragmatic: AI does not require abandoning the classical incident response discipline, but it does require organizations to prepare differently because AI systems introduce new assets, identities, evidence sources, business dependencies, and containment trade-offs.

Saul argued that the best-performing incident response teams are those that have done the governance work before the incident. He identified three areas that separate good response from great response: non-human identity governance, pre-positioned telemetry, and containment decision frameworks.

First, Saul emphasized non-human identity governance. AI systems create and depend on service accounts, Application Programming Interface (API) keys, authentication tokens, pipeline credentials, orchestration secrets, and agent-to-agent trust relationships. These identities often do not appear in traditional IR asset inventories. When compromised, response teams may not know what the identity does, what permissions it has, how to revoke it, or what business function will break when the trust layer is removed. Saul argued in favor of maintaining a living inventory of AI-created non-human identities mapped to business functions, permissions, behavior baselines, owners, and tested revocation playbooks.

Second, Saul emphasized pre-positioned telemetry. This includes prompt/completion logs, model version history, training-data access, embedding queries, agent decision traces, and drift

metrics. Saul stressed that logs should not merely be collected as technical artifacts; they should be mapped to investigation questions and the business functions they protect so that analysts can understand what they are defending.

Third, Saul emphasized containment decision frameworks. In AI systems, containment itself can cause damage. Killing an agent can corrupt downstream states; rolling back a model can break dependent services; and isolating a vector store can take down production. Saul recommended pre-built containment decision trees for AI components that are anchored in business functions. These decision trees should define available containment options (e.g., revoke, rollback, isolate, disable, fallback), cascading business effects, and approval authorities based on business impact.

Saul's practical call to action was to start with one AI system and build its governance record by cataloging its non-human identities, mapping its data flows end to end, documenting its failure modes, and building one containment decision tree. His core message was that teams that respond well are not simply faster; they did the governance work when it was still boring.

Call to Action

- **Create AI system governance records.** New guidelines should recommend that organizations maintain a living map for each significant AI system, including identities, data flows, failure modes, ownership, and business dependencies.
- **Inventory non-human identities as first-class IR assets.** Guidance should treat service accounts, API keys, OAuth tokens, agent credentials, inter-service trust chains, pipeline credentials, and orchestration secrets as core assets for incident response preparation.
- **Map effect areas in business terms.** Organizations should document areas of effect (i.e., "blast radius") for each AI-created identity and component both in technical terms as well as in terms of affected business functions.
- **Pre-position AI-specific telemetry.** Organizations should identify, collect, retain, and test AI evidence sources before an incident, including prompt/completion logs, model lineage, embedding queries, training-data access, agent traces, and drift metrics.
- **Build containment decision trees before incidents.** For each important AI component, organizations should define containment options (e.g., revoke, rollback, isolate, disable, fallback), document cascading effects, and specify who has authority to accept the resulting business risk.
- **Run tabletop simulations with IR, ML, and business owners together.** In addition to technical responders, Saul recommended practicing scenarios with the people who understand model behavior, business impact, and operational authority.

5. Session — Misuse and Malfunction: Understanding Emerging AI-Induced Incidents

This session was chaired by Jon Boyens of NIST and examined "Misuse and Malfunction: Understanding Emerging AI-Induced Incidents." The session shifted the workshop focus from AI systems as targets of attack to AI systems as sources, facilitators, or amplifiers of harm. It explored how organizations should understand and respond to AI-induced harms, including malicious misuse, non-adversarial malfunctions, hallucinations, and other frontier-model risks. Across the talks, the session emphasized the need to distinguish between information sharing, incident reporting, and incident response; to move beyond anecdotal incident narratives toward structured data and probabilistic risk measurement; and to develop practical response mechanisms for large-scale, cross-organizational GenAI incidents.

5.1. AI Incident Database: The Next 6,173 AI Incident Reports, by Sean McGregor (AIID)

Dr. Sean McGregor, Executive Director of the AI Incident Database, spoke about how AI incident management must move beyond isolated anecdotes toward structured incident data, shared evidence practices, and measurable risk. Drawing on the AI Incident Database's (AIID) first 6,173 incident reports, he emphasized that individual reports are useful but insufficient. The field needs data collection and modeling practices that can support learning, recurrence analysis, and risk measurement at scale.

McGregor began with an analogy to aviation safety, where accident and incident databases helped create a culture of systematic learning. He argued that AI needs a similar evidence base but noted that AI incidents are harder to define because they span intentional misuse, unintentional failures, security vulnerabilities, and broader harms. The AIID uses the term "incidents" because narrower terms (e.g., accidents, failures, vulnerabilities, threats, catastrophes, recalls) fail to capture intentional harms, are not event-oriented, or are too specific.

A central theme of the talk was that AI is too broad for any single organization or database to cover alone. McGregor argued that the field needs many databases that serve many purposes, are maintained by many organizations, share data, and use common fields where possible. He organized the landscape into three major categories: 1) security incidents in which the AI system is the target, 2) misuse involving intentional harms, 3) and accidents, failures, malfunctions, and hazards involving unintentional harms.

McGregor also emphasized the political and institutional difficulty of collecting AI safety data. Companies need accident and malfunction data to improve products, but the same data can create product-liability exposure. He noted that governments are extensive deployers of AI systems and have a first-party interest in collecting accidents, failures, and malfunctions related to the systems they use.

The latter part of the talk focused on moving from incident anecdotes to measurement. Dr. McGregor discussed nowcasting and pragmatic classification frameworks that estimate whether both harm and exposure are increasing or decreasing. This addresses the "denominator problem": without understanding how widely a system or class of systems is deployed, incident counts alone cannot tell whether the technology is becoming safer or more

dangerous on a "miles traveled" basis. He called on NIST to help clarify incident categories; develop common data fields; support hindcasting, nowcasting, and forecasting; advance incident management standards; and improve the naming of AI incident types.

