verAItech

AI Incident Response: Adapting Proven Complex Systems Engineering Practices for AI-Enabled Systems

Heather Frase, PhD

CEO, Veraitech

November 2025

Please cite as:

Frase, H. (2025). AI Incident Response: Adapting Proven Complex Systems Engineering Practices for AI-Enabled Systems. November 2025. Retrieved from https://veraitechus.com/ai-incident-response/

Preface

Systematic AI incident response requires capabilities that rarely exist in combination within single organizations or disciplines. This white paper draws on years of professional experience in operational testing of complex systems, financial crime enforcement, and AI safety assessment to address this gap. Work as a board member of the AI Incident Database, development of incident taxonomies and reporting frameworks, and analysis of over 1,000 AI incidents demonstrated that proven approaches from other complex systems domains can be adapted to meet AI-specific challenges.

These domains contribute distinct but complementary capabilities. Operational testing and evaluation provides methodologies for assessing complex systems in real-world contexts, understanding system-of-systems interactions, and systematically collecting and analyzing incidents so that known system problems can be tested. Financial crime enforcement demonstrates how structured reporting enables pattern recognition across organizations. It also reveals that an ecosystem of incident collectors, reporters, and response processes is necessary to detect malicious activity, understand systemic harm, and effectively analyze patterns. Direct work with AI incidents through the AI Incident Database, international standards development through European Telecommunications Standards Institute (ETSI) and Organisation for Economic Co-operation and Development (OECD) working groups, and red-team assessments for frontier models, including GPT-4 exposes the specific gaps in current AI incident response approaches.

The framework presented here synthesizes insights from all three domains, adapting proven approaches rather than inventing new methodologies. It emphasizes systems-thinking, context-focused evaluation, data usability, and the recognition that other complex systems domains have already solved many challenges AI incident response now faces.

Executive Summary

As AI-enabled systems integrate into critical applications across defense, financial services, healthcare, and other sectors, organizations face an urgent need for systematic incident response processes. Most lack the frameworks, procedures, and infrastructure to respond effectively when these systems fail or cause harm. This white paper presents a comprehensive framework adapting proven reliability engineering practices from complex systems domains to AI-specific characteristics. The framework provides both a generalizable seven-step process and tailored guidance for different stakeholders, enabling coordinated ecosystem response while allowing customization for specific operational contexts.

AI-enabled systems are complex systems[1] exhibiting characteristics familiar from aerospace, financial services, healthcare, and critical infrastructure: interconnected architectures, context-dependent behavior, cascading failure potential, and nascent system-level interactions. Like these other complex systems domains, AI systems benefit from systematic incident response processes. However, AI-specific characteristics including non-deterministic behavior,[2] adaptive responses, and transitory states require adapting rather than directly applying existing approaches.

Current Gaps in AI Incident Response

Organizations deploy AI-enabled systems to improve efficiency, enable new capabilities, and support critical decisions. Yet when these systems fail or cause harm, most lack systematic processes for responding effectively, learning from incidents, and improving reliability over time.[3]

AI-enabled systems exhibit probabilistic outputs, context-dependent behavior, and complex system-level interactions that complicate incident response. Existing approaches prove inadequate: model testing frameworks focus on pre-deployment validation rather than operational response, while cybersecurity incident response addresses adversarial attacks but not systematic errors, performance degradation, or unintentional harms.

Without systematic approaches, organizations cannot build institutional knowledge about failures, learn from incidents to prevent recurrence, or detect patterns visible only through analysis of multiple incidents. As AI systems become more autonomous and interconnected, these gaps become increasingly critical.

A Framework Grounded in Proven Practices

This white paper presents a systematic framework developed by adapting established methodologies from domains where incident response already works effectively. Rather than inventing new approaches, the framework draws on:

- **Aviation safety** for systematic investigation, identifying root causes in complex systems
- **Financial crime enforcement** for standardized cross-organizational reporting, enabling pattern recognition while protecting proprietary information

---

[1] Meadows, D.H. (2008) Thinking in Systems: A Primer. Chelsea Green, White River Junction
[2] Russell, S. J., & Norvig, P. (2022). Artificial Intelligence: A Modern Approach
[3] Engineering a safer world: Systems thinking applied to safety (Engineering Systems). NG Leveson. Mit Press Cambridge, 2011. 3854, 2011

- **Healthcare adverse event reporting** for blame-free investigation cultures surfacing human factors
- **Cybersecurity incident response**[4] [5] for rapid response protocols, clear escalation paths, and pre-defined containment procedures that enable swift action under pressure
- **Reliability engineering**[6] for tracking improvement over time through quantitative metrics

These proven approaches can be adapted for AI-specific challenges including non-deterministic behavior, context-dependent failures, and system-of-systems interactions. The framework complements existing AI incident and governance frameworks by providing operational detail for implementing the incident response capabilities these standards require.

The Seven-Step Process

The framework centers on seven interconnected steps forming a complete incident response cycle. The process is intentionally generalizable, enabling organizations to adapt severity criteria, investigation methodologies, and verification approaches to their specific contexts. Additionally, organizations may drop reorganize to repeat some of the steps.

1. **Detect**: Identify the incident through monitoring and user feedback
2. **Assess**: Evaluate severity and potential impact using established criteria
3. **Stabilize**: Execute pre-planned procedures to contain harm
4. **Report & Document**: Document incident details using standardized structures and notify stakeholders
5. **Investigate & Analyze**: Determine root cause through systematic analysis
6. **Correct**: Implement solutions to address root causes, reduce recurrence, and mitigate realized harm
7. **Verify**: Test and validate corrections, then monitor for effectiveness

Each step integrates response actions with a required preparedness infrastructure. This reflects a critical insight: effective incident response depends fundamentally on preparation before incidents occur. Organizations cannot respond systematically without pre-planned severity frameworks, stabilization procedures, trained personnel, monitoring infrastructure, investigation capabilities, and verification processes.

The Ecosystem Approach

Organizations can implement incident response independently, but ecosystem coordination unlocks capabilities no single entity can achieve alone. Multiple stakeholders bring complementary capabilities:

- **AI developers** possess deep technical knowledge of model internals but cannot see deployment contexts

---

[4] National Institute of Standards and Technology. Guide for conducting risk assessments. NIST Special Publication 800-30 Revision 1, U.S. Department of Commerce, 2012. URL https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-30r1.pdf.

[5] Cybersecurity & Infrastructure Security Agency (CISA), Incident Response Plan (IRP) Basics

[6] "IEEE Guide for General Principles of Reliability Analysis of Nuclear Power Generating Station Safety Systems," in ANSI/IEEE Std 352-1987 , vol., no., pp.1-118, 1987, doi: 10.1109/IEEESTD.1987.101069.

- **Deployers** understand operational environments but may lack expertise for technical root cause analysis
- **Users** experience real-world system performance but may not recognize AI involvement in decisions
- **Oversight bodies** can aggregate cross-organizational patterns, but depend on incident reports from others for visibility
- **Independent evaluators** provide transparency through public databases and research
- **Assurance organizations** offer verification that complements organizational self-assessment

Coordination enables capabilities impossible for individual organizations. Individual organizations see only incidents affecting their own users or systems. When incidents are aggregated across organizations, patterns become visible that no single entity could detect: systematic errors affecting specific populations, cascading failures propagating through interconnected systems, and sophisticated attacks distributed across multiple targets. This shared visibility accelerates learning about failure modes and spreads effective practices across the field.

Building this ecosystem requires standardized reporting structures.[7] Complex patterns of harm[8] remain invisible without structured, analyzable data that enables computational analysis across organizations. Financial crime enforcement demonstrates this works in practice: Suspicious Activity Reports enable government analysis of patterns across thousands of institutions while protecting the proprietary information of individual financial entities. These standardized structured reports[9] [10] enable collective defense while protecting competitive interests.

Implementation and Integration

Most organizations already have incident management capabilities through IT service management, cybersecurity operations, or risk management functions. Organizations can extend these existing capabilities for AI-specific characteristics rather than building entirely new parallel processes. Existing incident tracking systems, escalation procedures, and communication protocols provide a foundation that can be adapted with AI-specific severity criteria, root cause analysis methodologies, and verification approaches.

Organizations deploying AI systems should focus on detection infrastructure, severity assessment frameworks accounting for AI-specific factors, and standardized incident documentation procedures. Developers need technical response capabilities including rapid rollback mechanisms and model-level root cause analysis. Regulators should define clear reporting requirements while enabling information sharing that protects proprietary information. Standards bodies can develop technical frameworks for incident reporting, while professional organizations can facilitate information sharing.

---

[7] An upcoming companion paper will provide details for designing standardized reporting structures for AI incidents.
[8] Mia Hoffmann and Heather Frase, "Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework" (Center for Security and Emerging Technology, July 2023), https://doi.org/10.51593/20230022.
[9] Ren Bin Lee Dixon and Heather Frase, "AI Incidents: Key Components for a Mandatory Reporting Regime," (Center for Security and Emerging Technology, January 2025), https://doi.org/10.51593/20240023
[10] OECD (2025), "Towards a common reporting framework for AI incidents", OECD Artificial Intelligence Papers, No. 34, OECD Publishing, Paris, https://doi.org/10.1787/f326d4ac-en.

The frameworks and processes presented here can be implemented immediately. Organizations need not wait for perfect solutions or complete ecosystem development. Begin with capabilities appropriate to current systems and contexts. Organizations cannot respond effectively without a preparedness infrastructure established before incidents occur. Participate in ecosystem development by contributing to standardized reporting, sharing lessons learned where appropriate, and coordinating with other stakeholders.

The Path Forward

As AI systems become more capable, autonomous, and integrated into critical functions, the consequences of inadequate incident response increase. Systematic incident response is essential for continuous reliability improvement, regulatory compliance, and demonstrating due diligence. Organizations that implement these capabilities now will be better positioned to deploy AI safely, respond to incidents effectively, and show measurable improvement in system reliability over time.

Section 1: Introduction: Why AI Needs a New Kind of Incident Response

Organizations deploying AI-enabled systems face a fundamental challenge: traditional incident response approaches designed for deterministic software or standalone models prove inadequate when AI systems fail or cause harm. This section establishes why AI requires systematic incident response (1.1), describes what this framework provides to address that need (1.2), and defines core concepts, including what constitutes an AI incident and the range of incidents this framework addresses (1.3).

1.1 The Urgency of AI Incident Response

As AI-enabled systems become increasingly integrated into critical applications across defense, financial services, healthcare, law enforcement, and other sectors, the need for robust incident response processes has never been more urgent. Organizations deploy AI-enabled systems to improve efficiency, enable new capabilities, and support critical decisions. Yet when these systems fail or cause harm, many organizations lack systematic processes for responding effectively, learning from incidents, and improving reliability over time.

The consequences of inadequate incident response extend beyond individual failures. Without systematic approaches, organizations cannot:

- **Build institutional knowledge** about how their AI systems fail and why
- **Learn from incidents** to prevent recurrence
- **Demonstrate improvement** in system improvement to stakeholders and regulators
- **Detect patterns** that emerge only through analysis of multiple incidents
- **Coordinate responses** when incidents cascade across organizational boundaries

These challenges arise because AI-enabled systems are fundamentally complex systems. Like aerospace, financial services, and healthcare systems, they exhibit interconnected architectures, context-dependent behaviors, and complex system-level interactions. However, AI systems introduce novel characteristics, including non-deterministic behavior,[2] adaptive responses, and transitory states that require adapting rather than directly applying existing approaches.

Traditional incident response approaches designed for simpler systems prove inadequate when confronted with these complex system characteristics, particularly in AI systems incorporating agentic behaviors, multi-step reasoning, and tool use.

Software incident response processes assume deterministic behavior and reproducible failures. AI-enabled systems exhibit neither. The same input may produce different outputs. Incidents that occur in production may be impossible to reproduce in testing environments. Failures appear gradually through context-dependent performance degradation rather than manifesting as discrete, reproducible events.

Model testing frameworks focus on pre-deployment evaluation but provide limited guidance for responding to incidents in operational settings. They emphasize validating performance before deployment rather than establishing processes for detecting, analyzing, and correcting failures that emerge during real-world use. Organizations need both pre-deployment testing and operational incident response capabilities.

Cybersecurity incident response effectively addresses adversarial attacks and security breaches but does not cover the full scope of AI incidents. Output quality issues and context-dependent failures, performance degradation, system behavioral concerns, and unintentional harms from system limitations all demand response processes adapted for AI-specific characteristics while building on proven cybersecurity methodologies.

The stakes increase as AI capabilities advance. More capable systems create more consequential failures. More autonomous systems operate with less direct human oversight. More interconnected systems create cascading failure risks across organizational boundaries. Organizations cannot continue responding to AI incidents through ad hoc processes designed for traditional software or standalone machine learning models.

1.2 What This Document Provides

This white paper presents a comprehensive framework for AI incident response that adapts proven reliability engineering practices from complex systems domains to AI-specific characteristics. The framework addresses critical gaps limiting current incident response capabilities while building on methodologies that have proven effective across multiple sectors. It also provides stakeholder-specific recommendations for implementing AI incident response processes.

1.2.1 A Systematic Seven-Step Process

The framework centers on seven interconnected steps forming a complete incident response cycle: Detect, Assess, Stabilize, Report, Investigate, Correct, and Verify. Each step integrates response actions with required preparedness infrastructure. This design reflects a critical insight from mature domains: systematic incident response requires investments made before incidents occur, not capabilities improvised during crises. Section 3.2 details the specific Preparedness Recommendations for each step.

Section 3 provides details for each step. Throughout the section, comparative tables illustrate how each step operates in both existing complex system domains (particularly financial services fraud detection and transaction monitoring) and in agentic AI systems. These side-by-side examples demonstrate how proven approaches from financial crime enforcement, aviation safety, and other established domains can be adapted for AI-specific characteristics, including non-deterministic behavior, context-dependent failures, and system-of-systems interactions.

1.2.2 Integration with Existing Frameworks and Processes

Most organizations already have incident management capabilities through IT service management, cybersecurity operations, or risk management functions. This framework shows how to extend those capabilities for AI-specific characteristics rather than building entirely new parallel processes.

The framework complements existing standards, including NIST AI Risk Management Framework, NIST Cybersecurity Framework, and ISO standards for information security and risk management. Where these frameworks identify what organizations should do, this framework provides operational detail for how to implement incident response capabilities. Section 3.3 addresses integration approaches, showing how organizations can leverage existing processes while adapting them for AI.

1.2.3 Ecosystem Coordination Structures

While organizations can implement effective incident response independently, coordination across multiple stakeholders enables capabilities impossible for individual organizations. Each stakeholder brings distinct capabilities to the ecosystem. No single stakeholder possesses complete visibility or capabilities across the AI value chain. Section 4 describes six stakeholder categories and explains why their coordination is essential for effective incident response. Different stakeholders operate at different system levels; developers address component-level issues, deployers handle system-level problems, and oversight bodies analyze patterns across organizations. This distributed capability structure makes ecosystem coordination necessary rather than optional.

Section 4 clarifies roles and responsibilities across these six stakeholder categories, showing how coordination enables capabilities impossible for individual organizations: pattern recognition across incidents, shared learning about failure modes, and systematic reliability improvement across the field. The framework describes what each stakeholder can and cannot do at each step of the incident response process, explaining when to engage external parties and how to coordinate effectively.

Building this coordinated ecosystem requires infrastructure that does not yet exist in mature form. The framework identifies what needs to be built and how different stakeholders can contribute to ecosystem development.

1.2.4 Standardized Reporting Foundations

Complex patterns of harm remain invisible without structured, analyzable data. Individual organizations cannot detect three critical patterns alone. Section 2 analyzes diffuse harm distributed across organizations, intersectional effects targeting specific demographic combinations, and cascading failures through interconnected systems. The section also examines how sophisticated actors adapt techniques from financial crime, such as structuring and financial mules, to evade detection. Together, these patterns demonstrate why standardized reporting structures[7] represent a foundational requirement for effective AI incident response.[10]

1.2.5 Stakeholder-Specific Recommendations

Section 5 provides concrete action recommendations for each stakeholder group for supporting the AI incident response ecosystem. Rather than generic implementation guidance, this section specifies what AI deployers, AI developers, assurance organizations, standards bodies, regulators, and professional organizations should prioritize.

Deployers should prioritize detection infrastructure, severity assessment, and integration with existing IT processes. Developers should build technical response capabilities and coordinate with deployers. Assurance organizations should develop verification capabilities and audit programs. Standards bodies should harmonize international frameworks and create practitioner resources. Regulators should establish clear requirements, enable information sharing, and coordinate across jurisdictions. Professional organizations should develop information sharing and training programs.

These stakeholder-specific recommendations show how different entities can begin building incident response capabilities and contributing to ecosystem development based on their unique authorities and capabilities.

1.3 Defining AI Incidents

An AI incident occurs when there is harm that can be directly or indirectly linked to the behavior of an AI system.[8] [11] The harm can be experienced by people, organizations, systems, operations, human rights, property, communities, or the environment. This definition, formalized by the Organisation for Economic Co-operation and Development (OECD), provides a foundation for systematic incident response while remaining flexible enough to accommodate diverse deployment contexts and evolving AI capabilities.

Some organizations may also choose to apply their AI incident response process to AI hazards, vulnerabilities, flaws, or near-misses. All of these are conditions that could lead to harm. In contrast, incidents are events that did occur. While this framework focuses primarily on responding to actual incidents where harm occurred, organizations with mature incident response processes often expand to include proactive analysis of conditions that could lead to harm.

1.3.1 Scope of Harm

The breadth of entities that can experience harm reflects the reality that AI systems operate in complex sociotechnical contexts, where incidents can simultaneously have multiple types of harm.

