RadarFirst

ADDRESSING AI HARMS IS NOT JUST AN ETHICS PROBLEM

IT IS AN INCIDENT RESPONSE PROBLEM

BY LAUREN WALLACE AND ZACH BURNETT

ADDRESSING AI HARMS IS NOT
JUST AN ETHICS PROBLEM —
IT IS AN INCIDENT RESPONSE PROBLEM

RadarFirst

PART I - TURNING AI HARMS INTO OPERATIONAL RISK

This whitepaper series examines AI harm through an operational incident-response lens. Presented in three parts, the series moves from defining the problem to building a practical and defensible response capability.

Part I explains why AI harms and hazards should be treated as operational risks. Part II outlines a structured framework for identifying, documenting, investigating, escalating, and remediating AI incidents. Part III explores two of the most consequential and difficult issues in AI incident management: determining causality and assessing severity.

Together, the three parts provide organizations with a foundation for translating responsible AI principles into repeatable, auditable, and defensible practices.

AI risk is moving from an abstract concern to an operational reality. The most important question for organizations is no longer "Can AI cause harm?" It is "Can we recognize, classify, investigate, remediate, and report AI harm when it happens?"[1]

AI incident management mirrors existing incident management models and requires preservation within a durable, auditable system of record. Privacy and security teams already know how to manage ambiguous facts, timelines, jurisdictional triggers, affected individuals, evidence, mitigation measures, and regulatory notifications. AI governance should learn from that model.

But unlike privacy harms, which are enumerated in a mature body of regulatory and governance models, AI harms are an emergent category that requires deeper investigation to provide defensible reporting and useful feedback to developers and to internal and external stakeholders.

Traditional privacy and security incident response is optimized for data breaches, malware, outages, and unauthorized access; AI incident response must also handle drift, hallucinations and bias, training and data lineage, and lack of explainability. And while privacy incidents are necessarily bound by the requirements of protecting personal information, AI incidents may implicate a much broader scope of information and processes.

---

[1] National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework: AI RMF 1.0 (NIST AI Publication 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-; Section 1.2.3 Risk Prioritization: "Attempting to eliminate negative risk entirely can be counterproductive in practice because not all incidents and failures can be eliminated."

RadarFirst

PART I - TURNING AI HARMS INTO OPERATIONAL RISK

This paper argues that addressing AI harms is an incident response challenge, not merely a question of ethics. We examine how organizations must:

- Operationalize AI risk through structured intake, rigorous investigation, and documented remediation, and
- Ensure evidence is preserved for defensible reporting.

By defining AI incidents through the lens of both harms and hazards, teams can establish repeatable escalation pathways that bridge the gap between internal standards and external regulatory mandates.

We will analyze the complexities of causality and severity, emphasizing the need for auditable, consistent records within a cross-functional system of record. Ultimately, this framework enables privacy, legal, and technical stakeholders to identify incidents earlier and cultivate a continuous feedback loop for AI governance and development.

1. AI incidents are on the rise.

According to the Stanford 2026 Artificial Intelligence Index Report,[1] documented AI incidents continued to rise, with the AI Incident Database[2] recording 362 incidents in 2025, a 55% increase over 2024.

The risk of harm from AI incidents is sector-agnostic.

Any organizational function that architects, implements, or depends upon automated systems is susceptible to operational risks and realized incidents. The AI Incident Database,[3] the underlying system of record for many reports on AI risk, defines its scope broadly as tracking real-world harms or near-harms across deployed AI systems, not in any specific sector. And the NIST approach to AI risk management is explicitly non-sector-specific and use-case-agnostic.

---

[1] Sha Sajadieh et al., The AI Index 2026 Annual Report (Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2026), page 9, https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf.
[2] Responsible AI Collaborative, AI Incident Database, accessed June 29, 2026, https://incidentdatabase.ai/.
[3] ibid.

Organizations need to develop both internal and external reporting approaches that recognize unique risks in their AI deployments.

External reporting regimes such as the EU AI Act are helpful but not sufficient: while such frameworks help establish minimum expectations for AI logging, post-market monitoring, and serious-incident reporting, the cross-sectoral and context-dependent nature of AI risk means each organization must translate those baselines into internal standards.

This means defining what constitutes an AI incident that results in harm or hazard, what evidence must be preserved, who must be notified, and how records will be maintained for each material AI use case.

2. What is an AI Incident?

AI Incident means an alleged or substantiated real-world event, circumstance, or series of events involving the development, testing, procurement, deployment, use, misuse, malfunction, failure, or operation of one or more AI systems or AI models, in which the AI system or model directly or indirectly causes, materially contributes to, or fails to prevent:

1. Actual harm to people, groups, rights, property, infrastructure, operations, security, communities, or the environment (a “harm”); or

2. A concrete near-miss in which significant harm could reasonably have occurred absent external intervention, chance, or other circumstances outside the AI system (a “hazard”).

RadarFirst

PART I - TURNING AI HARMS INTO OPERATIONAL RISK

The primary frameworks for describing AI incidents are in the EU AI Act,[5] the OECD AI Monitor,[6] the MIT AI Risk Initiative,[7] the AI Incident Database,[8] and a recently proposed Department of Defense incident and vulnerability reporting program in the 2027 National Defense Authorization Act[9]. They each treat the distinction between "harms" and "hazards" somewhat differently, reflecting each contributor's cultural and regulatory goals.