Call to Action

- **Be specific about incident types.** Standards and frameworks should avoid treating an AI incident as a single undifferentiated category. Rather, guidelines should distinguish between AI security incidents, AI safety incidents, AI misuse incidents, accidents, failures, malfunctions, hazards, vulnerabilities, threats, and exposures.
- **Develop common data fields across incident databases.** Rather than creating one centralized repository, standards, frameworks, and guidelines should help define interoperable data fields so multiple databases and communities can share and compare incident records.
- **Support measurement, not just reporting.** Incident collection should enable hindcasting, nowcasting, and forecasting, including methods that estimate both harm and exposure so the community can assess whether risks are escalating, receding, concentrating, or being mitigated.
- **Encourage collection of accidents, failures, and malfunctions.** Governments and companies should collect unintentional AI failure data to improve deployed systems, while recognizing the product-liability and disclosure barriers that make such collection politically difficult.
- **Promote postmortems and response records.** In the Q&A, McGregor noted that AIID has a pathway for capturing responses and wants to encourage a culture of postmortems similar to service-outage learning. That should be reflected as a practical call to improve shared learning after incidents.

5.2. Information Sharing, Incident Reporting, and Incident Response for Frontier AI Risks, Chris Meserole (Frontier Model Forum)

Chris Meserole, Executive Director of the Frontier Model Forum (FMF), focused on the institutional design problem facing frontier AI safety and security. His central argument was that information sharing, incident reporting, and incident response are distinct but complementary tools for establishing a trusted AI ecosystem. Because these terms are often used interchangeably, organizations risk building mechanisms that are poorly matched to their intended function.

Meserole described the three concepts:

1. **Information sharing** is typically voluntary, proactive, ongoing, and bi-directional or multi-directional. Its purposes are collective learning and prevention.
2. **Incident reporting** is typically reactive and often mandatory, and information flows from industry to a regulator or oversight body after a qualifying incident occurs. Its purposes are awareness, transparency, and ecosystem oversight.

3. **Incident response** is the operational process of containing, mitigating, and recovering from an identified incident in real time.

The FMF has built an information-sharing mechanism for frontier AI companies but is not itself an incident-reporting body or an incident response function. FMF's information-sharing work focuses on frontier AI risks, such as chemical, biological, radiological, and nuclear (CBRN) risks; advanced cyber and security risks; and loss-of-control or autonomous AI R&D risks. The kinds of information shared include vulnerabilities and weaknesses, threat-related information, and capabilities of concern.

Meserole warned that institutional precision matters. If incident-reporting thresholds are vague, organizations may become reluctant to share sub-threshold information voluntarily because they fear triggering regulatory obligations or liability exposure. Conversely, incident reporting should not be mistaken for a response capability. Notifying a government authority that an incident has occurred does not itself create the operational capacity to contain or remediate the incident.

Call to Action

- **Define the mechanisms precisely.** New guidance should clearly distinguish between information sharing, incident reporting, and incident response, including their purposes, timing, participants, direction of information flow, and expected outputs.
- **Avoid chilling voluntary information sharing.** Reporting thresholds should be precise enough that organizations can participate in proactive information sharing without fearing that every shared signal will trigger formal reporting obligations.
- **Build response capabilities, not only reporting channels.** New guidelines should emphasize dedicated incident response capacity, including designated personnel, protocols, practice exercises, and escalation paths. Reporting mandates cannot substitute for actual response capabilities.
- **Set proportionate reporting and response timelines.** Timelines should be designed so that organizations can provide useful, evidence-based information rather than rushed, low-quality reports.
- **Leverage existing institutions where possible.** Frontier AI introduces new risks, but response mechanisms should build on existing cybersecurity and safety infrastructure where appropriate rather than creating new or unnecessary structures.

5.3. Expect the Unexpected: Incident Response Strategies for GenAI, by Jolanda Kumakaw (Google)

Jolanda Kumakaw, Manager of Strategic Command at Google Trust & Safety and lead for Google's cross-product AI incident response work, provided an operational, practitioner-focused view of how organizations can manage critical GenAI incidents. Her central message was that GenAI incidents differ from traditional cyber incidents because they often blur the boundaries between privacy, safety, security, trust, communications, and product operations.

In this environment, “the unexpected” becomes a normal operating condition, and organizations need crisis-response structures that can function under ambiguity, speed, and cross-product complexity.

Kumakaw distinguished between three categories of incidents:

1. **Cybersecurity incidents** focus on confidentiality, integrity, and availability, such as unauthorized access, malware, and distributed denial-of-service (DDoS) attacks.
2. **AI incidents** focus on the output and logic of the model itself, including hallucinations, safety-risk outputs, and generated misinformation.
3. **AI-connected cyber incidents** sit between these categories. AI acts as a force multiplier for traditional threats, such as AI-enabled automated phishing, adversarial prompt injection, or data exfiltration through connected email tools.

Kumakaw also identified three emerging GenAI risk areas:

1. Better generative media: Improved synthetic images, audio, and video increase risks (e.g., election manipulation, non-consensual intimate imagery).
2. AI personalization: Users may believe their personal information has been compromised if an AI-generated experience feels too intimate, even when the backend is technically secure.
3. Agentic era: Autonomous agents perform tasks with less human-in-the-loop oversight, which can expand potential areas of effect when an agent makes an incorrect inference, executes the wrong action, or is manipulated through prompt injection.