Tangible harm includes observable, verifiable impacts such as physical injury, financial loss, or property damage. A medical diagnosis system providing incorrect recommendations that lead to wrong treatment causes tangible harm. An AI agricultural management system providing incorrect irrigation or fertilization recommendations causes tangible harm through crop failures and financial losses for farmers. An autonomous vehicle system failing to detect obstacles that result in a collision causes tangible harm.

Intangible harm includes impacts that cannot be directly observed but cause real damage. Privacy violations, reputational harm, psychological harm, erosion of trust in institutions, and damage to professional credibility all constitute intangible harms.[8] [11] An AI system trained on customer service conversations inadvertently memorizing and reproducing sensitive personal information causes intangible harm through privacy violations, even when no specific individual can demonstrate direct injury. An AI risk assessment system used in consequential decisions exhibiting systematic performance variations across population segments creates legal liability risks, even when individual detention decisions appear procedurally correct.

Both tangible and intangible harms demand a systematic incident response. Organizations sometimes focus incident response on tangible harms because they are easier to measure and verify. However, intangible harms often prove more widespread and more difficult to detect without structured analysis across many incidents.

---

[11] OECD (2024), "Defining AI incidents and related terms", *OECD Artificial Intelligence Papers*, No. 16, OECD Publishing, Paris, https://doi.org/10.1787/d1a8d965-en.

1.3.2 Incidents, Hazards, and Near-Misses

Incidents are events that **did occur** and **did cause harm**.[8] [12] This distinguishes incidents from related concepts that organizations may also want to track:

Hazards are conditions that **could lead** to harm but have not yet caused actual harm. A vulnerability in an AI system that adversaries could exploit is a hazard. Organizations benefit from finding and addressing hazards before they result in incidents.

Vulnerabilities are specific weaknesses in AI systems that could be exploited or could lead to failures. A prompt injection vulnerability in a customer service chatbot is a vulnerability regardless of whether it has been exploited. Vulnerabilities are a type of hazard, but hazards that do not exploit a system weakness are generally not vulnerabilities.

Near-misses are events where harm almost occurred but was avoided through luck, human intervention, or system safeguards. An autonomous vehicle system that nearly failed to detect a pedestrian but corrected at the last moment is a near-miss.

Some organizations may choose to apply their AI incident response process to hazards, vulnerabilities, and near-misses. Proactive analysis of these conditions can prevent actual incidents. However, this framework focuses primarily on responding to actual incidents where harm occurred, recognizing that organizations with mature incident response processes often expand to include proactive hazard analysis.

1.3.3 Security and Safety Dimensions

AI incidents encompass both security and safety dimensions, and practitioners must recognize that many incidents involve both simultaneously. Traditional organizational structures often separate security teams focused on adversarial threats from quality and reliability teams focused on performance and safety. AI incidents frequently require coordinated actions across these organizational functions.

Security-related incidents involve harm resulting from the deliberate exploitation of system vulnerabilities,[13] including:

- **Adversarial attacks** where malicious actors manipulate inputs to cause misclassifications or bypass safety constraints
- **Data poisoning** where attackers inject malicious data into training sets to compromise model behavior
- **Unauthorized access** to AI systems, models, or training data
- **Prompt injection attacks** where users embed malicious instructions in prompt text to bypass system guardrails
- **Model extraction** where adversaries query systems to reconstruct proprietary models

---

[12] Heather Frase and Owen Daniels, "Understanding AI Harms: An Overview," Center for Security and Emerging Technology, August 11, 2023, https://cset.georgetown.edu/article/understanding-ai-harms-an-overview/.

[13] Two good resources for vulnerabilities are MITRE's Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS™) and ATT&CH®. These can be found respectively at https://atlas.mitre.org/ and https://attack.mitre.org/

Safety-related incidents involve unintentional harms:

- **Performance failures** where systems fail to meet functional or non-functional requirements, causing incorrect decisions, denied services, or degraded outcomes
- **Performance inconsistencies** where systems exhibit systematic variations across user populations, causing disparate service quality or outcomes that may trigger regulatory scrutiny
- **Unexpected system behaviors** in contexts not adequately represented in testing, causing harm through unanticipated actions or decisions
- **Human factors issues** where inadequate training, confusing interfaces, or misaligned incentives lead to harmful system use or misinterpretation of outputs

Many real-world incidents do not fall neatly into security or safety categories but represent hybrid events demanding coordinated response. A prompt injection attack (security) that causes an AI customer service system to provide harmful or dangerous recommendations (safety) requires both security containment and output quality safeguards. An adversarial attack (security) that degrades model performance (safety) causing service denial for legitimate users requires both security hardening and reliability improvement.[14]

Effective incident response requires bridging these traditional organizational divides. Cybersecurity teams bring expertise in rapid containment, forensic investigation, and coordination with security operations centers. AI safety and reliability teams bring expertise in quality assurance, fairness analysis, and operational performance. Both perspectives are essential for comprehensive incident response.

1.3.4 Complexity Range

AI incidents range from straightforward cases involving a single system and affected entity to complex, multifaceted events with cascading effects across systems and multiple affected populations.[12]

Simple incidents involve:

- Single AI system
- Clear failure mode
- Limited affected population
- Contained impact
- Straightforward root cause

Complex incidents involve:

- Multiple interconnected AI systems
- Cascading failures across system boundaries
- Large or diverse affected populations

[14] Depending upon specific incident characteristics, organizations may classify misuse as a security or safety incident. Misuse occurs when someone uses an AI system for harmful purposes, which may or may not involve exploiting vulnerabilities. A user employing a text generation system as designed to create phishing emails represents misuse without security compromise (primarily a safety incident). An attacker exploiting prompt injection to bypass guardrails represents both security attack and misuse. This distinction matters for incident response: pure misuse may require strengthening usage policies and guardrails rather than patching security vulnerabilities, while security-enabled misuse requires both security remediation and safety improvements.

- Systemic impacts
- Multiple contributing factors
- Novel behaviors from system interactions

Organizations should prepare incident response processes capable of handling both simple and complex incidents. Response procedures appropriate for simple incidents may prove inadequate when confronted with cascading failures across organizational boundaries or intersectional harms affecting specific demographic combinations.

This framework addresses the full complexity range, providing structured approaches that scale from straightforward single-system failures to complex system-of-systems incidents requiring coordination across organizational boundaries.

Section 2: The Challenge: Fragmented Capabilities and Missing Foundations

Effective AI incident response requires foundational structures that most organizations currently lack. This section first explains why AI-enabled systems present qualitatively different challenges from traditional software, requiring adapted approaches to proven incident response methodologies (2.1). It then identifies six critical missing structures that limit organizations' ability to respond systematically: severity classification frameworks, escalation criteria, investigation capabilities, standardized reporting structures, reliability metrics, and clear delineation of responsibilities (2.2). Addressing these gaps is essential for moving from ad hoc incident handling to systematic reliability improvement.

2.1 AI-Enabled Systems as Complex Systems

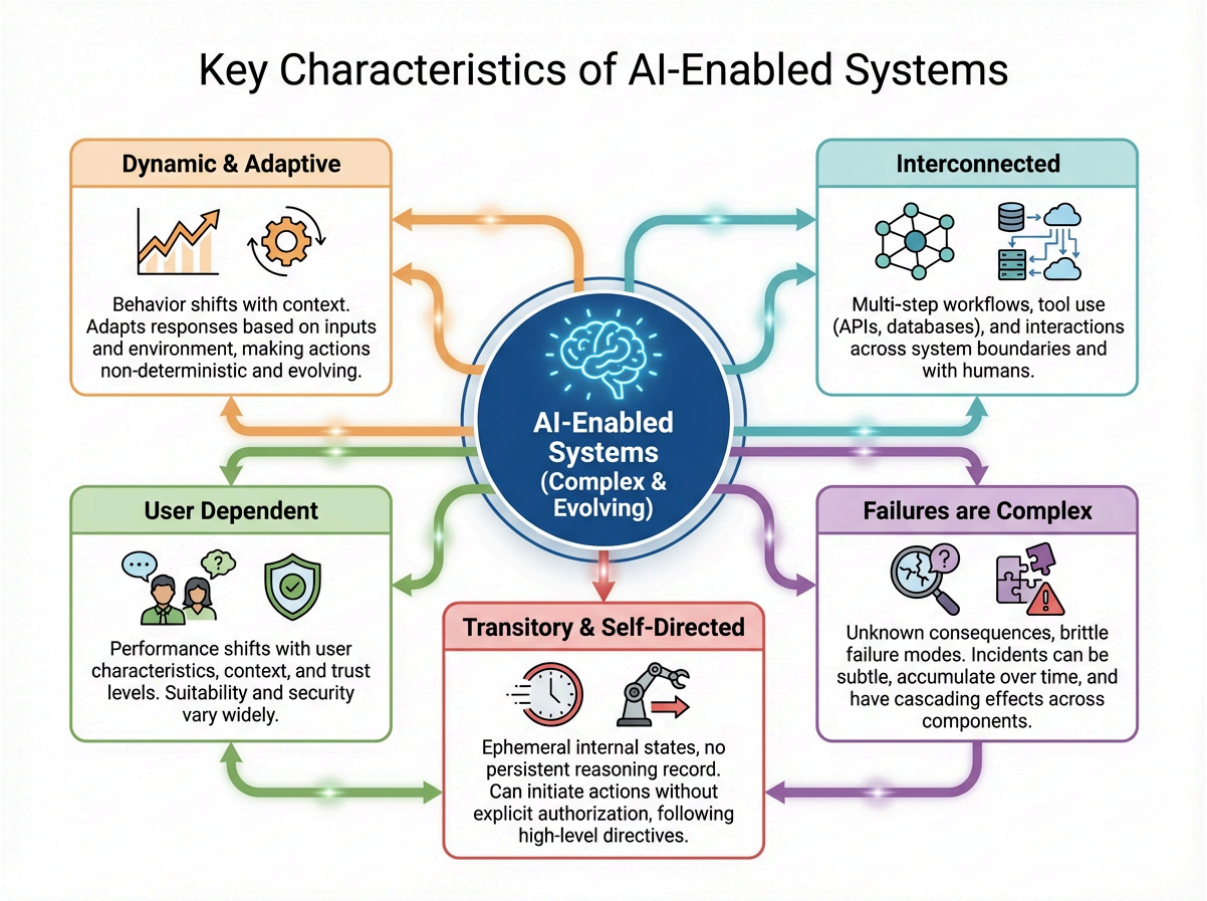


Figure 1: AI-enabled systems exhibit characteristics of complex systems with novel attributes.

AI-enabled systems represent a qualitatively different challenge from traditional software engineering or standalone machine learning models. They are complex systems that share characteristics with other complex systems in aerospace, healthcare, financial services, and critical infrastructure domains, while introducing novel attributes that require adapted approaches. Like

these other complex systems, AI-enabled systems can benefit from established incident response frameworks that have proven effective across multiple domains.

Understanding AI-specific characteristics is essential for adapting the execution of proven response processes. Traditional software incident response often assumes deterministic behavior, reproducible incidents, and clear system boundaries. AI-enabled systems often violate these assumptions. Their behavior shifts with context, incidents may be non-reproducible, and their boundaries extend across multiple interconnected components. These characteristics require adapting how we execute each step of the incident response process. The overall incident response framework remains constant. The specific techniques and tools for executing each step should account for AI's adaptive behavior, context-dependence, and uncertainty.

AI-enabled systems exhibit several characteristics that distinguish them from traditional software:

- Dynamic and adaptive behavior
- Interconnected architectures
- Complexity with unknown consequences
- User-dependent performance
- Transitory and self-directed operations
- Distributed and tool-mediated autonomy
- Novel failure modes

Dynamic and Adaptive

The functionality of AI-enabled systems shifts with context. Unlike traditional software with deterministic behavior, AI-enabled systems adapt their responses based on input variations, environmental conditions, and learned patterns. A financial audit assistant may perform adequately on standard transactions but fail unpredictably when encountering novel transaction structures. This adaptability creates challenges for incident response because the same system may exhibit different behaviors under similar conditions, complicating both incident reproduction and root cause analysis.

Interconnected

Modern AI-enabled systems, particularly agentic systems, exhibit multi-step behaviors and extensive tool use. They interact with databases, external APIs, other AI-enabled systems, and human users in complex workflows. An AI system designed to assist with financial analysis might query internal databases, fetch market data from external sources, invoke calculation engines, and coordinate with other specialized AI agents.

This interconnectedness means that failures and incidents can cascade across system boundaries, and root causes may lie far from where symptoms first appear. Consider an agentic customer service system using Model Context Protocol (MCP) to access multiple tools: a product database, order tracking system, refund processing system, and knowledge base. If the order tracking system experiences delays, the agent might make decisions based on incomplete information, leading to incorrect promises to customers. The incident manifests as customer service failures, but the root cause spans multiple systems and their interactions.

Complex with Unknown Consequences

AI-enabled systems can produce incidents with unknown consequences and exhibit brittle failure modes. Unlike traditional software, where edge cases can be systematically detailed and tested, AI-enabled systems can create incidents in unexpected ways when encountering novel inputs or combinations of conditions. These incidents may manifest subtly and develop gradually rather

than catastrophically. While traditional software incidents are typically discrete events with clear onset, AI-enabled system incidents can accumulate over time, with harm accruing across multiple user interactions before triggering incident response.

User Dependent

The suitability of AI-enabled systems shifts with real-world context and user characteristics. A system that performs adequately for expert users with deep domain knowledge may fail dramatically when deployed to novice users who lack context to recognize problematic outputs. Security properties may differ substantially between trusted users (authenticated employees with legitimate access) and untrusted users (external actors potentially attempting adversarial attacks). An AI legal research assistant that works well for experienced attorneys who can evaluate outputs critically may create incidents when used by pro se litigants who treat its outputs as authoritative legal advice.

Transitory and Self-Directed

Unlike other complex systems, AI-enabled systems (particularly generative systems and agentic AI) can be transitory and self-directed. Their internal states may be ephemeral, with no persistent record of reasoning processes that led to particular outputs. This makes post-incident forensics challenging: when an AI system's behavior results in harm, investigators may struggle to reconstruct why. AI-enabled systems can initiate actions without explicit human authorization for each step, following high-level directives through multi-step reasoning processes that may take unexpected paths.

Distributed and Tool-Mediated Autonomy

Agents increasingly rely on tools, APIs, and Model Context Protocols (MCPs) to perform actions. Both the agent and its tools may exhibit non-deterministic behavior, making the system's overall operation difficult to predict or audit. When an agentic system has access to email tools, database tools, and external API tools, the combinations of tool uses and sequences of operations grow exponentially. This distributed autonomy means that monitoring should extend beyond the agent itself to encompass the entire ecosystem of tools it can invoke.

Novel Failure Modes

These characteristics represent genuinely new challenges for incident response. New types of incidents continue to occur as AI capabilities expand. Prompt injection attacks (security attacks where malicious users embed instructions in prompt text) have no direct analog in traditional software. Data leakage, hallucinations, sycophancy, and agent handoff failures in multi-agent workflows exemplify failure modes that require incident response processes designed explicitly for AI-enabled systems.

2.2 Missing Foundational Structures

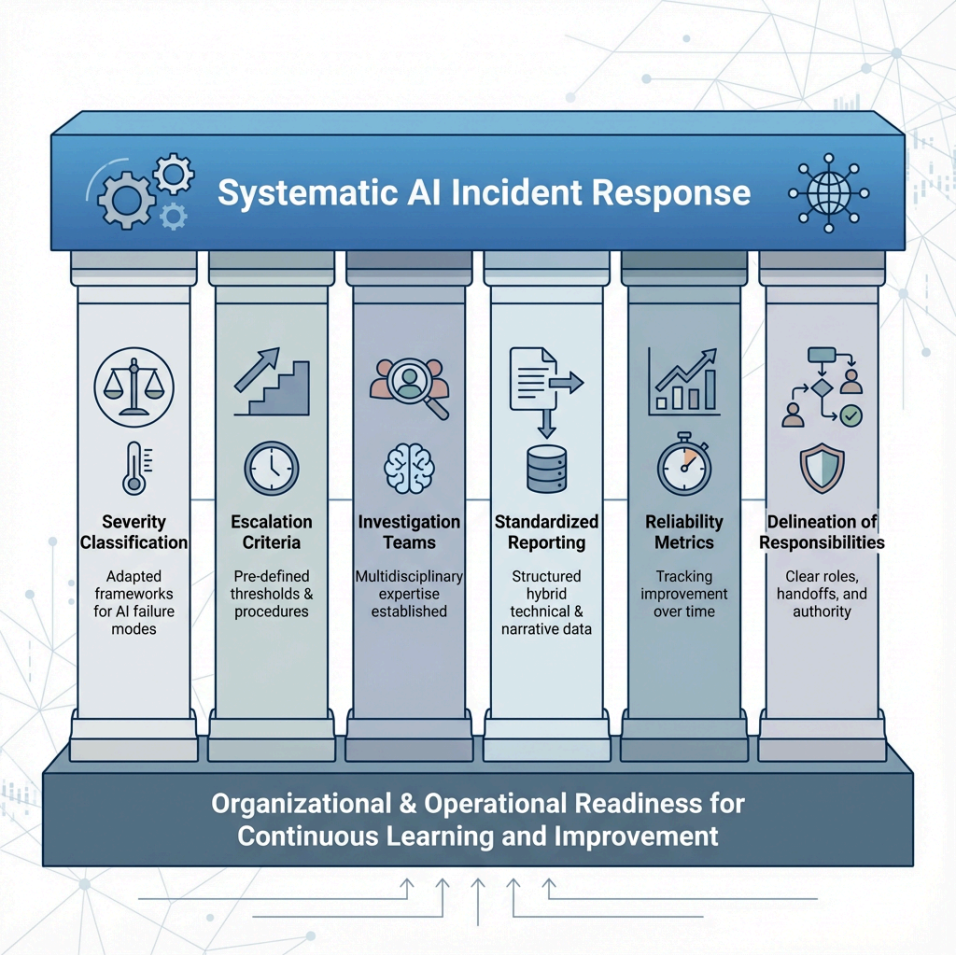


Figure 2: Six commonly missing foundational structures that organizations need to support systematic AI incident response.

