RadarFirst

ADDRESSING AI HARMS IS NOT JUST AN ETHICS PROBLEM

IT IS AN INCIDENT RESPONSE PROBLEM

BY LAUREN WALLACE AND ZACH BURNETT

RadarFirst

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

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.

Not every AI failure warrants the same response. Determining whether an AI system caused or materially contributed to an outcome and assessing the severity of the resulting harm are critical decisions that shape investigation, escalation, regulatory reporting, and organizational accountability.

Causality is a core challenge.

AI incidents often involve partial, indirect, or uncertain causal links. Teams need to document whether AI directly caused harm, contributed to it, failed to prevent it, or was misused or over-relied upon.

But-for causality and defensible postures in AI incident management.

But-for analysis should be the starting discipline, but AI incident recognition may also require a material-contribution lens where AI did not singlehandedly cause the harm but meaningfully shaped, accelerated, amplified, or failed to prevent it.

Litigation versus Regulatory review.

A litigation posture is usually retrospective, adversarial, and focused on liability, causation, damages, reliance, duty, foreseeability, and admissions. A regulatory posture is often more process-oriented. Regulators may care less about proving tort-style causation and more about whether the organization had a reasonable, timely, documented, and accountable process.

Auditable, consistent records are essential for both postures.

Severity must be tied to real-world impact.

“AI risk” is too broad. Leaders need to classify concrete harms: fairness, health, financial loss, infrastructure disruption, environmental damage, property damage, reputational impact, and rights infringement.

RadarFirst

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

The European Commission's Draft Guidance on Article 73 serious-incident reporting under the EU AI Act[1] sets out general principles for determining whether an incident in one of the high-risk categories should be considered "serious," as well as several examples of incidents that would merit the "serious" designation.

For organizations in regulated sectors, comparison to the harm standards in other, more mature reporting regimes may be helpful. Comparable reporting rules in privacy, cyber, finance, healthcare, and critical infrastructure include:

1. HIPAA, GDPR, and US State data breach notification laws
2. FDA Medical Device Reporting
3. US Banking Computer-Security Incident Notification Rule; NYDFS Cybersecurity Regulation, SEC cybersecurity incident disclosure
4. NIS2, DORA

Tailor data collection practices to your organization's harm standard.

While organizations must consider the risks of overcollection (through the lenses of data minimization, privilege, confidentiality, and security requirements), you'll want to ensure your organization collects enough incident information to meet any regulatory, contractual, or organizational requirements for incident reporting. The degree of detail that you collect will depend on whether the organization has mandatory reporting obligations and where you land on a "harms-only" or "harms+hazards" approach.

As you set up your data collection methodology, be sure to use a platform that meets your organization's legal privilege, confidentiality, retention, and discoverability requirements. While data collection will necessarily include stakeholders from across the organization, you may want to centralize the recording function where counsel can apply proper protections.[1]

---

[1] European Commission, "AI Act: Commission Issues Draft Guidance and Reporting Template on Serious AI Incidents, and Seeks Stakeholders' Feedback," Shaping Europe's Digital Future, September 26, 2025, last updated November 4, 2025, accessed June 29, 2026, https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks.

[1] Use of general-purpose AI (GPAI) tools to research or classify data elements during an AI-incident response may create potentially discoverable artifacts, including prompts, outputs, chat histories, account settings, tool logs, and related metadata, and may also risk waiver or loss of confidentiality if sensitive, privileged, or regulated information is entered into a public or non-enterprise tool. U.S. law remains unsettled: courts have compelled production of AI prompts, outputs, account information, and testing documentation where the AI work was relevant to claims or defenses, while other decisions have treated attorney-crafted prompts or AI-assisted litigation preparation as protected work product; recent courts have also split on whether use of a third-party GPAI tool defeats privilege or work-product protection. Accordingly, incident teams should assume GPAI interactions may later be sought in discovery and should use counsel-directed, enterprise-controlled, retention-limited workflows when analyzing incident data elements.

RadarFirst

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

The table in the Appendix provides granular explanations of the differences in data capture for a "harms-only" versus a "harms+hazards" approach. For a useful shorthand, apply this lens:

> -> harms-only asks, "What happened and who was harmed?" A harms+hazards standard asks that too, but also asks, "What could have happened, how plausible was it, and what stopped it? How can these lessons be applied to future development and AI governance?"

Evidence readiness will separate mature programs from reactive ones.