To manage this environment, Kumakaw introduced a practical crisis-response framework built around the Four Cs:

- **Connect.** Comprehend the issue and risks, identify cross-product impacts and trends, find gaps, connect the dots, and serve as the connective tissue across distinct response efforts.
- **Coordinate.** Organize the tactical response across teams, assemble the right response groups, lead the response strategy, and designate a single-threaded owner.
- **Communicate.** Act as the central hub for critical incident information, own the primary communication channels, maintain a real-time source-of-truth incident document, and notify executives and key stakeholders.
- **Control.** Steer the response effort, set priorities based on contextual understanding of risk, ensure alignment across teams, create ad hoc workflows where necessary, and accelerate response where bottlenecks emerge.

She illustrated the framework with two GenAI incident scenarios: non-consensual intimate imagery (i.e., “nudification”) using GenAI and personalized user experiences in GenAI products. In both cases, effective response requires more than a technical fix. It requires coordinated operational management, cross-functional communication, stakeholder alignment, and attention to user trust and product impact.

Call to Action

- **Recognize GenAI incidents as cross-functional concerns.** Guidelines should reflect that GenAI incidents may require coordinated response across engineering, security, trust and safety, legal, communications, product, and executive teams.
- **Define operational roles for AI incident command.** Organizations should designate a single-threaded incident owner, establish clear communication channels, and maintain a real-time source-of-truth document for high-severity GenAI incidents.
- **Address cross-product and cross-system escalation.** GenAI incidents can affect multiple products or services simultaneously. Guidance should help organizations assess whether an incident is isolated or has broader cross-product implications.
- **Include communications and user-trust response in AI incident planning.** Some GenAI incidents are not resolved by technical mitigation alone. Organizations should prepare processes for executive updates, user communications, legal review, and trust restoration.
- **Plan for agentic blast radius.** Agentic systems should be treated as higher-blast-radius deployments that require clearer escalation, containment, oversight, and human-authorization planning.

6. Mini-Talks: Landscape of Existing Guidance

The mini-talk sessions surveyed the landscape of some of the existing guidelines, standards, frameworks, and resources that are relevant to AI incident response. The session was intended to give participants a structured view of what already exists, where current resources may help, and where important gaps remain. Harold Booth, a group manager in the Computer Security Division of ITL, chaired this session. Speakers from NIST, In-Q-Tel, MITRE, Japan AISI (J-AISI), International Standards Organization (ISO)-related standards work, and Carnegie Melon University's Software Engineering Institute (SEI) highlighted some of the existing resources related to cybersecurity incident response, AI cybersecurity profiling, AI incident guidebooks, AI incident data elements, adversarial AI knowledge bases, AI incident response posture, and AI incident documentation practices.

These mini-talk sessions demonstrated that AI incident management has a substantial and rapidly growing resource base. The main challenge is the fragmentation of guidelines across risk management, cybersecurity, AI-specific controls, incident response practice, information sharing, reporting data, threat knowledge, and documentation. The resources in Table 2 are complementary rather than interchangeable.

Table 2. List of mini-talks covering some of the existing resources

Resource	Mini-Talk Presenter	What It Contributes
NIST SP 800-61r3	Alex Nelson (NIST)	Cybersecurity incident response as a CSF 2.0 Community Profile. Integrates preparation, response, recovery, roles, reporting, and continuous improvement across the CSF functions.
Cyber AI Profile	Katerina Megas (NIST)	AI-specific cybersecurity profile organized around CSF 2.0. Focus areas: securing AI system components (Secure), conducting AI-enabled cyber defense (Defend), and thwarting AI-enabled attacks (Thwart).
AI Incident Response Guidebook	Andrea Brennen (IQT)	Practitioner guide with four stages: 1) preparation; 2) identification; 3) containment, eradication, and recovery; and 4) lessons learned. Emphasizes non-adversarial AI failures and concrete policies, teams, documentation, and containment plans.
AI-IRS Approach Book	Kazuaki Nimura (Japan AISI)	Extends conventional IR using two AI-specific axes: observability for detection and analysis and controllability for containment, recovery, and localized action.
ISO/IEC 25870	Michael Garris (ISO/IEC)	Structured data elements for AI incident reports: incident, harm, stakeholders, AI system, autonomy, training/model links, response actions, and reproducibility.
MITRE ATLAS	Keith Manville (MITRE)	Living knowledge base of adversary tactics, techniques, mitigations, and case studies against AI-enabled systems, including GenAI and supply-chain attack vectors.
AI incident documentation	Violet Turri (CMU SEI)	Landscape of current practices; gaps in firsthand, technically informative documentation; and policy options for confidential reporting and proactive documentation.

6.1. NIST SP 800-61 Revision 3, by Alex Nelson

Alex Nelson introduced NIST Special Publication (SP) 800-61r3 (Revision 3), *Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile*, as the current evolution of NIST's foundational cybersecurity incident response guidance. He explained that earlier versions of SP 800-61 treated incident response as a specialized organizational function, while the Cybersecurity Framework (CSF) includes incident response as a part of broader cybersecurity risk management and organizational planning.

Nelson emphasized that SP 800-61r3 is structured as a CSF 2.0 Community Profile for the broad community of incident response teams. This makes it a useful starting point for AI incident response work but also creates a scoping challenge: SP 800-61r3 is focused on cybersecurity incidents, whereas AI incidents may also involve broader harm-based misuse or malfunction scenarios.

The key issue Nelson raised was how AI should be represented in future incident response guidance. AI may be the system under attack, functioning as the target or victim. It may also be used as a deceiver or force multiplier by an attacker. In other cases, the defending organization may not itself be an AI adopter but may still face AI-enabled threats. Nelson framed these distinctions as important inputs for deciding what AI-specific updates, companion guidance, or additional resources may be needed.