Entities developing an AI incident response plan should adopt a broad, framework-neutral definition of AI Incident that captures both actual harms and concrete near-misses involving the development, testing, procurement, deployment, operation, misuse, malfunction, or failure of an AI system. The definition should align with OECD's distinction between incidents and hazards, the EU AI Act's elevated category of serious incidents, MIT/AIID's inclusion of alleged harms and near-harms for learning and trend analysis, and DOD's emphasis on operational, safety, security, mission, guardrail, and vulnerability-related events.

---

[5] European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act), OJ L, 2024/1689, 12 July 2024, art. 3(49), art. 73. Article 3(49) defines "serious incident" as an incident or malfunctioning of an AI system that directly or indirectly leads to death or serious health harm, serious and irreversible disruption of critical infrastructure, infringement of fundamental rights obligations under Union law, or serious harm to property or the environment; Article 73 establishes reporting obligations for serious incidents.

[6] OECD.AI, Overview and methodology of the AI Incidents and Hazards Monitor, "Definitions" section. The OECD defines an "AI incident" as an event, circumstance, or series of events where the development, use, or malfunction of one or more AI systems directly or indirectly leads to specified harms, and defines an "AI hazard" as an event, circumstance, or series of events that could plausibly lead to an AI incident.

[7] MIT AI Risk Initiative, MIT AI Incident Tracker. The MIT AI Incident Tracker classifies real-world reported incidents from the AI Incident Database by risk, cause, harm, severity, and related dimensions, and uses MIT risk taxonomies and harm-severity classifications to structure incident analysis.

[8] AI Incident Database, Editor's Guide, "Definitions" section. The AI Incident Database defines an "AI incident" as an alleged harm or near-harm event to people, property, or the environment where an AI system is implicated; it also defines "AI issue," "AI incident variant," "implicated," and "nearly harmed," and states that it does not distinguish accidental from deliberate harm for incident-coding purposes.

[9] House Armed Services Committee, Cyber, Information Technologies, and Innovation Subcommittee, FY27 NDAA CITI Subcommittee Print, sec. 1501, proposed 10 U.S.C. § 2224b(j)(2)–(3). The print defines a "covered AI incident" to include events where an AI system causes unintended operational, safety, or security harm; operates outside approved safety, legal, or mission guardrails; materially degrades mission performance or reliability in real-world or operationally representative environments; or operates in a way that could foreseeably have resulted in significant unintended operational, safety, or security harm. It separately defines a "covered AI vulnerability" as an exploitable weakness, vulnerability, or systemic issue that could materially affect mission performance, compromise system integrity, create safety risk, or result in unauthorized or unintended behavior.

RadarFirst

PART I - TURNING AI HARMS INTO OPERATIONAL RISK

Plans should therefore classify events across a lifecycle spectrum: AI hazards or vulnerabilities before harm occurs; near-miss incidents where significant harm could reasonably have occurred; harm incidents where adverse impacts actually occur; and serious or reportable incidents where impacts meet legal, regulatory, mission, safety, rights, infrastructure, property, environmental, or security thresholds. This structure allows organizations to preserve a common internal reporting vocabulary while mapping specific events to applicable external reporting duties.

Recognizing AI harm as an operational risk is only the beginning. The next challenge is determining how to build a consistent, auditable, and defensible incident response framework that organizations can rely on during AI incidents. So where should that framework begin?

Learn more in our Part II - Creating a Defensible AI Incident Response Framework.

Disclaimer

This material is provided for informational purposes only and does not constitute legal advice. Organizations should consult qualified legal counsel regarding the application of data protection and AI governance laws to their specific circumstances.

RadarFirst

PART I - TURNING AI HARMS INTO OPERATIONAL RISK

About the Authors

Zach Burnett is CEO of RadarFirst, where he leads the company's vision for AI governance, regulatory risk management, and intelligent compliance. He is passionate about helping organizations operationalize trusted AI through responsible governance, intelligent automation, and enterprise-scale decision-making.

With more than 20 years of enterprise software leadership experience, Zach is a recognized voice on AI governance, agentic AI, and the future of regulatory technology.

Lauren Wallace is a technology and privacy attorney specializing in AI governance. Her background includes serving as general counsel for a venture-funded healthcare startup and a PE-backed SaaS software business, as a senior business lead for large enterprise technology companies such as Apple and Microsoft, and as lead negotiator for global enterprise SaaS and privacy agreements.

Lauren holds the designations of Certified Information Privacy Professional/US, Certified Information Privacy Manager, Fellow of Information Privacy, and AI Governance Professional. She writes extensively on the intersection of technology and privacy law, and previously served as RadarFirst's Chief Legal Officer for more than four years. Learn more at www.wallacetechlaw.com.

ABOUT RADARFIRST

RadarFirst is an AI-forward incident management platform for privacy and AI. Trusted by leading organizations, RadarFirst enables teams to manage incidents with speed, consistency, and defensibility by standardizing how incidents are captured, assessed, and actioned.