Despite the critical importance of effective incident response, most organizations currently lack several foundational structures that enable systematic response and continuous improvement. These gaps limit organizations' ability to respond consistently, learn from incidents, and build institutional knowledge. Addressing these missing structures is essential for moving from ad hoc incident handling to systematic reliability improvement.

2.2.1 Severity Classification Systems

The Gap

Organizations lack severity classification frameworks tailored to their operational contexts and AI-specific failure modes. Traditional software severity classifications based on system availability and data integrity prove insufficient for AI-enabled systems, for which subtle performance degradation, harmful or fabricated outputs, or privacy leakage can cause serious harm without catastrophic failure.

What's Needed

Organizations should develop severity classification frameworks by:

- Adapting established standards (such as MIL-STD-882E) that define severity levels: catastrophic, critical, marginal, and negligible
- Customizing severity definitions for their specific operational contexts and deployment scenarios
- Creating scoring criteria that account for AI-specific incident characteristics
- Developing assessment templates enabling consistent severity evaluation
- Training personnel on applying criteria to AI incidents

Section 3.2.2 (Assess) provides guidance on adapting established severity frameworks for AI-enabled systems.

2.2.2 Escalation Criteria and Response Procedures

The Gap

Clear escalation criteria and response procedures for different incident types are underdeveloped. Organizations struggle to decide when AI incidents warrant executive notification versus routine handling, when to engage external experts, when regulatory reporting becomes mandatory, and who has the authority to make containment decisions. Without pre-established criteria, escalation decisions become ad hoc, potentially delaying critical responses or creating unnecessary alarm about minor issues.

What's Needed

Pre-defined escalation frameworks that account for:

- Incident severity level
- Stakeholder impact (customers, employees, partners)
- Regulatory requirements and mandatory reporting thresholds
- System criticality and business dependencies
- Potential for escalation or cascading harm

These frameworks enable consistent decision-making under pressure, when time for deliberation is limited and consequences of delay may be significant.

2.2.3 Investigation Teams and Resources

The Gap

Effective investigation of AI incidents requires specialized expertise beyond traditional software debugging. AI incidents may be non-reproducible, have root causes in training data or fine-tuning processes far removed from deployment, or arise from complex interactions between components that function correctly in isolation. Organizations typically lack pre-established investigation capabilities with appropriate multidisciplinary teams.

What's Needed

Investigation teams with:

- **Data science expertise**: Understanding model architectures, training processes, fine-tuning approaches
- **Domain expertise**: Deep knowledge of the application area where AI operates

- **Operational expertise**: Understanding deployment contexts, user populations, real-world constraints
- **Human factors expertise**: Analyzing how users interact with systems, training adequacy, interface design
- **Systems engineering expertise**: Understanding component interactions, system-level behaviors, cascading effects

These capabilities should be established before deploying AI systems, not assembled after incidents occur.

2.2.4 Standardized Technical Reporting Structures

The Gap

Individual organizations responding to AI incidents within their own systems cannot detect specific critical patterns. These patterns become visible only through aggregation and analysis across organizations.

The European Telecommunications Standards Institute (ETSI) and OECD are developing common reporting standards for AI incidents.[10] Research organizations have proposed hybrid reporting frameworks combining structured technical specifications with narrative descriptions. However, broad adoption remains limited. Without standardized reporting, incident data remains siloed, preventing identification of patterns across organizations, systems, or deployment contexts.

What's Needed

Hybrid frameworks combining:

- **Structured technical specifications**: Prescribed fields, controlled vocabularies, consistent formats enabling computational analysis
- **Narrative descriptions**: Free-text context capturing circumstances, nuances, and novel factors
- **Privacy and security protections**: Multiple report versions with appropriate access controls
- **Cross-organizational compatibility**: Enable aggregation and pattern recognition while protecting proprietary information

Why This Gap is Particularly Critical

This gap affects the ecosystem's ability to learn collectively from incidents. Complex patterns of harm and misuse remain invisible when analyzing individual incidents in isolation. Diffuse harms distributed across many users and organizations become visible only through statistical analysis of aggregated datasets. Intersectional effects targeting specific demographic combinations require cross-tabulation of multiple variables. Cascading failures propagating through interconnected systems demand data capturing, timing relationships, and system interdependencies across organizational boundaries. Sophisticated misuse patterns deliberately obscure detection through techniques like structuring (breaking malicious objectives into many innocuous-appearing requests distributed across time and accounts).

Proven Approaches from Other Domains

Financial crime enforcement demonstrates that standardized reporting enabling pattern recognition works in practice. FinCEN (Financial Crimes Enforcement Network) aggregates Suspicious Activity Reports from financial institutions using highly structured formats

combining prescribed fields with controlled vocabularies alongside free-text contextual information. This hybrid structure enables pattern recognition across institutions, identifying money laundering networks and emerging threat patterns impossible to detect from individual reports. This approach from financial crime enforcement provides a proven model that can be adapted for AI incident response.

2.2.5 Metrics and Tracking

The Gap

Metrics and methodologies for tracking system improvement over time need adaptation for AI-enabled systems. Traditional software performance and reliability metrics (mean time between failures, defect density) assume deterministic behavior and may not translate directly to probabilistic AI systems. Organizations lack metrics enabling comparability across systems or system versions, methodologies for assessing whether corrective actions actually improve systems, and infrastructure for tracking metrics over time.

What's Needed

Organizations should:

- Adopt and adapt traditional reliability, cybersecurity, and incident metrics for AI contexts (Mean Time Between Incidents, Fix Effectiveness Rate, Mean Time to Respond)[15] [16] [17]
- Develop custom metrics specific to their AI systems, operational use, and sector
- Set up baseline measurements before implementing corrective actions
- Track trends over time to demonstrate improvement
- Report on reliability to stakeholders (internal leadership, regulators, users)

Without these capabilities, organizations cannot demonstrate whether their incident response efforts produce measurable improvements.

Reliability and safety in the context of incidents

Reliability in AI systems represents a system's ability to consistently perform its intended function under varied operational conditions, focusing on predictable and consistent performance. Safety, while related, addresses the potential for harm and unintended consequences that may arise from system operation. These concepts are distinct yet interconnected: a system can be reliable without being safe, and conversely, a system might be considered safe even with reliability issues.

An AI incident can manifest as a reliability issue, a safety concern, or a complex combination of both. In AI incident response, recognizing these subtle yet important differences is useful. The goal is not just to respond to incidents, but to build AI systems that are both dependably performant and fundamentally safe across diverse operational environments.

---

[15] Marvin Rausand and Arnljot Høyland. System Reliability Theory: Models, Statistical Methods and Applications. Wiley-Interscience, Hoboken, NJ, 2004.
[16] Hubbard, D.W., & Seiersen, R. (2023). How to Measure Anything in Cybersecurity Risk (2nd Edition). Wiley.
[17] Hubbard, D. W. (2007). How to measure anything: Finding the value of "intangibles" in business. John Wiley & Sons.

2.2.6 Delineation of Responsibilities

The Gap

Clear delineation of responsibilities within an organization's incident response process often remains underdeveloped. Many organizations lack clarity on who detects incidents, who assesses severity and authorizes escalation, who executes stabilization procedures, who leads root cause investigations, who implements corrective actions, and who verifies correction effectiveness. Without pre-defined handoffs between technical teams and operational teams, critical steps may be delayed or omitted entirely. Responsibility may shift between organizations in system-of-systems contexts, making pre-established coordination protocols essential.

What's Needed

Pre-established responsibility matrices defining:

- Roles across the complete response process
- Handoff procedures between teams
- Cross-organizational coordination protocols for system-of-systems
- Authorization levels for different response actions
- Escalation paths when primary responders are unavailable

Section 3: The AI Incident Response Framework

This section presents a systematic framework implementing AI Incident response in a manner that gaps discussed in Section 2. The framework builds on three foundations. First, it adapts established methodologies from cybersecurity incident response, systems engineering, reliability engineering, and failure mode analysis. These disciplines have decades of proven effectiveness in complex systems domains (3.1). Second, it presents a seven-step incident response process integrating response actions with required preparedness infrastructure (3.2). Third, it provides guidance for integrating AI incident response with existing organizational processes rather than building parallel systems (3.3).

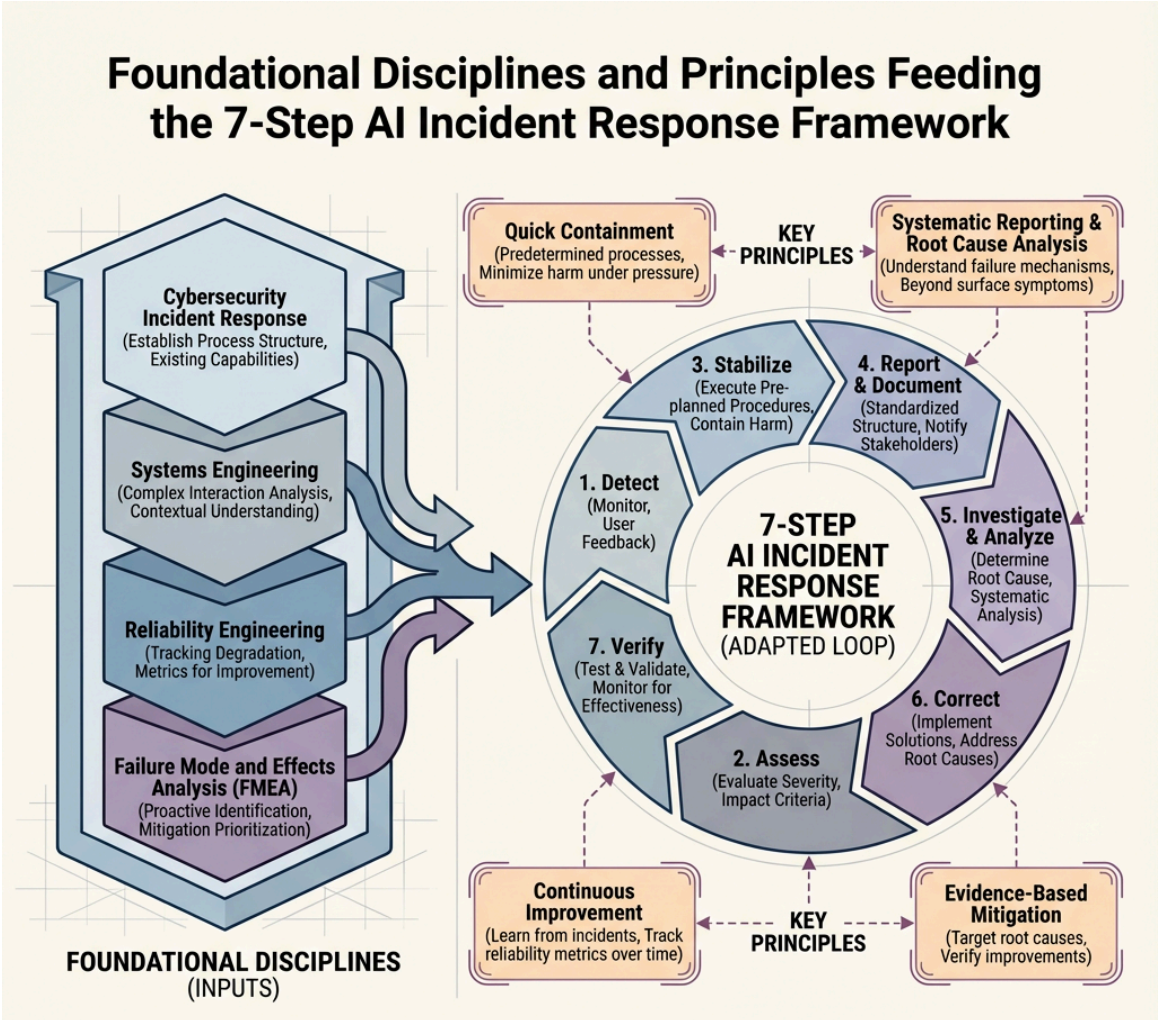


Figure 3: The proposed AI Incident Response Framework adapts proven methodologies and core principles from established disciplines.

3.1 Principles: Leveraging Proven Approaches

This framework reuses proven approaches from cybersecurity, systems engineering, reliability engineering, and failure mode analysis rather than inventing new methodologies. These domains have addressed incident response in complex systems for decades. The overall incident response structure remains constant across domains. The specific techniques for executing each step should be adapted for AI's non-deterministic behavior, context-dependence, and adaptive characteristics.

3.1.1 Foundation in Established Disciplines

Rather than inventing entirely new incident response methodologies, this framework adapts established processes from other complex systems domains. Four disciplines provide particularly relevant foundations:

Cybersecurity Incident Response[4] [5] [18] [19] provides the fundamental structure for the incident response loop. Frameworks like the NIST Cybersecurity Framework and ISO/IEC 27035 provide proven processes: preparation, detection and analysis, containment, eradication, recovery, and post-incident activities. Most organizations already have cybersecurity incident response capabilities with established infrastructure, trained personnel, escalation procedures, and reporting systems. This existing foundation provides enormous value, enabling organizations to extend and adapt what they already have rather than build entirely new processes.

However, cybersecurity incident response typically focuses on adversarial threats, such as malicious actors, data breaches, unauthorized access, and deliberate exploitation of vulnerabilities. AI incidents encompass a broader scope. Unexpected behaviors, performance failures, quality degradation, and unintentional harms from system limitations all demand incident response but often fall outside traditional cybersecurity frameworks. Many real-world AI incidents involve both security and safety dimensions simultaneously, requiring response coordination across organizational functions that traditionally operate separately.

This framework extends cybersecurity incident response to address AI's full incident spectrum while leveraging the proven loop structure and existing organizational capabilities that cybersecurity provides.

Systems Engineering[1] [3] [20] [21] emphasizes understanding how components interact to create system-level behaviors. It provides methodologies for analyzing complex interactions, understanding operational contexts, and recognizing that incidents often arise from interactions between individually functional components rather than from isolated failures. Systems engineering teaches us to look beyond component-level problems to system-level patterns, which is particularly valuable for AI systems, where failures may arise from context, deployment environments, or user characteristics rather than from model defects alone.

Reliability Engineering[22] [6] [23] provides methodologies for tracking system reliability over time, identifying degradation patterns before they cause incidents, and assessing whether corrective actions actually improve system performance. Reliability engineering gives us the metrics and analytical approaches to show measurable improvement rather than simply responding to individual incidents. For AI systems with non-deterministic behavior and probabilistic outputs, reliability engineering's focus on distributional shifts and statistical measures of improvement proves especially valuable.

---

[18] UK National Cyber Security Centre (NCSC), Incident Management
https://www.ncsc.gov.uk/collection/incident-management
[19] Carnegie Mellon University, Incident Management
https://www.cisa.gov/sites/default/files/c3vp/crr_resources_guides/CRR_Resource_Guide-IM.pdf
[20] Blanchard, B. S., & Fabrycky, W. J. (2011). Systems Engineering and Analysis (5th ed.)
[21] Erik Hollnagel, David D. Woods, Nancy Leveson, Resilience Engineering: Concepts and Precepts. Ashgate Publishing, Ltd., 2007. ISBN 978-0-754-68136-6.
[22] James McLinn. A short history of reliability. The Journal of Reliability Information, pages 8–15, 01 2011.
[23] Ebeling, C. E. (1997). An Introduction to Reliability and Maintainability Engineering

Failure Mode and Effects Analysis (FMEA)[24] [25] [26] provides a systematic approach to identifying potential failure modes, assessing their severity and likelihood, and prioritizing mitigation efforts. Initially developed for complex engineered systems, FMEA provides structured methods for anticipating how systems might fail and preparing responses before incidents occur. FMEA's emphasis on proactive analysis complements cybersecurity's reactive incident response, helping organizations prepare for both adversarial attacks and unintentional failures.

These disciplines have proven effective across aerospace, healthcare, financial services, critical infrastructure, and cybersecurity. They share common characteristics with AI-enabled systems: complexity, interconnectedness, context-dependency, and the potential for cascading failures across organizational boundaries.

3.1.2 Core Principles

Four principles adapted from these mature domains guide the framework:

1. **Quick containment using predetermined processes to minimize harm.** Time pressure during active incidents makes careful deliberation impractical. Organizations should establish stabilization procedures before deploying systems, enabling rapid response without complex decision-making under stress.
2. **Systematic reporting and root cause analysis to understand incidents and failure mechanisms.** Learning from incidents requires moving beyond surface symptoms to understand why failures occurred. Standardized reporting structures and systematic analysis methodologies enable this deeper understanding.
3. **Evidence-based mitigation to reduce recurrence and realized harm.** Corrective actions should target root causes identified through systematic analysis, not just address visible symptoms. Verification confirms that corrections actually improve reliability rather than introducing new problems.
4. **Continuous improvement based on real incidents and reliability metrics.** Each incident provides information about how systems fail in operational contexts. Organizations should track reliability metrics over time to demonstrate whether incident response efforts produce measurable improvements in system performance.