6.2. Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile), by Katerina Megas (NIST)

Katerina Megas presented NIST Interagency Report (IR) 8596, *Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile*, which is a CSF-based resource to help cybersecurity practitioners understand how AI changes their day-to-day responsibilities. She explained that NIST developed the Profile in response to community feedback from cybersecurity professionals who were already "fighting fires" and needed practical help identifying where to start with AI-related cybersecurity concerns.

Megas emphasized that the Profile does not seek to replace existing cybersecurity frameworks or create a new stand-alone regime. Instead, it uses the CSF as a familiar organizing structure and overlays AI-specific implications across CSF functions, categories, and subcategories. The Profile is designed to help cybersecurity and AI communities find common touchpoints, use shared framing, and connect to existing references and resources.

She described three focus areas in which AI affects cybersecurity practices:

1. **Securing AI systems** — Understanding how the attack surface changes when organizations introduce AI systems and what must be done differently to protect those systems
2. **Using AI for cyber defense** — Identifying where AI can assist or support cybersecurity functions and where organizations may get the most value from adoption

3. **Thwarting AI-enabled attacks** — Understanding how adversaries may use AI and how existing defensive activities may need to evolve in response

Megas also described the development process as community-driven. NIST conducted a landscape review, issued a concept paper, held workshops and community-of-interest meetings, received about 1,400 comments, and is consolidating that input toward an initial public draft. Important themes from the community included understanding how frameworks relate to each other, human-in-the-loop considerations, supply-chain security, and identity and access management.

6.3. IQT AI Incident Response Guidebook, by Andrea Brennen

Andrea Brennen described IQT's work on AI incident response as emerging from earlier AI assurance efforts. Beginning in 2021, IQT examined publicly reported AI incidents and found that many involved unintentional system failures (e.g., inadequate validation, spurious relationships in data, biased training data, brittle model behavior) rather than malicious attacks. This led IQT to define AI incidents broadly enough to include both intentional attacks and unintentional failures.

Brennen explained that IQT later developed AI red-teaming and assurance-audit practices as well as a library of tabletop exercises that involve AI-specific attacks, errors, and malfunctions. These exercises have revealed a recurring organizational gap: many institutions implicitly assume that existing cybersecurity incident response policies, procedures, and personnel will simply adapt to AI incidents when they occur. IQT's work suggests that this assumption is incomplete because AI systems can fail in technically unusual ways that may produce significant operational or mission impacts.

In 2024, IQT worked with Luminos Law to develop the AI Incident Response Guidebook, which helps practitioners bridge the gap between traditional information-security incident response and AI incident response. Brennen highlighted three key takeaways from the guidebook:

1. Organizations should adopt an operational definition for an AI incident. IQT defined an AI incident as an event caused by the undesirable or unexpected behavior of an AI system that directly or indirectly causes immediate or potential harm to people or property.
2. Organizations should prepare short-term containment plans for deployed AI systems before incidents occur so that responders are not forced to decide during the initial investigation whether to take a system offline, roll back to a previous version, use an alternative tool, or apply another contingency.
3. Organizations should develop AI incident response policies that address the guidebook's recommended topics, including the organizational, legal, technical, and operational questions that arise when AI systems fail or behave unexpectedly.

6.4. MITRE ATLAS: Mapping the Adversary, by Keith Manville

Keith Manville introduced MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) as a public, community-driven knowledge base for understanding adversary behavior against AI-enabled systems. ATLAS sits at the intersection of AI and cybersecurity and is intended to help organizations understand how adopting AI expands their attack surface and introduces attack techniques that differ from traditional IT systems.

Manville emphasized that ATLAS takes an adversary-centered view. Modeled heavily on MITRE ATT&CK, it captures both traditional and AI-specific adversary tactics, techniques, and procedures that can be used to attack AI systems. The ATLAS Knowledge Base is built from real-world evidence, including observed incidents and realistic exercises against production-like AI systems. It also identifies mitigations that organizations can use to improve their AI security posture.

Manville described two complementary parts of the ATLAS program. The threat-intelligence side focuses on developing and maintaining the knowledge base. The threat-operations side uses that knowledge base to build tools, red-team activities, and threat-emulation plans. These two sides reinforce one another: real-world evidence can inform emulation plans, and operational exercises can feed lessons back into the knowledge base.

Manville also highlighted several practical outputs built on top of ATLAS, including rapid-response analysis of emerging AI security issues, public case studies, attack-graph-style analytic products, incident-reporting formats based on ATLAS, and a tabletop exercise developed with CISA's Joint Cyber Defense Collaborative. The broader message was that ATLAS can help AI incident responders, red teams, and security practitioners move from abstract concern about AI attacks toward structured, evidence-based preparation and response.

6.5. Japan AI Safety Institute's AI Incident Response System, by Kazuaki Nimura (J-AISI)

Kazuaki Nimura presented the Japan AI Safety Institute's work on an AI Incident Response System (AI-IRS) as an AI-specific extension of existing cybersecurity incident response practices. His central message was that organizations should not assume that AI incidents can be fully prevented. Because prompt-injection defenses remain imperfect, agentic systems may continue acting autonomously under uncertain conditions. AI behavior is shaped during real-world operations through interactions with users, data, tools, and external systems, and organizations need to evaluate their ability to respond after deployment.

Nimura argued that pre-deployment testing and predefined assumptions are necessary but insufficient. Critical risks may emerge only in the operational phase, including accidents, failures, misuse, or other incidents. Therefore, AI incident response needs to assess whether both the system and the organization can detect, analyze, contain, and recover from incidents after deployment.

