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 II - CREATING A DEFENSIBLE AI INCIDENT RESPONSE FRAMEWORK

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.

Once organizations recognize AI incidents as operational risks, they need a structured, repeatable framework to capture critical information, investigate events, preserve evidence, and ensure every incident is managed consistently and defensibly.

AI incident management needs structured, repeatable intake, not ad hoc escalation.

Organizations deploying AI should assume that incidents, near misses, and potential harms are likely. A mature AI incident program needs a common intake model that connects AI inventories, AI risk registers, harm/hazard taxonomies, monitoring signals, and escalation pathways.

Intake should capture what happened, which AI system or use case was involved, who or what was affected, whether the event reflects a realized harm, a hazardous AI condition, or both, whether the system was used as intended, whether human oversight failed or was bypassed, what preliminary causal theory exists, what evidence must be preserved, and which governance, legal, privacy, security, compliance, product, and technical stakeholders should be engaged. Structured intake is the operational bridge between AI governance and AI incident response.

Organizations can improve AI risk visibility with structured, guided incident intake.

Structured incident intake processes capture concerns consistently across business units, functions, and AI use cases. Without a common, widely deployed intake mechanism, AI-related issues may remain fragmented or be handled informally by product teams, legal, compliance, security, customer support, or human resources, thereby failing to recognize broader incident patterns.

RadarFirst

PART II - CREATING A DEFENSIBLE AI INCIDENT RESPONSE FRAMEWORK

By standardizing how AI incidents are identified and recorded, organizations can detect recurring risks earlier, compare incidents across use cases, meet external reporting obligations more reliably, and build an internal evidence base for governance, remediation, and continuous improvement.

Guided intake foregrounds your organization's risk environment.

Just as each organization must set its own harm standards (see more on that below), so must each organization determine the incident information that it wants to collect and retain. Certain common elements should be gathered in the first instance, regardless of the incident's apparent profile, so that the most consequential facts are collected while they can still support containment, remediation, notification, and learning. A minimum incident intake process should collect the following fields:

Minimum Incident Information Intake	
Field	**Purpose**
System or model involved	Identifies the product, version, vendor, integration, or decision pipeline.
Incident trigger	Explains how the issue was detected: user report, monitoring alert, audit, regulator inquiry, media report, internal escalation.
Affected population or assets	Establishes who or what may have been harmed.
Time window	Determines when the incident began, whether it is ongoing, and whether exposure is expanding.
Deployment context	Distinguishes testing, internal use, customer-facing use, automated decisioning, advisory use, or human-in-the-loop use.

PART II - CREATING A DEFENSIBLE AI INCIDENT RESPONSE FRAMEWORK

Minimum Incident Information Intake	
Field	**Purpose**
Harm categories implicated	Harm categories implicated
Severity and reversibility	Determines whether the organization must act immediately to stop harm, notify affected parties, preserve evidence, or roll back the system.
Mitigation status	Records whether the system has been paused, rolled back, constrained, monitored, or left in production.

Once this minimum information set has been created and any necessary mitigation is underway, analysts can start to focus on pertinent sector- or industry-specific information.

For example:

- A hospital should gather patient safety and clinical reliability information before brand-impact information.
- A bank should gather information on financial loss, account impact, and disparate outcomes before the general narrative context.
- A water utility should gather service-disruption, public health, and environmental-containment information before reputational metrics.
- A public benefits agency should gather due process, eligibility, and vulnerable population information before operational efficiency metrics.

Developing minimum-intake profiles and organization-specific elements will narrow the scope of investigation while ensuring that essential information is collected and preserved.

Define your organization's harm standard as a starting framework for incident reporting.

Consider a bottom-up design in which your organization's regulatory obligations set the baseline. Do you recognize a reporting obligation under the EU AI Act? Does your industry require reporting to self-governing bodies? If so, capture the specific reporting requirements, which will necessarily be narrower than the scope of incident information needed for effective AI governance.

PART II - CREATING A DEFENSIBLE AI INCIDENT RESPONSE FRAMEWORK

Next, translate harm categories into reusable building blocks. These can be mapped to any incident to determine next steps. Blocks might include categories such as Protected Interest; Affected Stakeholders; AI System; Severity, Scale, and Causation; Provider vs. Deployer status; and Other-regime notifications.

Before proceeding, set this essential gate: will your organization record and investigate both harms and hazards, or only incidents where an actual harm has already occurred? The first approach will provide stronger signals for the safety of your AI deployment, but the second may be more manageable in the early stages of adoption.

Now, build a widening funnel from the bottom up (regardless of whether following a "harms only" or "harms+hazards" model, the more information you collect, the better).

Prepare to capture narrative information about the incident, and specify responses to the "building blocks" from the previous section.

Once you have comprehensive information about the incident - allowing for potential updates - assess for two main levers: Severity and Causality. The next paper in the series provides more detailed guidance on how to consider these factors in the context of AI incidents.

Finally, create a standardized per-incident severity ranking (again, this should be flexible, as the ranking may change over time as you learn more about each incident). Assign escalation triggers and processes for each assessed level.

For example:

Incident Severity Rankings & Actions

Severity	Label	Description	Example Action
S0	Signal	Unverified concern, anomaly, or complaint	Log and triage
S1	Hazard / near miss	Credible potential harm, no confirmed adverse effect	Assign risk owner

PART II - CREATING A DEFENSIBLE AI INCIDENT RESPONSE FRAMEWORK

Incident Severity Rankings & Actions			
Severity	**Label**	**Description**	**Example Action**
S2	Minor actual harm	Unverified concern, anomaly, or complaint	Log and triage
S3	Moderate harm	Meaningful adverse effect, limited scale, or reversible	Formal investigation
S4	Serious harm	Severe, systemic, rights-impacting, vulnerable-population, or hard-to-reverse harm	Executive escalation
S5	Mandatory / crisis	Legally reportable, catastrophic, material, or public-safety event	Legal / regulatory response

RadarFirst

PART II - CREATING A DEFENSIBLE AI INCIDENT RESPONSE FRAMEWORK

Even the strongest incident response framework depends on sound judgment. After an incident is documented and investigated, organizations must answer two of the most difficult questions in AI governance: Did the AI actually cause or contribute to the outcome, and how serious is the resulting harm?

Learn more in our Part III - Special Considerations: Causality and Severity.

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 II - CREATING A DEFENSIBLE AI INCIDENT RESPONSE FRAMEWORK

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.