3.1.3 Value of this approach

These established frameworks work because they address challenges that AI incident response now faces. They provide:

- **Proven methodologies** tested across decades and multiple domains
- **Structured processes** that function under pressure without requiring improvisation
- **Existing organizational capabilities** that can be extended rather than replaced
- **Emphasis on preparation** before incidents occur rather than reactive response alone
- **Measurement frameworks** enabling demonstration of improvement over time
- **Recognition of complexity** in interconnected systems where root causes may be distant from symptoms

---

[24] FEMA has identified different function levels for its operations: Primary MIssion Essential Functions, Mission Essential Function, and Essential Supporting Activities. For FEMA, a failure's severity level could be determined by which of these function levels was impacted.
https://www.fema.gov/sites/default/files/2020-07/Federal_Continuity_Directive-2_June132017.pdf

[25] McDermott, R. E., et al. (2009). The Basics of FMEA

[26] Stamatis, D. H. (2003). Failure Mode and Effect Analysis: FMEA from Theory to Execution

The framework adapts these proven approaches for AI-specific characteristics. The overall incident response structure remains constant across domains. The specific techniques and tools for executing each step should account for AI's adaptive behavior, non-determinism, context-dependence, and uncertainty. Organizations need not invent entirely new processes. Instead, they can build on what already works, adapting execution details for AI's unique attributes.

3.2 The Seven-Step Incident Response Loop

The framework consists of seven interconnected steps forming a complete incident response cycle. Each step description integrates both the response action and the preparedness infrastructure required to execute it effectively. A critical insight from mature incident response domains: systematic response depends fundamentally on preparation before incidents occur.

The seven steps are:

1. **Detect**: Identify the incident through monitoring and user feedback
2. **Assess**: Evaluate severity and potential impact using established criteria
3. **Stabilize**: Execute pre-planned procedures to contain harm
4. **Report & Document**: Document incident details using standardized structures and notify stakeholders
5. **Investigate & Analyze**: Determine root cause through systematic analysis
6. **Correct**: Implement solutions to address root causes, reduce recurrence, and mitigate realized harm
7. **Verify**: Test and validate corrections, then monitor for effectiveness

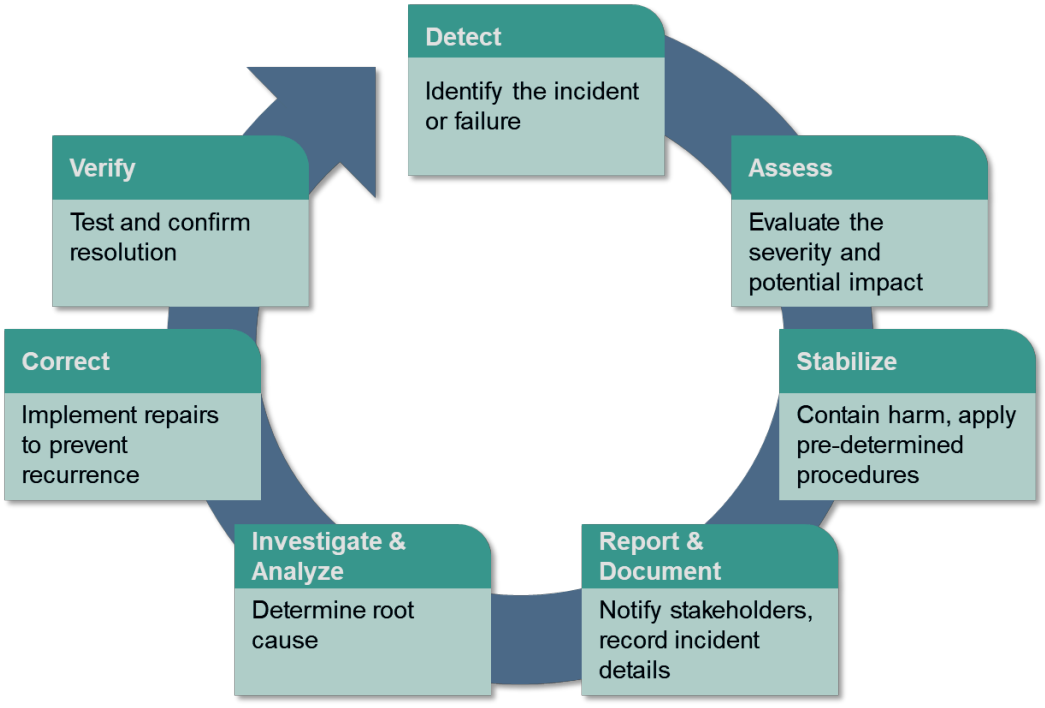


Figure 4: The circular process of the seven-step incident response loop

Critical Point: Preparedness Enables Response

Organizations cannot respond effectively to AI incidents without a preparedness infrastructure established before deployment. Each of the seven steps requires specific preparedness investments, described in the subsections below. The Preparedness Recommendations range from severity classification frameworks and monitoring systems to multidisciplinary investigation teams and verification processes. Organizations that attempt incident response without this foundation face reactive firefighting rather than systematic reliability improvement. This preparedness-focused approach moves organizations from reactive crisis management to systematic reliability improvement.

The loop is customizable

The specific ordering and naming conventions for this framework will vary across sectors and organizations; however, the fundamental intent, goal, and approach of this incident response cycle are generalizable.

This seven-step loop can be easily customized. For example, an organization might want multiple Investigate & Analyze steps. A cybersecurity loop might rename the Assess step as Containment, while replacing the Correct step with two steps, Eradicate and Recovery.

3.2.1 Step 1: Detect

Action: *Identify the incident through system monitoring and user feedback.*

AI system incidents often develop gradually through subtle performance degradation or context-dependent failures rather than manifesting as discrete events. No single detection approach can capture all incidents. Organizations need multiple complementary detection mechanisms working in concert. Additionally, some incidents are easier to detect when data is shared across organizations.

Multiple Complementary Detection Mechanisms

Automated Technical Monitoring provides continuous surveillance of system behavior, tracking response times, throughput, output formats, resource consumption, authentication patterns, and access behaviors. Automated monitoring excels at detecting deviations from established baselines and identifying anomalies in technical metrics. It provides objective, real-time data about system performance.[27] [28] However, automated monitoring may miss

[27] Tripathi, J., Gomes, H., Botacin, M. (2025). "Towards Explainable Drift Detection and Early Retrain in ML-Based Malware Detection Pipelines." In: Egele, M., Moonsamy, V., Gruss, D., Carminati, M. (eds) Detection of Intrusions and Malware, and Vulnerability Assessment. DIMVA 2025. Lecture Notes in Computer Science, vol 15748. Springer, Cham

[28] Leest, J., Gerostathopoulos, I., and Raibulet, C. (2023). "Expert Monitoring: Human-Centered Concept Drift Detection in Machine Learning Operations." In Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results

context-dependent failures, subtle quality degradation, or incidents whose symptoms appear only in specific user populations or use cases.

User Reports flow through help desks, feedback mechanisms, and surveys, identifying problems in real-world contexts that automated monitoring cannot capture. Users experience the actual consequences of AI system behavior. They detect when outputs are unhelpful, incorrect, or harmful in ways that technical metrics may not reveal. Users provide essential signals[29] [30] about how AI systems perform in operational environments with real-world variation and complexity. However, user reporting may introduce detection delays, depend on users recognizing problems and taking action to report them, and may produce inconsistent or incomplete information.

Internal Organizational Functions provide additional detection capabilities. Quality assurance teams identify issues during routine testing, often catching problems before they affect production users. Compliance officers discover problems during audits, particularly when AI system behavior creates regulatory concerns. Security teams detect anomalies during security assessments and penetration testing. Data scientists notice performance degradation during model monitoring and drift analysis. Each function contributes distinct perspectives and detection capabilities.

System-of-Systems Detection requires coordination across organizational boundaries. In deployments where AI systems interact across organizational boundaries, detection may extend beyond any single organization. Incidents may cascade through connected systems, with root causes manifesting far from initial indicators. A failure in one organization's AI system may trigger failures in partner organizations' systems that depend on its outputs or coordinate with its operations.

Effective detection in system-of-systems environments requires coordination between organizations to identify and trace incidents as they propagate. This coordination demands pre-established communication channels, shared understanding of system interdependencies, and agreements about how organizations will share incident information while protecting proprietary details.

Applying Detection Across Domains

Table 1 illustrates how detection operates in both a mature complex system domain (financial services transaction monitoring) and in agentic AI systems. Both domains face similar challenges: gradual degradation rather than discrete failures, multiple signal sources providing complementary information, and the need for both automated and human detection mechanisms.

---

[29] Senarath, Y., Mukhopadhyay, A., Vazirizade, S.M., Purohit, H., Nannapaneni, S., and Dubey, A. (2024). "Designing a Human-centered AI Tool for Proactive Incident Detection Using Crowdsourced Data Sources to Support Emergency Response." Digital Government: Research and Practice, Vol. 5, No. 1

[30] Y. Senarath, A. Mukhopadhyay, S. M. Vazirizade, H. Purohit, S. Nannapaneni and A. Dubey, "Practitioner-Centric Approach for Early Incident Detection Using Crowdsourced Data for Emergency Services," 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 2021, pp. 1318-1323, doi: 10.1109/ICDM51629.2021.00164.

Table 1: Comparative Example: Detection in Financial Services and Agentic AI. Effective incident detection for complex systems often requires multiple complementary approaches.

Comparative Example for Step 1 Detection in Financial Services and Agentic AI	
Financial Services: Fraud Detection System	**Agentic AI: Customer Service Agent - International Expansion**
Automated Monitoring: Transaction monitoring system flags unusual pattern, a customer makes 15 cash withdrawals of $9,500 each over two weeks (below $10,000 reporting threshold). Pattern deviates from customer's historical baseline of 2-3 monthly withdrawals averaging $500. System generates automated alert based on velocity and amount thresholds.	**Automated Monitoring:** System tracks agent performance metrics including resolution rates, average handling time, and tool usage patterns. Monitoring flags that agent's fallback-to-human rate in German market operations increased to 15% vs. 2% baseline in US market operations. Quality scores for German market interactions declined to 3.2 vs. 4.1 baseline (out of 5). Automated alerts trigger based on deviation from established baselines and cross-market performance comparison.
User Reports: Branch manager receives customer complaint about account access issues after fraud prevention system blocked legitimate transactions during travel. Customer reports calling help desk three times with unresolved problems. Other customers report similar experiences with blocked cards.	**User Reports:** German market customers submit feedback through post-interaction surveys reporting unhelpful responses and confusion about AI outputs. Help desk receives increased volume of complaints from German market about needing multiple interactions to resolve simple issues. Feedback mentions AI "not understanding" German address formats and product inquiries requiring human escalation.
Internal Functions: Quality assurance review of fraud alerts identifies false positive rate increase from 8% to 22% for specific customer segment (small business accounts with international transactions). Compliance team notices pattern during audit review. Security team identifies that recent rule updates interact poorly with legitimate cross-border payment patterns.	**Internal Functions:** Quality assurance team conducting weekly conversation sample reviews notices agent struggles with German postal address conventions (street number after street name, postal code city prefixes). Data scientists monitoring model performance observe increased uncertainty scores for German market interactions. Operations team notices product catalog mismatches between US terminology and German market product inquiries. Business development team raises concerns about service quality impacting market expansion success.
Detection Outcome: Multiple signals converge; automated monitoring flags suspicious patterns, users report blocked legitimate transactions, internal reviews identify systematic false positives affecting specific customer segment. Incident enters response process at Month 2 when pattern becomes clear across detection mechanisms.	**Detection Outcome:** Multiple signals converge: monitoring shows degraded performance metrics in German market, users report poor service quality in new market, internal reviews identify pattern of failures correlated with German address formats and product terminology. Incident enters response process at Month 3 when pattern becomes clear across detection mechanisms and threatens market expansion objectives.

Preparedness Recommendations

Ideally, organizations should establish detection infrastructure before deploying AI systems:

- **Monitoring infrastructure** deployed and configured to track relevant technical metrics
- **User feedback channels** established, monitored, and integrated into incident detection workflows
- **Clear procedures** for internal teams to escalate issues from their functional areas
- **Cross-organizational detection protocols** for system-of-systems deployments
- **Defined thresholds** for automated alerts calibrated to avoid both missed detections and alert fatigue
- **Integration points** connecting detection mechanisms, so patterns become visible across signal sources

Without this preparedness infrastructure, organizations rely on chance discovery of incidents rather than systematic detection.

3.2.2 Step 2: Assess

Action: *Evaluate the severity and potential impact using established criteria.*

Assessment determines incident severity, prioritizes response efforts, identifies required resources, and establishes escalation paths. During active incidents, organizations must make rapid decisions about response intensity: whether the incident requires executive notification, whether to activate emergency response procedures, whether external expertise is needed, and whether regulators must be notified. Assessment frameworks enable consistent, defensible determinations under time pressure.

Assessment differs from investigation (Step 5). Assessment makes initial severity determinations based on observable information to guide immediate response. Investigation conducts deeper analysis to determine root causes. Assessment asks "how bad is this?" Investigation asks "why did this happen?"

Multi-Dimensional Assessment

Effective assessment requires frameworks accounting for multiple dimensions simultaneously rather than relying on any single factor. Organizations should consider:

- **Magnitude and type of harm**: Physical injury, financial loss, privacy violations, service denial, or combinations thereof
- **Sensitivity of affected populations**: Incidents affecting vulnerable populations or distinct population segments warrant heightened concern beyond raw numbers affected
- **Criticality of affected systems**: Mission-critical systems, customer-facing services, and systems subject to regulatory requirements demand different response urgency
- **Regulatory implications**: Certain incident types trigger mandatory reporting obligations or compliance reviews
- **Potential for ongoing or escalating harm**: Active incidents causing continuing harm require more urgent response than contained historical incidents

Assessment synthesizes these dimensions into severity classifications that drive response decisions.

Developing Severity Classification Frameworks

Organizations should develop frameworks tailored to their operational contexts and system characteristics while referencing proven standards. Building from scratch invites inconsistency and errors in high-pressure situations. Adapting proven frameworks provides structure while enabling customization for specific organizational needs.

Three Approaches to Severity Classification

Organizations can develop severity classifications using different approaches, each emphasizing different aspects of incidents. Below are three common emphases (not an all-inclusive list) for developing incident severity criteria.

Harm-Based Severity focuses on the impact on affected entities: people, organizations, property, communities, or the environment. This approach evaluates the magnitude of harm the incident caused. Harm-based classifications work well for incidents with clear, measurable impacts such as physical injury, financial loss, environmental damage, or privacy violations. This approach aligns naturally with regulatory frameworks focused on consumer protection, civil rights, and safety.

Mission and Operations-Based Severity focuses on the impact on organizational operations and critical functions. This approach asks: How much did this incident disrupt what we need to do? Operations-based classifications work well for incidents affecting business continuity, service availability, or mission execution. The same technical failure may warrant different severity classifications depending on system criticality. A navigation system error that causes minor inconvenience in a personal vehicle represents a catastrophic failure in aircraft flight control.

Performance-Based Severity focuses on system reliability trends and performance characteristics rather than individual incident impacts. This approach asks: How is system reliability changing? How well does the system perform relative to requirements and baselines? Performance-based classifications work well for incidents involving quality degradation, increasing error rates, deviation from documented capabilities, or drift from acceptable performance baselines.

Performance-based severity considers multiple factors: gradual performance decline over time, variation across user populations, incident frequency and rate, increasing output variability, shifting false positive and false negative rates, declining robustness to edge cases, and scope of problems broadening across system functions. AI systems may reach severe conditions through accumulation rather than through any single discrete failure. Performance-based assessment captures these patterns that individual incident analysis would miss.

Organizations may need multiple severity frameworks for different incident types. A financial institution might use harm-based severity for incidents affecting customers, operations-based severity for incidents affecting transaction processing, and performance-based severity for monitoring fraud detection system reliability.

Incident Aggregation

Multiple incidents of lower individual severity can aggregate to higher severity when they exhibit systematic patterns. Ten marginal incidents affecting random customers may warrant marginal classification individually. Ten marginal incidents, all affecting customers from the same demographic group, aggregate to critical severity due to the systematic pattern indicating systematic performance disparities. Twenty performance degradation incidents distributed randomly across a system may remain marginal individually. Twenty performance degradation

incidents concentrated in specific operational contexts aggregate to critical severity because they reveal systematic reliability problems.

Organizations should consider both individual incident severity and aggregate pattern severity when classifying incidents and determining response priorities.

MIL-STD-882E as Structural Foundation

MIL-STD-882E,[31] the Department of Defense Standard Practice for System Safety, provides a proven structural foundation that organizations can adapt. Originally developed for defense applications, this standard has been successfully adapted across aerospace, healthcare, automotive, natural disasters,[32] and critical infrastructure domains.[33] [34] It defines four severity levels that provide starting points organizations can customize:

- **Catastrophic (Level I)**: Death, permanent total disability, irreversible significant environmental impact, or monetary loss exceeding $10M
- **Critical (Level II)**: Permanent partial disability, injuries requiring hospitalization of three or more people, reversible significant environmental impact, or monetary loss between $1M and $10M
- **Marginal (Level III)**: Injury or illness resulting in lost workdays, reversible moderate environmental impact, or monetary loss between $100K and $1M
- **Negligible (Level IV)**: Injury or illness not resulting in lost workdays, minimal environmental impact, or monetary loss under $100K

Organizations should adapt the specific dollar thresholds, harm definitions, and level names to reflect their operational contexts. However, the four-level structure spanning catastrophic to negligible applies broadly across complex systems domains and provides consistency, enabling comparison across incident types.

