AI Incident Management: Sketching the Problem Space

M S Raunak, ITL
Peter Cihon, CAISI

May 14, 2026

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

AI Action Plan

Promote Mature Federal Capacity for AI Incident Response

The proliferation of AI technologies means that prudent planning is required to ensure that, **if systems fail**, the **impacts** to critical services or infrastructure are minimized and response is imminent. To prepare for such an eventuality, the U.S. government should promote the development and incorporation of **AI Incident Response** actions into existing incident response doctrine and best-practices for **both the public and private sectors**.

AI Action Plan: NIST Tasking

Led by NIST at DOC, including CAISI, partner with the AI and cybersecurity industries to ensure AI is included in the establishment of standards, response frameworks, best practices, and technical capabilities (e.g., fly-away kits) of incident response teams.

AI Incidents – Examples

ANTHROP\C

Disrupting the first reported AI-orchestrated cyber espionage campaign

The Register

Datacenter | Security | Microsoft | AWS | Developer | Open Source | IT Careers | Columnists | Who, Me

AI + ML

Slack AI can be tricked into leaking data from private channels via prompt injection

Whack yakety-yak app chaps rapped for security crack

Thomas Claburn

Published Wed 21 Aug 2024 // 09:23 UTC

The Information

Inside Meta, a Rogue AI Agent Triggers Security Alert

By Jyoti Mann

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

Defining AI Incidents: Two proposed categories

1. AI Under Attack: AI Cybersecurity Incidents

- Adversarial compromise of security properties of AI
- AI as the object

2. Misuse or malfunction: AI-Induced Incidents

- Harm resulting from AI use or failure aside from compromise of security properties
- AI as the subject

AI Under Attack: AI Cybersecurity Incidents

Working definition:

- “An occurrence that actually or imminently jeopardizes, without lawful authority, the **confidentiality, integrity, or availability** of the AI system, any other system enabled and/or created by the AI system, or information stored on any of these systems" - JCDC Playbook

Examples/scenarios:

- adversarial inputs (e.g., prompt injection);
- model theft;
- data poisoning;
- model inversion or membership inference;
- supply-chain compromise (models, datasets, hardware, firmware);
- infrastructure compromise affecting AI systems (cloud, edge, ccelerators).

Misuse or Malfunction: AI-Induced Incidents

Working defintion:

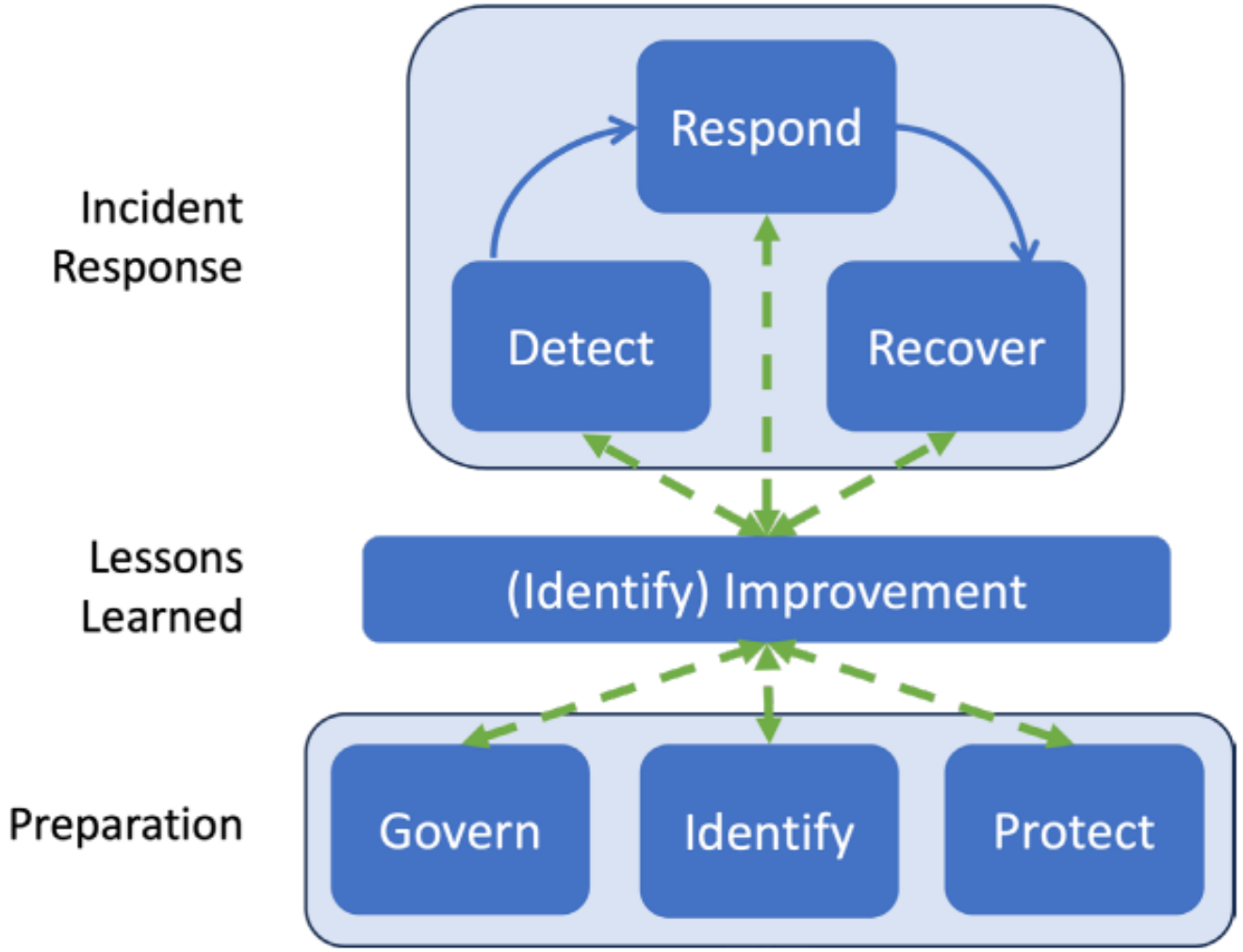
- Incident where the development, use, or malfunction of an AI system caused harm. Harms may be scoped, e.g., to national security and public safety.

Examples/scenarios:

- Malicious use or Misuse
  - Criminal assistance, e.g., using a model to assist in creating a bioweapon
  - State actor uses multiple accounts with an AI system provider to orchestrate offensive cyber operations,
  - A threat actor uses distillation to train an abliterated or uncensored version of a model.
- Malfunction
  - Hallucinations,
  - Reward hacking and specification gaming,
  - Sycophancy, encouraging user delusions or self-harm,
  - Operational failures in high-stakes settings (e.g., critical infrastructure).

Is the existing guidance sufficient?

Are the resultant incidents any different?



NIST 800-61R3 – Incident Response Lifecycle Model

The proposed road ahead

NIST ITL and CAISI plan to start with two workstreams

1. Updates to cybersecurity incident response practices for attacks on AI

   Scope: Incidents focused on cybersecurity

2. Recommendations for AI Misuse and Malfunction Incident Response

   Scope: Incidents focused on harm, particularly to national security and public safety

Do we need other Incident Response related guidance, standards, framewoks or best practices?

In today's workshop

Learn from the community

- Presentations from industry and civil society
- Roles and responsibilities panel on AI agents

Look at some of the existing resources

- NIST SP 800-61r3, ISO/IEC 25870, NIST Cyber AI Profile, IQT AI Incident Response Guidebook, Japan AISI Incident Response Approach Book, MITRE Atlas, and CMU Incident Documentaiton

Brainstorm what works, what breaks

- Breakout sessions to benefit from your expertise