The AI-IRS framework centers on two measurable capabilities: observability and controllability. Observability corresponds to detection and analysis in the incident response life cycle, which help organizations identify what happened and determine the cause of an incident quickly and

accurately. Controllability corresponds to containment and recovery, which help organizations localize the impact, take appropriate containment actions, and support recovery without unnecessarily disrupting the broader system.

Nimura emphasized that observability and controllability must work together. Observability without controllability may allow an organization to see that a problem is occurring but leave it unable to contain the problem with precision. Controllability without observability may lead to overly broad or incorrect actions, such as disabling more tools or functions than necessary because the underlying cause is unclear. He also stressed that AI incident response cannot be limited to a single organization. AI systems depend on models, data, external services, and other supply-chain components that may sit outside of one organization's direct control. Effective response requires cross-organizational coordination, common formats for sharing system and incident information, response protocols, and mechanisms for early warning across the AI supply chain.

6.6. ISO/IEC 25870: Data Elements for Reporting AI System Incidents, by Mike Garris

Mike Garris presented ISO/IEC 25870, which is an active ISO/IEC JTC 1/SC 42 project focused on data elements for reporting AI system incidents. He emphasized that the standard does not attempt to define the full AI incident response life cycle, assign reporting responsibilities, determine who should operate reporting systems, or decide which incidents must be reported. Instead, it addresses a more specific but foundational problem: how to describe AI system incidents in a consistent way.

The emerging definition of an AI system incident is an event associated with an AI system that results or can result in harm. Garris noted that this definition covers incidents across the AI life cycle and includes near misses. The standard does not define harm in the abstract but includes data elements for describing harm types.

Garris explained that the work originated from the OECD's common reporting framework for AI incidents, which included 29 data elements. ISO/IEC SC 42 took that work forward, adapted it into an ISO-compliant structure, and added a risk management framing aligned with related AI risk management standards. The current draft standard contains 38 data elements organized into five categories, including administrative information, incident information, AI system information, and incident analysis. The incident information category, where harm is described, is the largest category. Garris emphasized that the standard is intended to complement rather than replace or contradict existing incident reporting and sharing systems, such as cybersecurity, medical, and consumer-product reporting models. Its contribution is to make AI-system-specific incident reporting more consistent by describing the AI system itself and communicating the nature of harm more robustly. At the time of the workshop, the project was in working draft form and moving toward a committee draft.

6.7. Exploring the AI Incident Documentation Practice, by Violet Turri (Carnegie Mellon University Software Engineering Institute)

Violet Turri examined the current state of AI incident documentation and the gaps that limit the ability of engineers, organizations, and policymakers to learn from AI system failures. She argued that current incident accounts often capture the "who" and "what" of an incident but provide less detail about where in the AI system the failure occurred, when in the life cycle it was introduced, and why the system behaved as it did. These gaps are difficult to close because AI behavior is tightly bound to deployment context, data, model components, life cycle changes, and unresolved explainability challenges.

Turri compared AI incident documentation with reporting practices in aviation and cybersecurity. Aviation offers models for systematic incident collection, analysis, investigation, sharing, end-to-end timeline capture, and confidential or de-identified reporting. Cybersecurity offers models for standardized vocabularies, unique identifiers for known vulnerabilities, cross-organization sharing, and reporting requirements that support first-hand, validated data. She also discussed the potential value of AI system registries and controversial AI system catalogues, which could be linked to incident reports to trace AI system life cycles and support proactive risk monitoring.

Turri's recommendations centered on augmenting existing AI incident documentation practices through three mechanisms:

1. A federally overseen repository with mandatory reporting to enable first-hand, validated reports
2. Confidential or de-identified submission channels to encourage engineers and others to share detailed failure information without fear of repercussions
3. Proactive documentation of AI systems before incidents occur so that organizations can better track life cycles and detect early warning signs

7. Panel Discussion — Incident Response Roles and Responsibilities in the Era of AI Agents

Victoria “Vicky” Pillitteri of NIST moderated the panel discussion and examined how roles and responsibilities should be assigned when AI systems cause harm, are misused, or behave unpredictably. The panelists included Madhu Srikumar (Partnership on AI), Mikel Rodriguez (MITRE), Jim Reavis (Cloud Security Alliance), Apostol Vassilev (NIST), and Ian Reynolds (Hugging Face). Together, they focused on the institutional and operational challenge of managing AI incidents across a distributed value chain that may include foundation model developers, hosting platforms, API providers, framework or harness providers, deployers, operators, users, and affected parties.

A central theme, emphasized especially by Madhu Srikumar, was that AI value chains fragment both visibility and responsibility. A foundation model provider may see model-level behavior; a platform or API provider may see service activity; a framework or harness provider may see execution traces; a deployer may see inputs, outputs, business context, and user impact; and users or affected parties may be the first to observe harm. No single actor necessarily has all of the evidence, authority, or technical control needed to detect, characterize, contain, and report an AI incident.

The panelists also emphasized that AI agents complicate responsibility because the operational risk of an agent is not determined by the foundation model alone. Mikel Rodriguez and others highlighted that the surrounding harness or scaffolding (e.g., tools, memory, orchestrators, system prompts, retrieval systems, interfaces, credentials, delegated autonomy, downstream integrations) can determine what actions an agent can take and what evidence is available after something goes wrong. Therefore, incident response guidance needs to account for the full agentic system, not only the underlying model.

Ian Reynolds raised questions about adjudication, evidence, and responsibility in distributed systems, including who has authority to decide whether a reported event qualifies as an incident when developers, deployers, users, or platform operators disagree. This points to the need for evidentiary standards and clearer processes for resolving contested incident reports.