Organizations should be prepared to show what they knew, when they knew it, what actions they took, and why they concluded an incident was or was not reportable. Prompts, outputs, model versions, logs, confidence scores, monitoring alerts, user actions, and decision records should be preserved while the information is fresh and readily available upon request.

Consistent, time-stamped, and auditable formats will speed incident management in fast-changing scenarios and provide a ready response to urgent regulatory inquiries.

Strong AI governance rests on durable, auditable documentation.

AI governance will ultimately be judged not by the sophistication of its principles, but by the quality of the organization's response when an AI system causes, contributes to, or nearly causes harm.

Mature organizations will be able to reconstruct what happened, explain how they assessed causality and severity, show why an incident was or was not reportable, document the actions taken to remediate risk, and preserve those decisions in a durable, auditable system of record.

Consistent, comprehensive records are not administrative overhead; they are the foundation for regulatory credibility, contractual assurance, market trust, and continuous improvement.
As AI deployments scale, incident readiness becomes the operational proof that responsible AI commitments can withstand real-world scrutiny.

RadarFirst

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

Executive takeaways:

> ***"As AI continues to expand its capabilities, there is an urgent need to systematically collect AI incident data to enhance our knowledge of AI-related harm and help us develop safe, secure, and trustworthy AI systems." - CSET***

i. AI governance must become operational. Policies and principles are no longer enough; organizations need repeatable processes to detect AI incidents, classify harm, assess reportability, escalate decisions, remediate quickly, and preserve a defensible record.
ii. Organizations need their own harm-and-hazard standard. External regimes such as the EU AI Act provide important reporting baselines, but each organization must define what counts as an AI incident, what severity levels matter, which stakeholders must be engaged, and whether near misses or hazards will be investigated before harm occurs.
iii. Auditable evidence is the backbone of AI incident response. Prompts, outputs, model versions, logs, monitoring alerts, human-review records, causal assessments, severity determinations, and remediation decisions should be captured before they disappear and maintained in a durable cross-functional system of record.

Evaluating causality and severity transforms AI incident management from a reactive exercise into a repeatable governance capability. By applying consistent decision criteria, organizations can make more defensible response decisions, improve cross-functional coordination, and create an auditable record that supports both regulatory obligations and continuous improvement. But incident response should not end with documentation and remediation.

How organizations capture lessons learned and feed those insights back into AI governance ultimately determines whether each incident reduces future risk.

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.

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.

Appendix

Incident Information Collection Matrix

Editable table for organizations choosing a harms-only or harms+hazards reporting standard. Use this as a template: harms-only focuses on realized harm; harms+hazards adds hazards, near-misses, unsafe conditions, vulnerabilities, and plausible pathways to harm.

Information Category	Under a harms-only standard, collect…	Under a harms+hazards standard, also collect…
Event classification	Whether the event meets the organization's definition of a reportable incident; type of realized harm; date/time harm occurred; reporting threshold triggered.	Whether the event is an incident, hazard, near-miss, unsafe condition, vulnerability, evaluation finding, or misuse attempt; rationale for why harm was plausible even if not realized.
Plain-language event summary	A concise description of what happened, who or what was harmed, and how the AI system was involved.	A concise description of what could have happened, the causal pathway to harm, and what prevented or limited harm.
Timeline	Time of deployment, detection, harm occurrence, escalation, mitigation, notification, and resolution.	Exposure window before harm occurred or was prevented; how long the hazard existed; when the organization first had reason to know of the hazard.

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

Information Category	Under a harms-only standard, collect…	Under a harms+hazards standard, also collect…
AI system identification	System name, model name/version, vendor or developer, deployment environment, release version, affected feature, API or product surface.	Experimental model/version, pre-release system, internal tool, evaluation environment, red-team target, unreleased capability, or system component that created the hazard.
Use case and context	Intended use, actual use, sector/domain, geography, user population, operational setting, and whether the system was used as designed.	Plausible deployment contexts in which the hazard could materialize; assumptions about scale, user behavior, adversarial use, or environmental conditions.
Affected parties	Individuals, groups, organizations, customers, workers, users, bystanders, or communities actually affected.	Parties that could plausibly be affected; vulnerable groups; downstream users; people indirectly exposed through integrations, automation, or third-party reliance.