Harm-Based Severity Classification

Table 2 presents entity-focused severity criteria adapted from MIL-STD-882E, illustrating how organizations can customize definitions while maintaining structural consistency. This harm-based approach provides three parallel classification schemes focusing on individual health, environmental impact, and financial loss. Organizations can develop additional columns reflecting their specific contexts, such as impact on civil rights, damage to critical infrastructure, or harm to vulnerable populations.

---

[31] United States of America Department of Defense. Department of defense standard practice, system safety (mil-std-882e). Department of Defence, 2012. https://safety.army.mil/Portals/0/Documents/ON-DUTY/SYSTEMSAFETY/Standard/MIL-STD-882E-change-1.pdf

[32] H. J. Caldera and S. C. Wirasinghe. A universal severity classification for natural disasters. Natural Hazards, 111:1533–1573, 2021. doi: 10.1007/s11069-021-05106-9

[33] https://nij.ojp.gov/sites/g/files/xyckuh171/files/media/document/draft-failure-definitions-and-scoring-criteria.docx

[34] FEMA has identified different function levels for its operations: Primary Mission Essential Functions, Mission Essential Function, and Essential Supporting Activities. For FEMA, a failure's severity level could be determined by which of these function levels was impacted. https://www.fema.gov/sites/default/files/2020-07/Federal_Continuity_Directive-2_June132017.pdf

Table 2: Entity-focused severity criteria provide standard definitions spanning multiple harm dimensions. Organizations adapt dollar thresholds and harm definitions to their operational contexts while maintaining the four-level structure enabling consistent severity assessment across incident types.

Harm Severity Scoring Criteria Based on Entity Impact				
Description	Severity Category	Incident Result Criteria Examples		
		Individual Health Focused	Environment Focused	Financial Loss Focused
Catastrophic	I	Death or permanent total disability	Irreversible significant environmental impact	Monetary loss (or equivalent property damage) equal to or exceeding $10M
Critical	II	Permanent partial disability, injuries, or occupational illness that may result in the hospitalization of at least three personnel	Reversible significant environmental impact	Monetary loss (or equivalent property damage) equal to or exceeding $1M but less than$10M
Marginal	III	Injury or occupational illness resulting in one or more lost work day(s)	Reversible moderate environmental impact	Monetary loss (or equivalent property damage) equal to or exceeding $100K but less than $1M
Negligible	IV	Injury or occupational illness not resulting in a lost workday	Minimal environmental impact	Monetary loss (or equivalent property damage) less than $100K
Severity criteria from MIL-STD-882E[31]				

Mission and Operations-Based Severity Classification

Table 3 illustrates operations-focused severity criteria emphasizing mission and task impact rather than entity harm. This table demonstrates a critical insight: **operational context determines severity classification**. The same technical issue warrants different severity levels depending on system criticality and operational consequences.

Consider degraded navigation output. In an airborne radar system, degraded output prevents safe flight, warranting catastrophic classification because the mission cannot be safely executed. In a personal vehicle GPS, degraded output causes missed turns and inconvenience, warranting

marginal classification because the vehicle remains safely operable through alternative navigation methods. The **technical failure is similar**, but the **operational consequences differ dramatically**.

This context-dependency means organizations deploying AI systems across multiple operational contexts may need different severity classifications for technically similar failures. An image generation error producing six-fingered hands represents a marginal issue for a consumer entertainment application but could represent a critical issue for a medical imaging system where anatomical accuracy matters for diagnosis.

Table 3: Examples of operations-focused severity criteria demonstrate that operational context determines severity classification

Severity Category	Incident Result Criteria Examples		
	Airborne Radar	Continuously Operating AI	Personal Vehicle
1	Engine failure prevents safe flight.	Image generators produce child sexual abuse material (CSAM).	A tire blows out and needs to be replaced.
2	Some radar antenna elements are not working. The radar is operable, but its performance is degraded.	Image generators cannot consistently remove types of objects (e.g., dogs, airplanes, cars, etc.) when requested through a text prompt.	The internal GPS navigation system has an old map and needs updating. The system usually works well, but more recent maps would prevent wrong or missed turns.
3	An overhead interior light needs replacing, but operations are not impacted.	Created images sometimes have hands with 6 fingers.	A small dent in the passenger door.
Table taken from [35]			

Performance-Based Severity Classification

Performance-based approaches assess severity through system reliability trends rather than individual incident impacts. A single misclassification in a fraud detection system may represent negligible severity. A pattern of increasing misclassification rates over time, even with each individual incident remaining negligible, may aggregate to critical severity because the pattern indicates systematic reliability degradation.

[35] Boston, M. F., Frase, H., & Georgala, E. (2025). Reliability and Repair for Agentic Systems. Reins AI Technical White Paper v1.0. October 2025. Retrieved from www.reinsai.com/articles/reliability-and-repair-for-agentic-systems

Performance-based severity proves particularly valuable for AI systems exhibiting gradual drift or degradation. Organizations should track metrics including error rates over time, performance gaps across user populations, quality score trends, and incident frequency patterns. When these metrics show consistent degradation, the aggregate pattern may warrant a higher severity classification than any individual incident would receive.

Performance-based assessment requires baseline measurements and a continuous monitoring infrastructure. Organizations cannot assess whether performance has degraded without knowing previous performance levels and tracking changes over time.

Number of Severity Levels

Organizations should use at least four severity levels rather than fewer. Three-level frameworks (such as High-Medium-Low or Critical-Moderate-Minor) create operational problems that undermine effective incident response.

The least consequential level tends to be ignored. Organizations rarely allocate resources to investigating or correcting incidents classified at the lowest severity. These incidents get logged but often receive no systematic response. When only three levels exist, this means one-third of the severity range receives minimal attention.

The most severe level rarely occurs or may never occur for many systems. Truly catastrophic incidents with deaths, permanent disabilities, or losses exceeding $10M happen infrequently. When only three levels exist, the highest level may remain empty or nearly empty for extended periods, making it operationally irrelevant for most day-to-day prioritization decisions.

Most incidents of operational interest fall in the middle range between ignored negligible incidents and rare catastrophic incidents. With only three severity levels, this entire middle range collapses into a single category. Organizations cannot differentiate between incidents requiring urgent executive attention and incidents that can wait for scheduled maintenance cycles. All incidents in this broad middle category compete equally for resources and attention despite having different actual urgency and impact.

Dividing the middle range into two or more levels dramatically improves operational utility. With four levels, organizations can distinguish between:

- Critical incidents requiring immediate executive notification and emergency response
- Marginal incidents requiring systematic investigation and correction but not emergency procedures

This distinction enables better prioritization and clearer escalation criteria. Response teams know which incidents demand immediate action and which can follow standard processes. Executives receive notifications about genuinely urgent issues rather than being overwhelmed by all non-negligible incidents.

Five or more levels provide even finer gradation, though organizations should balance granularity against consistency. More levels enable more precise prioritization but require more detailed criteria and may introduce classification inconsistencies if criteria are not sufficiently clear.

Table 3 presents a three-level framework for illustration purposes, demonstrating how operational context affects severity. Organizations adapting this approach for their own use

should consider adding a fourth level to improve operational utility. Incident aggregation, discussed earlier, addresses some prioritization needs by elevating multiple lower-severity incidents to higher aggregate severity when patterns indicate systematic problems. However, aggregation does not fully substitute for appropriate severity level granularity in the base framework.

AI-Specific Severity Criteria Factors

AI-enabled systems introduce factors requiring explicit consideration during the development of the severity levels and the Assess step:

- **Differential Performance Patterns:** Incidents exhibiting systematic performance variations across distinct user populations warrant elevated severity assessment, even when aggregate harm magnitudes might suggest a lower classification. Systematic performance disparities create both direct harm to affected users and legal compliance risks under anti-discrimination regulations in many jurisdictions.
- **Potential for Harm Accumulation**: AI incidents may cause harm that accumulates across multiple interactions rather than manifesting in single discrete events.A financial advisory system making small calculation errors in each transaction may create significant cumulative financial harm over thousands of transactions with hundreds of users. Assessment should consider total accumulated harm, not just individual interaction impacts.
- **Context-Dependency of Failure**: Systems that fail for specific user populations, in particular operational contexts, or under certain environmental conditions may show incident severity that varies dramatically by deployment scenario. Assessment should evaluate severity across the full range of actual operational contexts, not just average or typical usage scenarios.
- **Cascading Potential**: AI systems integrated into workflows or system-of-systems architectures may propagate failures across organizational boundaries. Assessment should consider not just immediate direct harm but also potential for cascading effects through interconnected systems.
- **Regulatory Reporting Thresholds**: Certain jurisdictions and sectors mandate incident reporting above specified thresholds. EU AI Act requirements, sector-specific regulations in financial services or healthcare, and data breach notification laws all establish reporting obligations triggered by specific incident characteristics. Assessment should explicitly evaluate whether incidents cross these regulatory thresholds, as crossing them affects required response actions and timelines.

Leveraging Organizational and Regulatory Context

Assessment does not occur in isolation from organizational realities and regulatory requirements. Effective assessment integrates:

- **Business impact analyses:** quantifying operational and financial consequences
- **Operational dependencies:** identifying which business functions depend on affected systems
- **Mandatory reporting thresholds:** establishing regulatory obligations
- **Compliance obligations:** defining required response actions and timelines
- **Stakeholder agreements:** specifying contractual responsibilities

Organizations should develop assessment templates with standardized fields enabling rapid, consistent evaluation while accommodating context-specific factors. Templates should capture affected population size and characteristics, type and magnitude of harm, system criticality,

regulatory implications, and escalation requirements. Standardized templates enable consistent assessment across different incident types and different assessors while ensuring all relevant factors receive explicit consideration.

Applying Assessment Across Domains

Table 4 illustrates how assessment operates in both a mature complex system domain (financial services fraud detection) and in agentic AI systems. Both domains require multi-dimensional assessment considering harm magnitude, affected populations, operational impact, regulatory implications, and potential for ongoing harm.

Table 4: Comparative Example: Assessment in Financial Services and Agentic AI. Multi-dimensional assessment synthesizes harm magnitude, population sensitivity, system criticality, regulatory implications, and ongoing harm potential into severity classifications.

Comparative Example for Step 2 Assessment in Financial Services and Agentic AI	
Financial Services: Fraud Detection System	**Agentic AI: Customer Service Agent - International Expansion**
Initial Observations: Fraud detection system blocking legitimate transactions for customers traveling internationally. Pattern identified through detection mechanisms: automated monitoring shows 22% false positive rate (up from 8% baseline) for small business accounts with cross-border transactions. Customer complaints about blocked cards during legitimate travel. Branch managers reporting multiple cases.	**Initial Observations:** Customer service agent exhibiting degraded performance in German market operations. Pattern identified through detection mechanisms: automated monitoring shows 15% fallback-to-human rate in German market vs. 2% baseline in US market. Quality scores lower for German market interactions (3.2 vs. 4.1 out of 5). User complaints disproportionately from new market operations. QA reviews confirm agent struggles with German address formats and product terminology. Approximately 50,000 customer interactions affected over 3-month period.
Harm Type & Magnitude: Tangible harm: Customers unable to access funds during travel (stranded travelers, emergency payment failures). Legitimate business transactions declined causing operational disruption of the affected business and potential business relationship damage. **Intangible harm:** Customer trust in bank eroded, reputational damage. **Magnitude:** Approximately 200 customer accounts affected over 6-week period. Estimated financial impact: $50K in direct costs (expedited card replacements, fee reversals), $500K in estimated business relationship impact.	**Harm Type & Magnitude: Tangible harm:** Extended resolution times for German market customers, requiring multiple interactions vs. single interaction for US market customers. Increased operational costs from elevated human escalation rates. Slower market penetration due to poor customer experience. **Intangible harm:** Brand reputation damage in strategically important expansion market, customer frustration affecting market adoption rates. **Magnitude:** Approximately 50,000 customer interactions affected over 3-month period representing 12% of German market customer base. Estimated 30-40 customer accounts lost. Revenue impact from delayed market penetration.
Affected Population Sensitivity: Small business owners during business-critical travel. International transaction patterns common for import/export businesses. Timing-sensitive transactions (payroll, vendor payments). Population includes established	**Affected Population Sensitivity:** Early adopter customers in newly launched international market. Market expansion represents significant business investment and strategic growth initiative. Customer experience in launch phase shapes brand reputation in new geography. Service

Comparative Example for Step 2 Assessment in Financial Services and Agentic AI	
Financial Services: Fraud Detection System	**Agentic AI: Customer Service Agent - International Expansion**
long-term customers with significant account balances. Some customers are sole proprietors for whom personal and business finances are closely linked.	failures during market entry create amplified reputational risk. Market represents key growth opportunity for company's international expansion strategy.
Criticality: Customer-facing system affecting payment access and account services. Not mission-critical infrastructure but the incident affects customer satisfaction and retention. Moderate business criticality, because customers can use alternative payment methods or request manual override. However, brand reputation impact is high. Failure affects customer trust in bank's reliability and fraud protection capabilities.	**Criticality:** Customer-facing system handling 100% of initial customer service inquiries in German market. Business criticality high: customer service quality directly impacts market penetration success, affects company's public commitments to international expansion, and determines competitive positioning in new geography. Service quality failures threaten strategic business objectives and revenue growth targets. Market expansion timeline at risk.
Regulatory Implications: Consumer Financial Protection Bureau (CFPB) oversight of unfair or deceptive practices. No mandatory immediate reporting threshold crossed, but pattern could trigger examination scrutiny. State banking regulators may consider pattern in safety and soundness examinations. Potential civil litigation exposure. Internal compliance review required.	**Regulatory Implications:** Potential violation of consumer protection laws prohibiting discrimination in service provision. Federal Trade Commission (FTC) has jurisdiction over cross-border trade practices. German consumer protection regulations (BGB §312a) may apply to automated customer interactions. No mandatory immediate reporting threshold but company's own service level commitments and market entry representations create accountability obligations. Legal review required for market-specific compliance.
Ongoing Harm Potential: Incident is active and ongoing (false positives continue affecting new customers until system fixed). Each day increases affected customer count and accumulates additional harm. However, customers can request manual overrides and problem is understood by branch staff who can provide workarounds. Escalating harm potential is moderate, because the problem is contained to known customer segment and temporary workarounds exist.	**Ongoing Harm Potential:** Incident is active and ongoing. Degraded service continues for all German market customers until correction implemented. Each day, additional interactions accumulate with varying service quality. Harm is continuous and affecting large customer volume. Escalating harm potential is moderate to high: pattern may worsen through feedback loops, threatens market expansion timeline, and damages brand reputation in strategically critical geography. Workaround (elevated human review) is activated but increases operational costs and delays market profitability.
Severity Classification: Critical (Level II) Rationale**:** Magnitude affects 200 customers with significant operational impact but not catastrophic. Harm is tangible and immediate but temporary and reversible. Sensitive population (business owners during critical travel) elevates concern. System is customer-facing with high reputational stakes. Regulatory exposure exists but no immediate mandatory reporting. Ongoing active incident requires urgent response but not emergency procedures. Monetary impact <$1M justifies Critical rather than Catastrophic classification.	**Severity Classification: Critical (Level II)** Rationale**:** Magnitude affects 50,000 interactions in strategically important new market launch. Both tangible harm (service quality, operational costs) and intangible harm (brand reputation in expansion market) are present. Customer-facing system affecting strategic business growth initiative. High reputational stakes in critical market expansion. Regulatory exposure requires legal review of market-specific requirements. Ongoing active incident affecting business-critical expansion requires urgent response and executive engagement. Classification as Critical (Level II) triggers executive notification, immediate corrective action, strategic business review, and consideration of market entry timeline adjustments.

Preparedness Recommendations

Organizations should establish assessment capabilities before deploying AI systems:

- **Severity classification framework** developed, documented, and adapted to organizational context
- **Assessment templates** created with standardized fields capturing all relevant dimensions
- **Personnel trained** on applying criteria consistently across incident types
- **Escalation thresholds** defined linking severity levels to required notifications and response actions
- **Integration with regulatory requirements,** ensuring assessment explicitly evaluates reporting obligations
- **Clear decision authority** designating who can make severity classifications and authorize escalations
- **Documentation requirements** specifying what information should be captured during assessment

Without this preparedness infrastructure, organizations make ad hoc severity determinations that may be inconsistent across incidents, do not account for critical factors, or miss regulatory obligations until it is too late.

3.2.3 Step 3: Stabilize

Action: *Execute pre-planned procedures to contain harm.*

Stabilization and correction serve different purposes in the incident response cycle. Stabilization focuses on immediate mitigation of active harm and stopping ongoing harm through predetermined procedures that can be executed rapidly without complex decision-making under pressure. Correction, addressed later in Step 6, seeks to address harm already realized, improve the system, and reduce incident recurrence through systematic fixes targeting root causes.

This distinction matters because they have different appropriate actions. Stabilization prioritizes speed and harm containment, even if solutions are imperfect or temporary. Correction prioritizes thoroughness and long-term reliability improvement, even if solutions take time to develop and deploy. Organizations that conflate these purposes may either respond too slowly during active incidents while looking for perfect solutions or implement quick fixes that fail to prevent recurrence.

Why Pre-Planning is Critical

Time pressure during active incidents makes careful deliberation impractical. When an AI system is actively causing harm, responders should act quickly to contain the situation. Deciding what actions to take, who has authority to authorize them, how to execute them technically, and what communication protocols to follow cannot happen effectively in the moment. Pre-planning is not optional for effective stabilization.

Organizations should identify likely incident types and failure modes, design stabilization responses for each, document procedures in accessible formats, train staff on execution, and test these actions before deploying AI systems. This preparation transforms stabilization from improvisation under pressure into execution of intentional, vetted procedures.