Apostol Vassilev emphasized the importance of taxonomies, controls, and technical evidence for securing AI systems while also cautioning that telemetry and logs can introduce new risks. Detailed logging is essential for traceability, but if logs containing malicious prompts or adversarial content are later consumed by agents or automated systems, they can become a new attack surface.

Jim Reavis highlighted the need for continuous monitoring, deception infrastructure, and practical mechanisms for learning from emerging threats before they scale. He argued that organizations need to move beyond one-time monitoring or “patch once and move on” thinking toward continuous tracing, reporting, and adaptation.

The panel also stressed that mature end-to-end AI incident response capabilities are not yet widely in place. Existing cybersecurity practices provide useful building blocks, but AI-agent incidents may unfold across multiple steps, systems, organizations, and time horizons. A hallucination, unsafe tool call, data exposure, or harmful downstream action may not have a single obvious root cause. Organizations need trace-based investigation methods, continuous

evaluation, operational rehearsals, clear evidentiary standards, and processes for incorporating lessons learned quickly.

Finally, the panel framed the AI incidents as a new set of challenges that demand collaborative community response instead of being addressed separately by individual enterprises, frontier labs, or cybersecurity teams. Effective response will require participation from model providers, deployers, platforms, open-source communities, sector specialists, standards bodies, security teams, legal and compliance functions, communications teams, and affected communities. The panel's practical message was that AI incident response must become a coordinated, continuously improving operational discipline.

Call to Action

- Establish coordinated information-sharing and reporting infrastructure across the AI value chain so that model providers, deployers, platforms, model adapters, framework providers, hosting services, and affected organizations can contribute relevant incident data.
- Define evidentiary standards and adjudication processes for AI incidents, including who decides whether an event qualifies as an incident when developers, deployers, platforms, or reporters disagree.
- Improve telemetry and traceability across AI-agent workflows, including inputs, outputs, tool calls, third-party resources, provenance, agent-to-agent interactions, and human approvals or decision points.
- Treat logging as essential but risky, and design evidentiary systems that preserve traceability without turning logs into new prompt-injection, data-exposure, or agent-consumption attack surfaces.
- Develop context-specific risk taxonomies and playbooks for high-impact AI use cases rather than relying only on generic frameworks.
- Invest in deception infrastructure (e.g., honeypots or honeynets for AI-enabled threats) to gather telemetry on emerging attacks before they scale.
- Conduct operational tabletop exercises and simulations for AI-agent failures, including multi-step failures in which an error at one stage leads to harmful action later.
- Move from one-and-done security monitoring toward continuous evaluation, monitoring, tracing, and reporting.
- Broaden participation beyond frontier labs and cybersecurity experts. AI incident response requires participation from platforms, deployers, open-source communities, sector specialists, standards bodies, researchers, and affected communities.

8. Breakout Session Synthesis

The workshop concluded with breakout sessions designed to gather participant feedback on NIST's potential next steps for AI incident response. The sessions followed the workshop's two-track framing: one track focused on AI systems as targets or components of cybersecurity incidents, and the other focused on AI misuse, malfunction, and harm-based incidents.

To include both in-person and remote participants, the workshop held two in-person breakout sessions and two virtual breakout sessions, broken down into two more. Each format covered the two major tracks. Participants discussed gaps in existing guidance, needed playbooks, evidence and logging challenges, roles and responsibilities, value-chain coordination, and priorities for future NIST work.

The synthesis that follows is not a verbatim transcript. Rather, it is a consolidated summary based on the breakout-session chairs' reports and facilitator readouts, organized around the major themes that emerged across the in-person and virtual discussions.

8.1. Track 1: AI Systems as the Target of Cybersecurity Incidents

The in-person cybersecurity breakout focused on incidents in which the AI system is itself the target or victim of a cybersecurity or information-security event. Participants identified AI systems as new targets and attack surfaces, including compromised or manipulated agents, credential harvesting and privilege escalation in agent environments, AI-assisted phishing, shadow AI deployments, AI-enabled vulnerability discovery and exploitation, and cloud-native ephemeral workloads that may disappear before investigators can understand what happened. Participants emphasized that core security practices (i.e., system hardening, benchmarking, logging, identity and access management) remain essential. However, these practices must be extended to agents, harnesses, models, AI-specific infrastructure, and non-human identities.

The virtual cybersecurity breakouts reinforced similar concerns. Participants highlighted risks associated with shadow MCP servers, "vibe-coded" applications created outside IT oversight, exposed API keys and credentials, data spills, insecure AI-generated code, and uncertain liability hand-offs among model developers, deployers, and users. Across these discussions, visibility emerged as the prerequisite for all later response activity. Organizations cannot apply threat modeling, identify stakeholders, preserve evidence, or respond effectively to systems that security teams do not know exist.

Participants also stressed that AI incident response exposes a workforce and terminology gap between cybersecurity and AI communities. Cybersecurity practitioners and AI practitioners often use the same terms differently (e.g., "red teaming" may mean penetration testing to cyber professionals but guardrail bypass testing to AI professionals). Cyber responders may understand traditional artifacts, logs, and malware investigation but lack familiarity with the AI pipeline from data ingestion through model training, deployment, retrieval, inference, and agent tool use. This gap complicates triage, evidence collection, root-cause analysis, and coordination during incidents.

8.2. Track 2: Misuse, Malfunction, and Harm-Based Incidents

The in-person harm-focused breakout emphasized that AI incident prioritization depends heavily on context and consequence. Participants discussed categories of harm (e.g., privacy breaches, financial harm, discrimination, client-facing errors, infrastructure damage, property damage, rights-related harms, malicious use, agentic behavior) that may affect safety-critical environments. A recurring point was that a technically secure AI system can still create an AI incident if it produces incorrect outputs, enables unacceptable decisions, or causes harm in a specific operational context.