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

Information Category	Under a harms-only standard, collect…	Under a harms+hazards standard, also collect…
Nature of harm	Type of actual harm: physical injury, financial loss, rights violation, discrimination, privacy breach, psychological harm, reputational harm, safety degradation, infrastructure disruption, environmental harm, or other impact.	Potential harm type; worst credible harm; foreseeable secondary harms; whether the hazard could contribute to systemic, cumulative, or cascading impacts.
Severity and scale	Severity of realized harm; number of affected people or entities; monetary loss; duration; reversibility; geographic scope; confidence in measurement.	Estimated severity if harm occurred; estimated population exposed; probability or likelihood band; uncertainty; whether the hazard could scale rapidly.
Evidence of harm or hazard	Complaints, logs, screenshots, audit records, model outputs, user reports, medical/legal/financial records, monitoring alerts, or third-party reports showing actual harm.	Red-team results, evaluation failures, vulnerability reports, near-miss records, unsafe outputs, simulations, threat intelligence, misuse attempts, benchmark failures, or internal warnings.

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

Information Category	Under a harms-only standard, collect…	Under a harms+hazards standard, also collect…
Causal role of the AI system	Whether the AI system caused, contributed to, amplified, failed to prevent, or was merely present during the harm.	How the AI system could cause or amplify harm; necessary preconditions; trigger conditions; dependency on human action, automation, integration, or adversarial exploitation.
Human and organizational factors	Human review failures, misuse, poor training, unclear responsibility, inadequate escalation, operator error, or flawed deployment decisions that contributed to harm.	Latent organizational weaknesses: missing controls, inadequate testing, unsafe incentives, incomplete documentation, unclear ownership, or insufficient monitoring.
Technical failure mode	Hallucination, bias, privacy leakage, security compromise, model drift, unsafe autonomy, overreliance, misclassification, prompt injection, tool misuse, data error, or integration failure involved in the harm.	Capability or vulnerability that could lead to harm: jailbreakability, dangerous capability, data exfiltration path, unsafe tool access, excessive autonomy, inadequate guardrails, or robustness failure.

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

Information Category	Under a harms-only standard, collect…	Under a harms+hazards standard, also collect…
Inputs, outputs, and data	Relevant prompts, inputs, outputs, training/fine-tuning data issues, retrieval sources, logs, and data-processing steps linked to the harm.	Inputs or conditions that elicit hazardous behavior; synthetic test cases; adversarial prompts; edge cases; data gaps; trigger patterns; reproducibility instructions where safe to store.
Safeguards and controls	Controls that were in place at the time of harm: human review, monitoring, rate limits, access controls, content filters, escalation paths, fallback procedures.	Controls that failed during testing or could be bypassed; missing controls; safety margins; compensating controls; reasons harm did not occur.
Detection and reporting source	How the organization learned of the harm: user complaint, internal monitoring, audit, media report, regulator, researcher, partner, or employee	How the hazard was discovered: red team, evaluation, bug bounty, employee report, external researcher, simulation, anomaly detection, tabletop exercise, or near-miss report.

RadarFirst

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

Information Category	Under a harms-only standard, collect…	Under a harms+hazards standard, also collect…
Mitigation and remediation	Actions taken to stop ongoing harm, help affected parties, correct outputs/decisions, suspend features, patch systems, notify stakeholders, or provide redress.	Preventive actions: delaying release, narrowing deployment, adding guardrails, changing access levels, improving evaluations, updating safety cases, or revising launch criteria
Recurrence and pattern analysis	Whether similar harms have occurred before; links to prior incidents; frequency; trend; known repeat failure mode.	Whether similar hazards have been observed in testing, competitors' systems, academic literature, red-team exercises, or previous near-misses.
Accountability and ownership	Responsible internal team, system owner, incident commander, vendor contacts, legal/compliance owner, and business owner.	Hazard owner, risk-acceptance authority, release approver, model governance forum, and unresolved risk decision-maker.

PART III - SPECIAL CONSIDERATIONS: CAUSALITY AND SEVERITY

Information Category	Under a harms-only standard, collect…	Under a harms+hazards standard, also collect…
External reporting and notification	Whether users, customers, regulators, insurers, vendors, law enforcement, or the public were notified; dates and content of notifications.	Whether hazard disclosure, coordinated vulnerability disclosure, regulator consultation, partner notification, or pre-deployment risk communication is appropriate.
Confidentiality and sensitivity	Personal data, trade secrets, security-sensitive details, legal privilege, contractual limits, and evidence-preservation needs.	Additional controls for dual-use details, exploit instructions, dangerous capability information, unreleased model details, or vulnerabilities that could enable misuse.
Lessons learned	Root cause, corrective actions, control improvements, policy changes, and post-incident review findings.	Preventive lessons, changes to hazard thresholds, launch gates, safety evaluations, monitoring criteria, and near-miss reporting culture.