Each **stabilization procedure should specify** at least five key elements:

1. **Trigger conditions** define what circumstances activate this procedure. Clear triggers enable rapid decision-making about which procedure to execute. Ambiguous triggers lead to delays while responders debate whether conditions warrant action.
2. **Execution steps** detail what to do, in what order, with sufficient technical specificity that responders can execute without improvisation. Steps should be concrete and actionable rather than general guidance requiring interpretation.
3. **Authorization requirements** specify who can authorize stabilization actions. Some stabilization procedures may be pre-authorized for immediate execution by on-call responders. Others may require executive approval due to business impact or customer-facing consequences. Clear authorization paths prevent delays while seeking approval and prevent unauthorized actions with significant consequences.
4. **Exit criteria** define conditions under which organizations should remove stabilization actions and return to normal operations. Temporary measures like disabling features or implementing rate limits may create their own problems if left in place too long. Exit criteria specify when to remove temporary measures based on incident resolution, elapsed time, or business need.
5. **Communication protocols** specify who needs notification about stabilization actions and what information they need. Internal stakeholders, including executives, operations teams, and customer service, need awareness. External stakeholders, including affected users and potentially regulators, need appropriate notification. Communication protocols reduce the risk that critical stakeholders learn about incidents through channels other than official notification.

Organizations can prepare multiple types of stabilization actions. **Common stabilization actions** include:

- **Rolling back to previous system versions** returns systems to last known good states. This works well when recent updates or changes correlate with incident onset. Rollback requires maintaining previous versions in deployment-ready states and having tested procedures for version switching.
- **Activating fail-safes that constrain AI behavior**[36] implements automatic safety boundaries limiting what actions AI systems can take without turning them off entirely. Fail-safes might restrict AI to read-only operations, limit transaction amounts below specified thresholds, disable specific tools or capabilities while maintaining core functions, or require additional confirmation for high-risk actions. This works well when the AI can continue operating safely within narrower boundaries. Fail-safes require pre-configured constraints that can be activated quickly.
- **Switching to backup systems or manual processes** takes AI systems offline entirely and maintains critical services through alternatives. This requires maintaining backup capabilities and regular testing to ensure backups function when needed.

[36] Hanmer, R.S. (2007). Patterns for Fault Tolerant Software. Wiley Software Patterns Series. John Wiley & Sons.

- **Temporarily disabling affected functionality while maintaining critical services** contains harm without complete service disruption. This works well for systems with modular functionality where some features can be disabled without affecting core services. Partial disabling requires architectural design supporting feature flags and graceful degradation.[37] [38]
- **Implementing rate limiting or access restrictions** reduces incident velocity without stopping service entirely. This works well when incidents scale with usage volume or when specific access patterns correlate with failures. Rate limiting requires monitoring infrastructure, detecting when limits engage, and mechanisms for adjusting limits dynamically.
- **Escalating human oversight** keeps AI systems active but adds mandatory human review before outputs take effect. This works well when AI outputs are usually acceptable, and human review can catch problematic cases. Human oversight slows processing but leverages AI capabilities, ensuring problematic outputs are caught before causing harm. This approach requires available personnel with appropriate expertise and clear criteria for approving or rejecting AI outputs.

Context-Dependent Action Selection

The appropriate stabilization actions for any system depend on operational context, system criticality, assessed risk profile, and available alternatives. Mission-critical systems with no acceptable backup may require different stabilization approaches than systems with redundant alternatives. Customer-facing systems with high availability requirements face different constraints than internal productivity tools.

Organizations should develop stabilization procedures specific to their deployment contexts rather than applying generic procedures across all systems.

Overlap with Risk Management

Much stabilization pre-planning overlaps with organizational risk management processes. Risk management identifies potential failures, assesses their likelihood and impact, and develops mitigation strategies. These risk mitigation strategies often become stabilization procedures when incidents occur. Organizations can leverage documentation and procedures from risk management efforts for incident response, reducing duplicated work and ensuring consistency between risk planning and incident response.

Applying Stabilization Across Domains

Table 5 illustrates how stabilization operates in both a mature complex system domain and in agentic AI systems. Both domains require pre-planned procedures enabling rapid response without improvisation, clear authorization for executing procedures under pressure, and communication protocols ensuring stakeholders receive timely notification.

---

[37] Amazon Web Services. (2022). "REL05-BP01 Implement graceful degradation to transform applicable hard dependencies into soft dependencies." AWS Well-Architected Framework - Reliability Pillar.
[38] Edwards, Tamsyn & Lee, Paul. (2017). Towards Designing Graceful Degradation into Trajectory Based Operations: A Human-Machine System Integration Approach. 10.2514/6.2017-4487.

Table 5: Comparative Example: Stabilization in Financial Services and Agentic AI. Stabilization executes pre-planned procedures enabling rapid harm containment without improvisation under pressure.

Comparative Example for Step 3 Stabilization in Financial Services and Agentic AI	
Financial Services: Fraud Detection System	**Agentic AI: Customer Service Agent - International Expansion**
Incident Context: Fraud detection system generating 22% false positive rate for small business accounts with international transactions. Blocking legitimate payments during customer travel. Assessment classified as Critical (Level II) requiring urgent response.	**Incident Context:** Customer service agent performance degradation incident affecting German market operations. Stabilization measures active with mandatory human review for German market interactions. Investigation proceeding. Documentation and notification required.
Pre-Planned Procedure Activated: Procedure FD-STAB-003: "False Positive Surge Protocol" pre-approved for execution by fraud operations manager when false positive rate exceeds 15% for any customer segment. Procedure documented, tested quarterly, and stored in incident response playbook. Authorization is pre-approved because the procedure does not disable fraud detection entirely.	**Pre-Planned Procedure Activated:** Internal procedure CS-STAB-004: "Market Launch Performance Protocol" pre-approved for execution by international operations director when quality metrics in new market operations fall below acceptable thresholds. Procedure developed during market expansion planning, tested in staging environment, stored in incident response playbook. Authorization pre-approved because procedure adds oversight without stopping service in expansion market.
Execution Steps: Lower blocking threshold for affected customer segment to reduce false positives. Route all blocks affecting small business international accounts to manual review queue with 3 additional fraud analysts assigned. Configure monitoring to alert if legitimate fraud increases. Notify branch managers of temporary manual review process.	**Execution Steps:** Activate mandatory human review for all German market customer interactions before final response delivery. Route German market queries to specialized review team with German language capability and market knowledge. Assign 5 additional German-speaking customer service specialists to review queue. Configure monitoring to track review queue depth and response time impacts. Notify German market operations team and business development stakeholders of temporary elevated oversight process.
Exit Criteria: Remove manual review process and restore automated thresholds when: false positive rate returns below 10% for affected segment for 7 consecutive days, OR permanent fix deployed and verified through A/B testing showing sustained improvement, OR 30 days elapsed requiring executive decision on extending temporary measures vs. alternative approaches.	**Exit Criteria:** Remove mandatory human review and restore normal agent operation when: quality score gap between German and US markets narrows below 0.3 points for 14 consecutive days, OR permanent fix deployed and verified through staged rollout in German market, OR 45 days elapsed requiring executive decision on extending elevated oversight vs. modifying system functionality or adjusting market entry timeline.
Communication Protocol: Internal: Fraud operations team, branch managers, customer service, executive dashboard updated within 1 hour. **External:** Affected customers receive proactive notification: "We are reviewing international transaction patterns to improve service. You may experience brief delays on international payments while we verify transactions. Contact us immediately if urgent payment needed." **Regulatory:** Internal compliance review determines no immediate	**Communication Protocol: Internal:** German market operations team, international business development, customer service leadership, executive leadership notified within 2 hours. **External:** No immediate public notification during stabilization phase. German market customers experiencing delays receive explanation: "Your inquiry requires specialized review to ensure highest quality response. A team member will respond within 4 hours." Messaging emphasizes quality assurance, not

Comparative Example for Step 3 Stabilization in Financial Services and Agentic AI	
Financial Services: Fraud Detection System	**Agentic AI: Customer Service Agent - International Expansion**
reporting obligation, but incident logged for examination disclosure.	system failure. **Regulatory:** Legal team reviews whether German consumer protection notification is required and determines that stabilization measures sufficient pending permanent correction.
Stabilization Outcome: False positive rate for affected segment drops to 12% within 24 hours of procedure activation. Manual review queue processing 40 cases per day with 2-hour average turnaround. Zero legitimate fraud incidents detected in manual review process during first week. Customer complaints about blocked payments decrease 80%. Stabilization holds while investigation determines the root cause and develops a permanent correction.	**Stabilization Outcome:** All German market interactions receive human oversight before delivery. Average response time increases from 3 minutes to 35 minutes for German market customers (still within service level agreements). Quality scores for German market improve to 3.8 (vs. 4.1 for US market, gap narrowed from 0.9 to 0.3 points). Customer complaints from German market decrease 60%. Review queue depth stabilizes at manageable levels. Market expansion timeline extended by 6 weeks to allow for permanent correction. Stabilization holds while investigation proceeds and permanent correction is developed.

Preparedness Recommendations

Organizations should develop stabilization capabilities before deploying AI systems:

- **Stabilization procedures documented** for likely incident types, stored in accessible formats with version control
- **Clear authorization and escalation paths** defining who can execute procedures and under what conditions
- **Backup systems or manual processes available** and tested regularly to ensure functionality when needed
- **Personnel trained** on procedure execution through tabletop exercises and periodic drills
- **Testing of procedures** before system deployment to identify gaps or impractical steps
- **Integration with existing incident response** connecting AI stabilization procedures to broader organizational response capabilities
- **Communication templates** prepared for internal and external stakeholders enabling rapid notification
- **Monitoring infrastructure** providing visibility into stabilization effectiveness and detecting unintended consequences

Without this preparedness infrastructure, organizations must improvise stabilization responses during active incidents, leading to delays, inconsistent actions, and increased risk of stabilization measures causing additional problems.

3.2.4 Step 4: Report and Document

Action: *Document incident details using standardized structures and notify appropriate stakeholders.*

Reporting and documentation serve multiple critical purposes simultaneously throughout the incident lifecycle. During active response, reports enable coordination across distributed teams and time zones. For investigation and analysis, they preserve evidence and system states that might otherwise be lost. For organizational learning and regulatory compliance, they create structured records enabling pattern recognition across incidents and demonstrating due diligence. Effective reporting should balance standardization for computational analysis with contextual narrative for human comprehension, while serving audiences ranging from immediate responders to future investigators to external regulators.

Multiple Purposes Throughout the Lifecycle

Reporting serves distinct purposes at different stages of incident response. Understanding these multiple purposes clarifies why reporting matters and what information different audiences need.

- **Response Coordination:** Reporting provides shared understanding of incident status without requiring constant synchronous communication. Teams can reference documentation to understand completed actions, remaining tasks, and handoff points. This structured record proves essential when incidents span multiple shifts or require escalation to personnel not involved in initial response.
- **Investigation and Analysis:** Reports preserve evidence supporting root cause analysis. Documentation of initial observations, system states, user reports, and environmental conditions provides essential data that might otherwise be lost as systems change, logs rotate, or memories fade. This proves particularly critical for AI incidents where non-deterministic behavior means incidents may not be reproducible.
- **Pattern Recognition:** Standardized reporting structures enable computational analysis identifying patterns invisible in narrative-only reports. Organizations can detect recurring failure modes, common root causes, and systemic issues affecting multiple systems. This structured approach transforms individual incident reports into strategic intelligence about system reliability and risks.
- **Regulatory and Compliance:** Reporting fulfills mandatory obligations, demonstrates due diligence, and supports regulatory oversight. Certain jurisdictions and sectors mandate incident reporting above specified thresholds. Documentation shows organizations responded appropriately and took required actions.
- **Stakeholder Communication:** Reporting maintains trust through transparency, keeps affected parties informed, and documents remediation actions. Users affected by incidents need appropriate notification. Executives need awareness of significant incidents. Partners and customers may have contractual rights to incident information.
- **Institutional Learning:** Reporting builds organizational knowledge, informs future prevention efforts, and contributes to ecosystem-wide learning. Organizations can review historical incidents when designing new systems, training personnel, or updating procedures.

The value of incident reports extends far beyond immediate response.

Standardized Structures Enable Analysis

Effective reporting requires standardized structures balancing machine-readability with human comprehension. Purely narrative reports work well for human readers but cannot support computational analysis. Purely structured reports with rigid fields may miss important contextual

nuances. Hybrid structures combining both approaches provide optimal utility. Key elements of effectiveness standardized reporting structures include:

- **Standardized fields** enable computational analysis. Fields with consistent data types, controlled vocabularies,[39] and defined formats allow automated processing. Organizations can query incidents by type, filter by affected populations, aggregate by severity, and analyze trends over time.
- **Controlled vocabularies** ensure consistent terminology across incidents. When all reporters use the same terms for similar concepts, analysis becomes reliable. Controlled vocabularies prevent one person describing an incident as "hallucination" while another calls a similar incident "confabulation" while a third calls it "fabricated output" and a fourth describes it as "factually incorrect generation."
- **Consistent data formats** allow integration across systems. Dates in standard formats, numerical values with defined units, and identifiers following consistent patterns enable data sharing and aggregation. Inconsistent formats require manual cleaning before analysis, introducing errors and delays.
- **Narrative fields** capture context, circumstances, and nuances that structured categories cannot fully represent. Free-text descriptions allow reporters to explain unusual factors, describe unexpected observations, and provide qualitative assessments. Narrative fields give humans the flexibility to communicate what matters without forcing information into predefined categories.

This hybrid approach transforms individual incident reports into datasets supporting systematic analysis while preserving the contextual richness humans need for understanding and decision-making.

Multiple Report Versions

Reports should exist in multiple versions with different sensitivity levels, each protected by appropriate access controls. Different audiences need different information and have different legal rights to access incident details. Below are some examples of different report versions:

- **Full Technical Reports** serve internal incident response teams. These include sensitive technical details, proprietary information, complete forensic data, system architecture specifics, and security vulnerability details. Access should be limited to incident response team members, relevant technical staff, and individuals with operational need to know.
- **Sanitized Versions** support regulatory reporting requirements or information-sharing agreements with other organizations. These remove commercial- or security-sensitive details, protect proprietary information, and limit security vulnerability exposure while preserving information needed for compliance and oversight.
- **Aggregate Statistical Summaries** serve public transparency and sector-wide learning. These present patterns and trends without exposing individual organizational vulnerabilities or competitive information. Public summaries contribute to ecosystem-wide learning without creating security risks.
- **Access Control Infrastructure** supports appropriate information sharing through role-based permissions ensuring individuals access only authorized report versions, audit trails tracking access for accountability, secure sharing channels protecting reports during

[39] Chipangila, B., Liswaniso, E., Mawila, A. et al. (2024). "Controlled vocabularies in digital libraries: challenges and solutions for increased discoverability of digital objects." International Journal on Digital Libraries, Vol. 25, pp. 139–155

transmission, and version control tracking how understanding progresses as investigations advance.

Each report version serves distinct audiences with different information needs, legal rights to access, and responsibilities in the incident response ecosystem.

Applying Reporting Across Domains

Table 6 illustrates how reporting operates in both a mature complex system domain and in agentic AI systems. Both domains require standardized structures[40] enabling analysis, multiple report versions for different audiences, and clear protocols for stakeholder notification.

Table 6: Comparative Example: Report and Document in Financial Services and Agentic AI. Standardized reporting structures enable both immediate stakeholder communication and long-term pattern analysis

Comparative Example for Step 4 Report & Document in Financial Services and Agentic AI	
Financial Services: Fraud Detection System	**AAgentic AI: Customer Service Agent - International Expansion**
Incident Context: Fraud detection system false positive incident affecting small business international accounts. Stabilization measures active. Investigation proceeding. Multiple stakeholders need notification and documentation.	**Incident Context:** Customer service agent performance degradation incident affecting German market operations. Stabilization measures active with mandatory human review. Investigation proceeding. Documentation and notification are required.
Full Technical Report (Internal): Complete system logs, rule configuration details, test data composition analysis, false positive pattern analysis by transaction type, customer segment characteristics, geographic distribution. Security team analysis of whether vulnerability exploitable. Technical root cause investigation findings (added after the Investigate and Analyze step). Complete timeline of detection, assessment, and stabilization actions. Personnel involved and decisions made.	**Full Technical Report (Internal):** Complete conversation logs (anonymized but analyzable), model uncertainty scores by market, address parsing failure analysis by postal format type, database lookup patterns, tool use sequences. Analysis of training data market representation. Performance metrics by operational geography. Complete timeline of detection, assessment, and stabilization. Technical investigation of model behavior, prompt processing, tool interactions, and product catalog integration.
Cross-Institution Sharing (Sanitized) Incident information shared with other financial institutions through Section 314(b). Enables other institutions to identify similar patterns affecting their customers. Uses standardized codes for activity types and behavioral indicators to enable pattern matching across institutions This case does not involve any detection of suspicious	**Regulatory Report (Sanitized):** Incident summary for consumer protection authorities in German jurisdiction: incident type, affected market operations and interaction volume, service quality impact, timeline of detection and response, stabilization measures including elevated human review, investigation status, expected correction timeline. Removes: proprietary model architecture, training data specifics, competitive service delivery information, detailed system implementation. Report demonstrates due diligence in market entry compliance.

[40] Akbari Gurabi, M., Nitz, L., Bregar, A., Popanda, J., Siemers, C., Matzutt, R., & Mandal, A. (2024). Requirements for Playbook-Assisted Cyber Incident Response, Reporting and Automation. Digital Threats: Research and Practice, 5(3), 1–11.