Participants also discussed the difficulty of identifying AI incidents before human users recognize that something is wrong. AI systems can behave probabilistically, may not reproduce the same failure twice, and may be manipulated through rare, context-dependent, or adversarially selected inputs. Participants raised the possibility of AI-assisted or agentic defenses operating at machine speed, particularly in scenarios in which human responders cannot keep pace with automated attacks. At the same time, they recognized that governance, authorization, provenance, and accountability mechanisms are not yet mature enough to support fully automated response without careful safeguards.

The harm-focused discussion also underscored that existing taxonomies and frameworks do not always map neatly to AI harms. Participants noted that AI incidents may combine cybersecurity, privacy, safety, consumer protection, civil rights, and sector-specific regulatory concerns. This creates a need for crosswalks among existing frameworks and clearer guidance on how harm-based incidents should be categorized, prioritized, escalated, and remediated.

8.3. Cross-Breakout Themes

- **From policy to implementation.** Participants emphasized the need for technical recommendations, controls, implementation examples, and operational playbooks that map high-level policy and frameworks to day-to-day engineering and incident response activities. Strategic frameworks (e.g., NIST AI RMF, CSF) were viewed as useful, but participants wanted more guidance on how to translate those frameworks into practical controls, procedures, and response actions.
- **Logging, retention, and detection.** Participants agreed that logs are essential but insufficient for AI incident response. Logs must be strategically selected, retained, protected, searchable, and paired with usable detection logic. Participants warned that collecting large volumes of logs without the ability to identify relevant signals can overwhelm analysts and produce false positives rather than actionable evidence.
- **Privacy, compliance, and logging tension.** Several participants highlighted a direct tension between the evidence needed for AI incident response and the privacy or compliance constraints that may limit evidence collection. The most useful logs for AI IR may include prompts, outputs, data access records, retrieved content, tool calls, and prompt-injection traces. Those same logs may contain personal data, regulated information, proprietary data, or other sensitive content subject to PCI, GDPR, DLP,

retention, or privacy obligations. Organizations need guidance on how to preserve useful evidence while meeting privacy, compliance, and records management requirements.

- **Asset and dependency management.** AI asset management must extend beyond traditional hardware and software inventory. Participants identified the need to track deployed models, data sources, agents, harnesses, MCP servers, retrieval systems, identities, credentials, business functions, vendor dependencies, and downstream systems. Unknown or ungoverned AI systems (e.g., shadow AI tools, shadow MCP servers, vibe-coded applications) substantially increase response difficulty because responders may not know what exists, who owns it, what data it accesses, or what business process it supports.
- **Vendor, procurement, and value-chain constraints.** Participants emphasized that many organizations do not fully control the AI systems they rely on. Model providers, resellers, cloud providers, managed service providers, API platforms, and third-party vendors may control logs, model documentation, technical artifacts, and remediation actions. Government and enterprise procurement channels may further limit what information can be compelled or obtained during an incident. This creates a need for contract language, audit rights, incident notification clauses, evidence commitments, and clear contact paths across the AI value chain.
- **Terminology and framework crosswalks.** Participants noted the presence of inconsistent terminology and role definitions across NIST, ISO, the EU AI Act, MITRE, OWASP, CSET, the Coalition for Secure AI, and other frameworks. Rather than calling only for new definitions, participants asked for crosswalks that help organizations understand how existing definitions, roles, taxonomies, and incident categories relate to one another. This was viewed as especially important for organizations trying to align AI governance, cybersecurity, legal, compliance, and operational response functions.
- **AI transparency artifacts and provenance.** Participants debated whether AI SBOMs, model cards, or other machine-readable transparency artifacts can support incident response. Some participants saw AI SBOMs as a potential way to standardize information about model versions, scaffolding, fine-tuning, and dependencies. Others were skeptical, noting that current SBOMs are often used for third-party risk questionnaires rather than actual incident analysis and that there is not yet consensus on what an AI SBOM should contain. Provenance tracking was also described as difficult once inputs, model behavior, and emergent system interactions are combined.
- **Forensic correctness and AI artifacts.** Breakout participants raised unresolved questions about how to prove that an AI system has been corrected after an AI-specific incident, such as data poisoning, model compromise, prompt injection, or unsafe agent behavior. Unlike traditional malware analysis in which investigators may identify a specific malicious file or signature, AI failures may be distributed across data, weights, prompts, retrieval systems, or agent workflows. Participants requested DFIR guidance on what AI

artifacts investigators should collect while recognizing that NIST may need to provide broad artifact categories rather than prescribe proprietary vendor formats.

- **Business impact and area of effect.** Responders need to understand what an AI system supports before an incident occurs. Participants stressed that an AI failure that affects manufacturing, medical diagnostics, billing, financial onboarding, aviation, critical infrastructure, or student-facing educational tools will require different escalation paths and containment decisions. AI asset documentation should include business functions, safety relevance, fallback procedures, acceptable disruptions, and potential downstream impacts.
- **Machine-speed response and human governance.** Participants repeatedly raised the concern that AI-enabled attacks or agentic failures may unfold faster than traditional human-centered response processes. However, they also recognized that automated or agentic defenses require governance, authorization, provenance, and accountability mechanisms that are not yet fully mature. Response models are needed to define what actions can be automated, what actions require human authorization, and how decisions are logged and reviewed.
- **Examples and learning resources.** Participants requested practical case studies, best-practice repositories, tabletop scenarios, and examples of AI incident response in operational settings. This need was especially acute for smaller organizations that lack specialized AI security staff, legal resources, or internal tooling to develop AI-specific safeguards on their own. Participants viewed practical examples as necessary to help organizations conceptualize what AI incident response looks like in practice.

9. Conclusion

The workshop established that SI incident management is not an abstract future problem. Organizations are already training, evaluating, and deploying SI systems that may be susceptible to compromise or take unintended actions while handling sensitive data, supporting decisions, automating processes, customer interactions, security operations, or critical business functions. Existing cybersecurity incident-response foundations remain indispensable, but SI introduces new challenges in observability, control, evidence, governance, the value chain, the speed of incident occurrence, identification, the assessment of harm, and response.

The near-term opportunity is to translate the workshop's findings to develop focused, connected, and implementation-ready guidelines that organizations can use. The guidelines should be built on existing in-practice documents and frameworks, such as SP 800-61r3, the upcoming document IR 8596, MITRE ATLAS, ISO/IEC incident-reporting work, the IQT Incident response guidebook, JCDC resources, SI incident databases, and other resources. NIST's continued work will extend traditional cybersecurity incident response to address the operational realities of SI systems, including model life cycle events, telemetry, provenance, traceability, observability, forensic analysis, information sharing, and coordinated response across the SI value chain. The medium-term opportunity is to strengthen the ecosystem for incident data, documentation, reporting, and sharing. The long-term challenge is to resolve deeper research questions about telemetry, reproducibility, recovery, machine-speed response, attribution, and multi-party responsibility.

Recognizing how quickly SI is becoming a part of everyday life, business, critical infrastructure, and government services, NIST's SI Incident Response work will focus on staying flexible and working closely with all stakeholders. As SI capabilities and risks evolve, NIST will engage with the community to identify new challenges and develop practical guidelines for emerging issues while also expanding and improving a broader set of clear, useful advice that keeps pace with technological advances, real-world experiences, and stakeholder needs. This approach ensures that guidelines remain relevant and effective amid rapid technological change.

10. Next Steps

The workshop covered multiple topics across incident response in the era of SI. NIST plans to move forward with the following set of interconnected activities to translate stakeholder inputs into practical SI incident-response guidelines. New resources will build on existing NIST publications and other established incident-response practices. The scope and specific guidelines are expected to evolve with future feedback and the emerging needs of stakeholders.

Possible next steps and deliverables include:

1. **Establish an SI Incident Management Community of Interest**

   For a transparent, evolving process, NIST plans to establish an SI Community of Interest (COI) that brings together government, industry, academia, standards organizations, and other relevant stakeholders. This will provide an appropriate mechanism for identifying emerging challenges, exchanging lessons learned, validating proposed approaches, and informing future NIST activities and publications.

   a. **Identify Priority Gaps and Develop Targeted Guidelines**

      The COI will offer access to a variety of stakeholders. It will utilize different engagement methods (e.g., workshops, tabletop exercises, community discussions) to identify gaps in current SI incident response practices and guidance. The findings from these engagements could help inform targeted and practical guidelines that address priority needs and emerging scenarios in SI incident response.

   b. **Convene NIST Task Group on Agent Capabilities, Monitoring, and Containment**

      Task Group 8, “Agent Capabilities, Monitoring, and Containment,” will be set up to address agent security. The task group will meet on an ad hoc basis through the NIST AI Consortium. Organizations that are developing and evaluating the highest capability SI agents on long-horizon tasks will discuss challenges and strategies for SI incident response, among other topics.

2. **Update the Existing SP 800-61 Framework**

   NIST plans to update NIST SP 800-61 to ensure that broader incident-response guidelines appropriately address SI-specific considerations and clearly align with SI-focused incident-response, including a potential Part A focused on cybersecurity incidents involving SI systems and a possible Part B focused on the misuse, malfunction, or other unintended actions of SI systems that may *lead to* harm. Two initial steps are considered below:

   a. **Develop SI Cybersecurity Incident Response Guidelines**

      Develop appropriate SI-specific incident-response guidelines that complement the existing SP 800-61 framework. The guidelines would address cybersecurity incidents involving SI systems, including considerations associated with training and testing data as well as models, autonomous agents, SI infrastructure, and

the SI supply chain. The primary audience for this document would be all stakeholders at risk of an SI-involved cyber incident, including deployers and adopters of SI systems that range in capability and function.

b. **Develop Key Practices for SI Control Incidents**

This document would feature voluntary principles to guide emerging practices in the management of incidents that occur during the training and evaluation of SI agents due to agents taking unintended actions outside of the operator's intended task scope. The primary audience for this document would be organizations training or evaluating highly capable SI agents on long-horizon tasks with limited human supervision.

3. **Strengthen SI-Specific Monitoring and Log Management Practices**

Examine how current monitoring and logging practices need to evolve to support SI incident detection, investigation, and response. This work could inform updates to or additional resources for SP 800-92, *Guide to Computer Security Log Management*, including considerations for model life cycle events, SI telemetry, provenance, and forensic evidence.

4. **SI Incident Information Sharing**

Examine SI-specific requirements for incident information sharing and reporting, and consider how existing guidelines — including SP 800-150, *Guide to Cyber Threat Information Sharing* — could address SI-related information. Considerations could include model life cycle events, telemetry, provenance, incident context, and forensic information.

5. **Develop Conceptual SI Traceability and Observability**

Address the identified gap in SI traceability and observability related to incident response. This work could result in a dedicated NIST publication or another appropriate resource that establishes common concepts, operational practices, and considerations for technical standards for tracing and observing SI system behavior.

6. **Continue to Identify and Address Emerging Gaps**

Maintain a flexible body of work that integrates lessons from exercises, real-world incidents, stakeholder feedback, and advancements in SI technologies. Additional guidelines, frameworks, best practices, and research activities may be developed to continue supporting emerging needs for incident response.