Comparative Example for Step 4 Report & Document in Financial Services and Agentic AI	
Financial Services: Fraud Detection System	**AAgentic AI: Customer Service Agent - International Expansion**
or illicit activity. It is unlikely to be shared with other organizations.	
Stakeholder Notifications: Executive: Critical incident brief with business impact, customer segment affected, stabilization status, investigation timeline. **Branch Managers:** Operational guidance for handling affected customer inquiries, manual override procedures, expected resolution. **Customers:** Proactive notification of transaction review process and alternative payment options. **Compliance:** Incident logged for regulatory examination preparation.	**Stakeholder Notifications: Executive:** Critical incident brief with market expansion impact, affected geography and customer volume, stabilization through human review, business and reputational impact, investigation timeline. **German Market Operations:** Operational guidance for elevated review process, specialized team assignments, expected workload and response time targets. **Affected Customers:** Explanation during delayed response times emphasizing quality assurance. **Legal/Compliance:** Assessment of German consumer protection notification requirements and market-specific regulatory obligations. **Business Development:** Impact analysis on market expansion timeline and competitive positioning.
Pattern Recognition Value: Incident report structure enables future analysis. Similar incidents affecting different customer segments can be identified through query on "false positive rate increase" + "customer segment." System can detect if rule updates correlate with incident patterns across multiple deployments. Aggregate analysis across financial institutions (if shared) could reveal industry-wide test data gaps.	**Pattern Recognition Value:** Incident report structure enables future analysis. Similar incidents affecting different international markets identifiable through query on "market expansion" + "localization failure" + "address parsing." System can detect if training data gaps correlate with new market launches. Aggregate analysis across organizations could reveal systematic market readiness issues affecting international expansion initiatives, informing best practices for future market entries and highlighting common technical debt in US-trained systems expanding internationally.

Preparedness Recommendations

Organizations should create reporting infrastructure before deploying AI systems:

- **Standardized reporting templates** developed with required fields, controlled vocabularies, and narrative sections
- **Access control infrastructure** implemented with role-based permissions and audit capabilities
- **Clear procedures** for different report versions specifying what information each version contains
- **Training** on reporting requirements, ensuring personnel understand what to document and how
- **Integration** with regulatory reporting systems enabling efficient compliance
- **Secure storage** and sharing mechanisms protecting sensitive information
- **Incident tracking systems** managing reports through the complete lifecycle from initial documentation through closure
- **Stakeholder notification protocols** defining who gets notified, what information they receive, and timing requirements

Without this preparedness infrastructure, organizations produce inconsistent reports that cannot support pattern analysis, fail to fulfill regulatory obligations on time, or share inappropriate information with audiences lacking proper authorization.

3.2.5 Step 5: Investigate and Analyze

Action: Determine root cause through systematic analysis.

Investigation and Analysis identify the causes of incidents to inform effective corrective actions. Without understanding root causes, organizations cannot implement corrections that prevent recurrence. Analysis must be systematic and comprehensive, examining multiple levels from individual components through system-of-systems interactions.

Investigate & Analyze: Using your preferred methods

This framework frequently discusses Root Cause Analysis (RCA) as an investigative tool. However, organizations should select and apply the methods best suited to their specific operational needs, technical context, existing processes, and organizational culture. Root cause analysis represents one of many valid investigative methodologies. Other common approaches include Failure Modes and Effects Analysis (FMEA),[41] Systems Theoretic Accident Modeling and Processes (STAMP),[42] [43] and Fault Tree Analysis (FTA).[44] [45] [46] The investigation techniques described here are illustrative rather than prescriptive, and practitioners should adapt or substitute methodologies that align with their established practices and requirements.

It can be advantageous to use multiple investigation and analysis methods. Each method has strengths and weakness. You may need to leverage multiple tools to effectively implement Step 6 Correct.

Multiple Levels of Analysis

Effective incident investigation and analysis require analysis at three distinct but interconnected levels. Root causes may reside at any level or span multiple levels simultaneously. Organizations that focus exclusively on one level may miss critical contributing factors.

Component-Level Issues

Individual components may fail or perform incorrectly. Component-level analysis examines:

---

[41] Stamatis, D.H. (2003). Failure Mode and Effect Analysis: FMEA from Theory to Execution (2nd Edition). ASQ Quality Press.
[42] Leveson, Nancy & Daouk, Mirna & Dulac, Nicolas & Marais, Karen. (2003). Applying STAMP in accident analysis. Workshop Investigation Reporting Incidents Accidents (IRIA).
[43] Nancy G. Leveson, Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press, 2011. ISBN 978-0-262-01662-9.
[44] Vesely, W. E., et al. (2002). Fault Tree Handbook with Aerospace Applications Content
[45] Vesely, W. & Goldberg, F. & Roberts, N. & Haasl, D.. (1981). Fault Tree Handbook. 216.
[46] NASA Software Engineering Handbook, Section 8.07 - Software Fault Tree Analysis, Created by Haigh, Fred, last modified on Jun 30, 2023.

- Specific model failures including incorrect predictions, hallucinations, or outputs inconsistent with training objectives
- Training data issues such as insufficient representation, labeling errors, or distribution shifts
- Configuration errors in system parameters, thresholds, or deployment settings
- Individual tool or plugin failures in agentic systems where a specific tool malfunctions

Integration-Level Issues

Components that work correctly in isolation may fail when connected. Integration-level analysis examines:

- API failures where component interfaces do not communicate correctly
- Data format mismatches where one component's output cannot be correctly interpreted by another
- Communication protocol errors between components
- Version incompatibilities where component updates break existing connections
- Authentication or authorization failures in component-to-component communication

System-Level Problems

The integrated system may exhibit behaviors not predictable from individual components or their connections. System-level analysis examines:

- Workflow design issues that lead to cascading effects or inappropriate sequencing
- Behaviors arising from component interactions rather than individual component or integration failures
- Resource contention or bottlenecks where multiple components compete for limited resources
- Timing dependencies where correct function requires specific ordering or synchronization
- Feedback loops where system outputs influence subsequent inputs in unexpected ways

System-of-Systems Concerns

Incidents may propagate across organizational boundaries or result from interactions between independently operated systems. System-of-systems analysis examines:

- Cascading failures that propagate through connected systems across organizational boundaries
- Behaviors arising from complex interactions between systems that each function appropriately within their own context
- Cross-organizational dependencies where one organization's system performance depends on another organization's system
- Timing and synchronization issues in distributed systems where coordination across organizations proves difficult
- Assumptions about external system behavior that prove incorrect under certain conditions

The Critical Human Factors Dimension

Humans and users are part of the system. Root causes or important contributing factors may reside in human factors rather than technical issues. Organizations that focus investigation and analysis exclusively on technical components miss critical failure modes that originate in how humans interact with AI systems,[47] how organizations structure work around AI systems, or how incentives shape AI system use.

Common Human Factor Issues

Investigation and analysis should examine whether human factors contributed to the incident:

- Inadequate training where users lack understanding of appropriate system use, capabilities, or limitations
- Confusing interfaces that fail to communicate system capabilities and limitations clearly
- Insufficient guidance on appropriate use contexts, operational boundaries, or when human judgment should override AI outputs
- Procedures that incentivize workarounds when official processes prove too slow, cumbersome, or incompatible with operational realities
- Organizational pressures prioritizing speed over accuracy, productivity over safety, or efficiency over quality
- Misaligned incentives where rewards do not align with safe, appropriate, or responsible use
- Inadequate staffing or time pressure that prevents proper oversight or verification of AI outputs
- Authority gradients where junior staff feel unable to question or override AI recommendations

Why Human Factors Prove Harder in AI Than Traditional Software

Unlike traditional software with deterministic behavior, AI systems create unique challenges for users that make human factors more difficult to identify and address.

AI systems produce outputs that may appear plausible while being incorrect. Users cannot rely on obvious error signals. Traditional software typically fails in detectable ways, displaying error messages or producing clearly wrong outputs. AI systems can confidently produce incorrect outputs that appear reasonable, requiring users to exercise judgment about when to trust system outputs.

AI systems shift responsibility from system to user in ways that may not be clearly communicated. When an AI system produces a recommendation, users must often decide whether to accept, modify, or reject that recommendation. The system may not clearly indicate its confidence level or the circumstances under which its outputs should not be trusted. Users bear responsibility for decisions based on AI outputs without clear guidance on when human judgment should prevail.

The boundary between appropriate and inappropriate use shifts with context. What constitutes appropriate use may depend on user expertise, operational conditions, and specific circumstances that change over time. Organizations cannot simply monitor for "incorrect use" because the definition of correct use proves context-dependent in ways that resist simple specification.

---

[47] Garcia-Martin, R., et al. (2024). "The impact of AI errors in a human-in-the-loop process." Cognitive Research: Principles and Implications, 9(1).

Automation bias creates over-reliance[48] [49] on AI recommendations. Users tend to trust automated systems even when they should apply skepticism. This automation bias proves particularly problematic with AI systems because their outputs appear authoritative and may be difficult to verify without significant effort.

Blame-Free Investigation Culture

Organizations should foster blame-free investigation cultures that encourage honest reporting and surface human factors issues that purely technical analysis might miss.[50] When people fear blame, they will not report near-misses, will hide workarounds they have developed, will not admit confusion about system operation, and root causes remain hidden.

Blame-free culture recognizes that most incidents result from system design issues rather than individual failures. If a user misunderstands how to use a system, the interface or training may be inadequate. If users develop workarounds, official procedures may be impractical. If users ignore safety guidelines, incentives may be misaligned. Investigation should ask what system changes would prevent similar incidents rather than assigning individual responsibility for following inadequate procedures or working around poorly designed systems.

This approach does not eliminate accountability for deliberate misuse or violation of clearly communicated policies. It distinguishes between mistakes that reveal system design problems and intentional violations of known requirements.

Ongoing Learning

Investigation and analysis extend beyond determining immediate root causes. Organizations should systematically collect information that improves future incident response, expands understanding of potential failure modes, and builds institutional knowledge about AI system behavior.

- **Staying Current on Failure Modes:** Novel AI failure modes continue to emerge as capabilities expand and deployment contexts diversify. Organizations should systematically review research publications, industry reports, security advisories, and practitioner communities. This informs testing strategies, provides context for investigations, and updates understanding of AI system risks. Technical staff should allocate dedicated time for this knowledge gathering.
- **Capturing Insights for Process Improvement:** During investigation and analysis, teams should document observations and insights that will inform process improvements in Step 6: Correct. These observations include what worked well in the response, what hindered effectiveness, and what assumptions prove incorrect. This documentation enables systematic improvement of both AI systems and incident response processes.

Applying Multi-Level Investigation Across Domains

Table 7 illustrates how investigation and analysis operate in both mature complex system domains and agentic AI systems. Both domains require systematic examination at multiple levels, from individual components through integration, system, and system-of-systems interactions. Human factors analysis proves essential in both contexts. Root cause determination at the appropriate level informs effective corrective actions.

---

[48] Parasuraman, R., & Manzey, D.H. (2010). "Complacency and Bias in Human Use of Automation: An Attentional Integration." Human Factors, 52(3), 381-410.
[49] Singh, A., et al. (2025). "Exploring automation bias in human–AI collaboration: a review and implications for explainable AI." AI & Society.
[50] Dekker, S. (2012). Just Culture: Balancing Safety and Accountability (2nd Edition). CRC Press.

Table 7: Comparative Example: Investigation and Analysis in Financial Services and Agentic AI. Investigation and analysis require examination at multiple levels, from individual components through system-of-systems interactions.

Comparative Example or Step 5 Investigation & Analysis in Financial Services and Agentic AI	
Financial Services: Structuring Detection	**Agentic AI: Customer Service Agent - International Expansion**
Incident Context: Multiple banks report unusual transaction patterns. Individual transactions appear innocuous but collectively suggest money laundering through structuring. 15 different customer accounts across 3 banks, $2M+ over 3 months. No single bank saw enough activity to trigger high-priority investigation.	**Incident Context:** System monitoring detects performance degradation in German market operations (15% fallback-to-human rate vs. 2% baseline). Quality scores lower for German market. Approximately 50,000 customers affected over three months. Multiple detection signals converged indicating systematic technical issues in new market launch.
Component Level: Individual transaction monitoring examines each transaction for fraud indicators: amount exceeds threshold, transaction type inconsistent with account history. Transaction flagging rules operate correctly, each transaction evaluated independently. Components function as designed. Transactions under $10K threshold pass without flags. Analysis confirms: individual transactions appear legitimate when examined in isolation.	**Component Level:** Natural language understanding model, address parser, and database lookup tool each function correctly within specifications when examined in isolation. Address parser performs well on US address formats (training distribution). Database lookup executes exact-match queries correctly. Product catalog returns valid results for US product terminology. Failures occur with inputs outside training distribution: German postal conventions (street number after street name, postal code with city prefix), German product terminology. Components work as designed but training data lacks representation of German market patterns.
Integration Level: Account-level pattern analysis integrates transaction history to calculate customer risk scores. Customer risk scoring updates regularly based on integrated behavioral patterns. Analysis reveals: individual customer accounts show normal risk scores, transaction patterns consistent with stated business activities, no single account has high-risk indicators. Risk scores remain in normal ranges because activity distributed across multiple accounts prevents concentration that would elevate individual scores	**Integration Level:** Address parser outputs incomplete data for German postal formats. Incomplete data fails database exact-match requirement. Lookup failure signals agent uncertainty. Integration lacks fallback mechanisms for partial matches or format variations. Tool chain amplifies initial parsing limitations: parsing error leads to lookup failure leads to agent escalation. Workflow design assumes US address format consistency. No graceful degradation for international address variations. Product catalog integration uses US product codes as primary keys, German market equivalents treated as exceptions requiring manual mapping.
System Level: Bank's fraud detection system analyzes coordination patterns across different accounts and customers within the institution. Network analysis identifies: timing correlations between seemingly unrelated accounts, complementary transaction patterns, shared beneficiaries or intermediaries across multiple customers. Investigation reveals systematic coordination: 5 accounts show synchronized transaction timing, value patterns suggest deliberate structuring to keep individual accounts below risk thresholds,	**System Level:** Agent workflow designed for confidence-based routing: high certainty enables direct resolution, uncertainty triggers human escalation. Workflow creates performance disparity because parsing struggles correlate with German market operations, resulting in 15% escalation rate versus 2% baseline. Training data predominantly contained US addresses and product terminology. Workflow assumption that ambiguity indicates request complexity proves incorrect when ambiguity correlates with market geography. Model

Comparative Example or Step 5 Investigation & Analysis in Financial Services and Agentic AI	
Financial Services: Structuring Detection	**Agentic AI: Customer Service Agent - International Expansion**
relationship mapping identifies shell company connections. Pattern indicates organized activity rather than independent customers.	confidence calibration optimized for US market distribution, not internationally diverse inputs. System-level testing focused on US market scenarios. Market expansion planning underestimated localization requirements across multiple integrated components.
System-of-Systems Level: Cross-institutional information sharing through Section 314(b) agreements reveals broader network. Bank A identifies structured deposits and shares the pattern. Bank B recognizes matching withdrawal patterns in their systems. Bank C identifies international wire transfers with similar timing. Aggregated analysis across institutions reveals multi-national criminal network: 15 accounts across 3 banks, coordinated by same organization, systematic money laundering operation totaling $2M. Pattern invisible to any single institution becomes clear through cross-organizational data sharing and analysis	**System-of-Systems Level:** Agent connects to customer database, shipping coordination system, refund processing system, and product catalog across organizational boundaries. Multiple systems share similar limitations (exact-match requirements, US format optimization), creating compounding effects. Training data pipeline sourced from historical customer service conversations that already exhibited US market bias in address and product terminology. Agent learned and potentially amplified existing patterns. Data pipeline, multiple integration points, and workflow design each optimized for US market without considering international expansion requirements. Pattern spans data science team, integration engineering team, and operations team, each optimized for their context without visibility into cross-system impacts on international scalability.

Preparedness Recommendations

Organizations should establish investigation capabilities before deploying AI systems. Effective investigation cannot be improvised after incidents occur.

- **Investigation and analysis team with multidisciplinary expertise:** Assemble teams including data science expertise, domain expertise in the application area, operational expertise in deployment context, human factors expertise, and systems engineering expertise. Teams should be established and trained before incidents occur.
- **Investigation and analysis methodologies selected and documented:** Choose appropriate investigation approaches (RCA, FMEA, STAMP, Fault Tree Analysis, or others) based on organizational context. Document procedures so teams can execute consistently during incidents.
- **Blame-free culture established and communicated:** Create organizational norms that encourage honest reporting and surface human factors issues. Communicate explicitly that the investigation seeks system improvements rather than individual blame.
- **Procedures for literature review and failure mode collection:** Establish processes for regularly reviewing research publications, security advisories, and practitioner discussions. Allocate time for technical staff to maintain current knowledge of AI failure modes.
- **Access to incident database:** Provide investigation teams with access to previous incident reports within the organization. Analysis benefits from understanding whether current incidents match previous patterns or represent novel failure modes.

- **Documentation systems:** Implement systems for capturing investigation findings, root causes, and insights. Structure documentation to support future analysis and organizational learning.

Without this preparedness infrastructure, organizations lack the expertise, processes, and institutional knowledge needed to conduct systematic investigations that identify true root causes and inform effective corrective actions.

3.2.6 Step 6: Correct

Action: *Implement solutions to address root causes, reduce incident recurrence, and mitigate realized harm.*

Correction transforms investigation findings into tangible improvements.[51] [52] Organizations should address both future prevention and past harm. Effective correction requires multiple types of actions: repairing systems to prevent recurrence, mitigating harm already caused, and updating organizational knowledge to improve both systems and processes.

Not all incidents get corrected

Not every AI incident requires or warrants correction. During the assessment phase, organizations carefully evaluate the severity, potential impact, and resources required to address an incident. Some incidents may be deemed too minor (low severity) to correct. Alternatively, the investigations phase could identify a correction whose cost may outweigh the potential risks. Severity classification frameworks help organizations make strategic decisions about which incidents demand immediate attention and which can be monitored or accepted as acceptable system variations.

System Repair: Preventing Future Incidents

Organizations reduce incident recurrence through several types of repairs. Each type serves different purposes based on incident severity, root cause certainty, implementation complexity, resource requirements, and time needed for permanent solutions. Possible repairs include:

- **Corrective Actions** provide fundamental improvements addressing underlying failure modes rather than treating symptoms. Examples include retraining models with augmented datasets, redesigning workflows to eliminate problematic interaction patterns, changing system architectures to remove structural vulnerabilities, and implementing new monitoring capabilities. Corrective actions typically require significant resources and time but provide lasting reliability improvements.

[51] International Organization for Standardization. (2015). ISO 9001:2015 Quality Management Systems - Requirements. ISO.
[52] Wilson, P.F., Dell, L.D., & Anderson, G.F. (2023). Root Cause Analysis: A Tool for Total Quality Management (3rd Edition). ASQ Quality Press.

- **Workarounds** provide temporary mitigation while permanent fixes are developed when corrective actions require extensive resources or lengthy development cycles. Examples include implementing additional human review steps for affected use cases, restricting system use to reliable contexts, adding verification procedures for problematic scenarios, and establishing approval requirements for sensitive decisions. Workarounds are typically process-oriented.
- **Patches** address immediate symptoms when root cause investigation continues or underlying causes require extended time to resolve. Organizations should clearly identify patches as temporary measures requiring eventual replacement with corrective actions. Examples include input filters screening out problematic patterns, output validators checking for failure signatures, rate limits reducing exposure to failure conditions, and threshold adjustments that modify system behavior.
- **Guardrails** establish boundaries or constraints**,** reducing conditions leading to incidents. Guardrails may remain permanent, serving as defense-in-depth safety measures. Examples include restricting tool access to reduce potential harm, limiting system autonomy in high-stakes decisions, requiring human approval for actions exceeding thresholds, and implementing hard constraints preventing certain failure modes.

Realized Harm Mitigation

Beyond system repair, the correction step addresses harm that has already occurred. This differs from Step 3: Stabilize, which stops ongoing harm. Step 6: Correct addresses harm that happened before stabilization and provides remediation to affected parties. Possible harm mitigation approaches include:

- **Affected Party Notification** informs affected parties of the incident and its resolution, demonstrating accountability and transparency. Effective notification includes an explanation of what happened, a description of corrective actions taken, communication of how similar incidents will be prevented, and establishment of contact points for questions or concerns. Notification timing and content depend on incident severity, regulatory requirements, and affected party characteristics.
- **Correction of Erroneous Decisions** reviews and corrects AI decisions that affected individuals or organizations where appropriate. This may include reversing incorrect loan denials, updating inaccurate records, re-reviewing impacted cases with corrected processes, and providing alternatives where original determinations were inappropriate. Organizations should establish criteria balancing the harm from incorrect decisions against the operational burden of review.
- **Remediation** provides compensation or redress for harm suffered. Financial remediation may include refunds for incident-related charges, service credits for poor quality, expedited processing to compensate for delays, or direct compensation where appropriate and legally permissible. Non-financial remediation may include priority access to improved services, extended support for affected parties, or procedural accommodations. Remediation decisions should consider incident severity, harm magnitude, affected party circumstances, and organizational capabilities.
- **Documentation** records all remediation actions for accountability to affected parties, stakeholders, and regulators. Documentation should record what notifications were sent and when, which decisions were reviewed and corrected, what remediation was provided to whom, and how affected parties were supported through the resolution process.

Organizational Knowledge Updates

During the Correct step, organizations should update policies, training, documentation, and testing protocols based on lessons learned during incident response, transforming individual incidents into institutional learning. Organizations should consider doing the following:

- **Lessons Learned Collection** systematically documents what worked well (detection mechanisms, assessment criteria, stabilization procedures, investigation approaches), what hindered response (preparedness gaps, inadequate tools, team handoff delays, information access difficulties), and surprises encountered (incorrect assumptions, unexpected failure modes, unanticipated interactions). Documentation supports analysis across multiple incidents to identify recurring issues.
- **Policy Updates** revise usage policies clarifying appropriate system use, update operational procedures incorporating incident lessons, enhance oversight requirements for sensitive applications, and document new safeguards. Updated policies are communicated through training, documented in accessible formats, and incorporated into operational guidance.
- **Training Program Revisions** address identified knowledge gaps, document new workarounds or procedures, incorporate lessons about system behaviors discovered during investigation, and update guidance on appropriate system use based on observed failure modes.
- **System Documentation Updates** document newly discovered failure modes and triggering conditions, update descriptions of known limitations based on operational experience, revise operational guidance to avoid identified risks, and clarify capability boundaries that may have been misunderstood.
- **Assessment and Testing Protocol Modifications** add test cases reproducing incident conditions, revise monitoring thresholds based on observed failure patterns, adjust severity criteria if incidents revealed inadequacies, and expand assessment scope to cover previously unconsidered scenarios.
- **Cross-Organizational Sharing** spreads lessons across internal teams deploying similar systems and enables external sharing through industry groups, standards bodies, or research collaborations.[53] External sharing requires agreements protecting sensitive information but enables field-wide learning about common failure modes and effective corrective approaches.

Applying Correction Across Domains

Table 8 illustrates how correction operates in both a mature complex system domain and in agentic AI systems. Both domains require system repairs to prevent recurrence, mitigation of harm already caused, and organizational knowledge updates to improve both systems and processes. Effective correction addresses past harm while preventing future incidents.

[53] Garvin, D.A. (1993). "Building a Learning Organization." Harvard Business Review, 71(4), 78-91.

Table 8: Comparative Example: Correct in Financial Services and Agentic AI. Correction requires multiple types of actions addressing system reliability, harm to affected parties, and organizational learning.

Comparative Example for Step 6 Correct in Financial Services and Agentic AI	
Financial Services: Structuring Detection	**Agentic AI: Customer Service Agent - International Expansion**
Incident Context: Multiple banks detected coordinated structuring pattern: 15 accounts across 3 banks, $2M over 3 months, deliberate distribution to avoid detection thresholds. Investigation revealed sophisticated money laundering network exploiting lack of cross-bank visibility. 200+ legitimate small business accounts incorrectly frozen during initial detection attempts.	**Incident Context:** System monitoring detected performance degradation in German market operations (15% fallback-to-human rate vs. 2% baseline). Quality scores lower in new market (3.2 vs. 4.1 out of 5). Approximately 50,000 customers affected over three months. Root cause: US-centric training data, address parsing optimized for US formats, workflow design assumptions based on US market patterns, insufficient localization planning for market expansion.
Corrective Action: Implement cross-bank pattern detection algorithms. Deploy network analysis capabilities analyzing account relationships and coordination patterns. Lower thresholds for structured transaction detection. Enhance customer due diligence procedures for high-risk account types.	**Corrective Action:** Augment training data with German market examples through targeted data collection and synthetic data generation representing German postal conventions and product terminology. Implement international address parsing with format detection and country-specific parsing rules. Add fuzzy matching for address database lookups with cascade strategies (exact match → fuzzy match → manual review). Redesign workflow to reduce escalation triggered by format ambiguity. Add market-specific confidence calibration. Retrain model with balanced dataset including representative international market data. Complete product catalog localization mapping German terminology to system codes.
Workarounds & Guardrails: Enhanced manual review for transactions matching structuring indicators while automated systems updated. Mandatory human approval for account freezes affecting small business accounts. Temporary lower thresholds with elevated false positive review.	**Workarounds & Guardrails:** Continue enhanced human review for German market interactions during retraining and deployment cycle. Proactive outreach for escalated cases to maintain customer relationships. Expedited resolution pathways for German market queries. Ongoing market-specific performance monitoring with automatic alerts for performance disparities exceeding thresholds. Mandatory international market analysis in all future system testing and deployment planning. Established market readiness assessment checklist for future geographic expansions.
Notification & Remediation: Customer notification explaining transaction review process, alternative payment options during review, expected resolution timeline. Branch manager guidance for handling inquiries. Account unfreezing with expedited processing. Business interruption compensation where	**Notification & Remediation:** Proactive customer outreach to German market customers apologizing for service quality issues and explaining improvement initiatives. Expedited resolution of pending customer service issues. Service credits offered to affected customers. Clear escalation path to specialized German

Comparative Example for Step 6 Correct in Financial Services and Agentic AI	
Financial Services: Structuring Detection	**Agentic AI: Customer Service Agent - International Expansion**
appropriate. Compliance documentation for regulatory examination.	market support team established. Priority handling for German market queries during transition period. Account management outreach to key early adopter customers in market expansion.
Decision Correction: Review all frozen accounts from affected period. Reverse incorrect fraud determinations. Restore account access. Clear negative marks from internal systems. Provide documentation to customers confirming accounts in good standing.	**Decision Correction:** Review all German market cases from affected period involving extended resolution times or customer dissatisfaction. Re-process queries that received suboptimal responses with corrected system. Proactive outreach with corrected information where original responses were inadequate. Update customer interaction records to remove negative service indicators that resulted from system limitations rather than customer issues. Financial remediation where appropriate for service failures.
Policy & Training Updates: Update fraud detection policies to include network analysis requirements. Revise alert investigation procedures for distributed patterns. Train analysts on structuring detection techniques. Document new detection capabilities and procedures. Update testing protocols to include multi-account scenarios.	**Policy & Training Updates:** Revise AI development practices mandating market representation analysis in training data for all international deployments. Require international market readiness testing before launches. Establish market-specific performance monitoring as standard practice for geographic expansion. Train customer service staff on recognizing system limitations in international contexts and appropriate escalation procedures. Empower staff to override system recommendations immediately when market-specific issues identified. Update market expansion procedures to include comprehensive localization assessment across all system components. Create market entry checklist covering data representation, address format handling, product catalog localization, and workflow testing.
Cross Organizational Sharing: Share structuring pattern indicators with other financial institutions through Section 314(b) agreements. Contribute to industry-wide detection capability improvement. Participate in information exchange agreements. Pattern descriptions shared without customer identifiers or proprietary detection methods	**Cross Organizational Sharing:** Share lessons about training data representation challenges and international market localization requirements through AI industry professional organizations. Contribute findings to standards development organizations working on international AI deployment best practices. Participate in information sharing about common technical debt in US-developed systems expanding internationally. Publish anonymized case study on market expansion preparation requirements for AI systems. Contribute to development of market readiness assessment frameworks that benefit industry-wide international expansion efforts.

Preparedness Recommendations

Organizations should establish correction capabilities before deploying AI systems. Effective correction requires processes, authorities, and resources that are difficult to create during incident response.

- **Defined processes for each repair type:** Document procedures for implementing corrective actions, workarounds, patches, and guardrails. Define approval authorities and resource allocation for each type.
- **Clear criteria for selecting repair:** Establish guidelines for selecting among repair types based on incident severity, root cause certainty, implementation complexity, resource requirements, and time constraints.
- **Harm notification and remediation procedures:** Create templates and protocols for notifying affected parties. Establish criteria for decision correction and remediation. Define authorities who can approve financial remediation. Document processes for tracking remediation through completion.
- **Documentation and knowledge management systems:** Implement systems capturing correction actions, remediation provided, and lessons learned. Structure documentation to support analysis across incidents. Ensure documentation accessible to personnel who need it for future incident response and system improvement.
- **Training update processes:** Establish procedures for revising training programs based on incident findings. Define authorities who approve training changes. Create mechanisms ensuring updated training reaches all relevant personnel.
- **External sharing agreements:** Where appropriate, establish agreements with industry groups, standards bodies, or peer organizations for mutual information exchange. Define what information can be shared externally and approval processes for sharing. Protect sensitive information while enabling ecosystem learning.

Without this preparedness infrastructure, organizations struggle to implement timely corrections, fail to mitigate harm to affected parties effectively, and miss opportunities to transform incidents into organizational learning that prevents recurrence.

3.2.7 Step 7: Verify

Action: *Test and validate corrections, then monitor for effectiveness.*

Verification closes the incident response loop by ensuring corrective actions actually improve system reliability rather than introducing new issues or failing to address root causes. Without verification, organizations cannot know if their corrections work as intended or if similar incidents will recur.

Testing Validates Corrections

Testing confirms that corrections mitigate incident recurrence without creating new problems. For AI-enabled systems, verification cannot rely solely on pass/fail tests due to non-deterministic behavior and context-dependency. Systems may produce different outputs for the same inputs, behaviors shift with operational conditions, and system-level behaviors cannot always be predicted from component testing. Organizations should verify corrections through both immediate validation and sustained monitoring. Verification approaches may include:

- **Assessing Distributional Shifts** rather than expecting deterministic improvement.[54] Organizations should measure reduced incident rates, fewer incidents of the same type over time, improved accuracy metrics on previously challenging input categories, reduced false positive and false negative rates in production environments, improved performance on previously problematic input types, and increased robustness to edge cases that previously caused failures. This approach requires statistical analysis over multiple trials to determine whether observed improvements are statistically significant rather than random variation.
- **Immediate Validation** tests corrections before full deployment using staged rollouts that gradually expand to the full user base, A/B testing capabilities comparing corrected versus uncorrected versions where ethically appropriate, close monitoring during initial deployment with defined rollback criteria, and rapid reversion capability if verification reveals problems.[55] Immediate validation catches obvious problems before they affect all users, though some issues only become apparent over extended operation.
- **Sustained Monitoring** confirms corrections remain effective over time and detects delayed issues. Sustained monitoring tracks whether correction effectiveness degrades over time, identifies new failure modes introduced by corrections, and confirms long-term reliability improvement. **Sustained monitoring should integrate with ongoing detection processes in Step 1, creating a continuous improvement loop.**[35]

Processes for Closing Incident Reports

Organizations should require verification confirming correction effectiveness before marking incidents resolved. Closure criteria should include correction implemented and deployed, validation testing completed successfully, monitoring period completed without recurrence, documentation updated to reflect changes, and lessons learned captured and disseminated.

Simply implementing a correction does not constitute incident resolution. Verification provides evidence that the correction actually works and improves system reliability and reduce incidents as intended.

Tracking Reliability and Incident Metrics Over Time

Tracking reliability and incident metrics provides empirical evidence of system improvement,[56] transforming incident response from reactive firefighting into proactive reliability engineering that demonstrates systematic progress to executives, stakeholders, and regulators. Recommendations for reliability tracking include:

- **Select Appropriate Metrics** matching system types and operational contexts. Common metrics include Mean Time Between Incidents, Fix Effectiveness Rate (the proportion of corrected incidents that do not recur), Mean Time to Respond (time from detection to stabilization), and system-specific metrics tailored to applications or use cases. Organizations should select metrics that meaningfully capture system reliability and incident reduction in their operational contexts.
- **Establish Baselines** by measuring system performance before implementing corrective actions. Organizations should document current incident rates, performance metrics, and reliability indicators. Organizations should set improvement targets based on incident severity and organizational risk tolerance to enable comparison showing whether corrections actually improve reliability.

---

[54] Montgomery, D.C. (2019). Introduction to Statistical Quality Control (8th Edition). Wiley.
[55] Kim, G., Humble, J., Debois, P., Willis, J., & Forsgren, N. (2021). The DevOps Handbook (2nd Edition). IT Revolution Press.
[56] O'Connor, P., & Kleyner, A. (2012). Practical Reliability Engineering (5th Edition). Wiley.

- **Track Trends** by monitoring selected metrics over time, comparing performance before and after corrections. Organizations should identify gradual degradation indicating declining correction effectiveness. Organizations should track trends across multiple incidents to determine whether overall system reliability and incident metrics improves beyond resolving individual incidents.
- **Report on Trends** through internal reporting to leadership, demonstrating system improvement, regulatory reporting where required by compliance obligations, public transparency where appropriate to maintain stakeholder trust, and future prevention efforts showing which types of corrections prove most effective.

What if metrics do not improve?

It is possible that an organization completes Step 5 (Investigate & Assess) and Step 6 (Correct) but not system performance is seen during Step 7 or during ongoing monitoring efforts. This most likely means the correction was not effective or the investigation and assessment came to the wrong conclusion.

When this happens, organizations should revisit Step 5 (Investigate & Assess) and Step 6 (Correct). They should consider using a different analysis method for Step 5. They should develop a hypothesis for why the correction was ineffective and develop a new corrective approach.

Applying Verification Across Domains

Table 9 illustrates how verification operates in both a mature complex system domain and in agentic AI systems. Both domains require immediate validation of corrections before full deployment and sustained monitoring to confirm long-term effectiveness. Reliability metrics provide empirical evidence of improvement.

Table 9: Comparative Example: Verify in Financial Services and Agentic AI.Verification requires both immediate validation and sustained monitoring across domains.

Comparative Example for Step 7 Verify in Financial Services and Agentic AI	
Financial Services: Structuring Detection	**Agentic AI: Customer Service Agent - International Expansion**
Incident Context: Implemented cross-bank pattern detection, network analysis capabilities, real-time information sharing through Section 314(b) agreements. Enhanced customer due diligence procedures. Corrected harm to 200+ incorrectly frozen accounts.	**Incident Context:** Implemented training data augmentation with German market examples, international address parsing with format detection and fuzzy matching, workflow redesign to reduce format-triggered escalation, product catalog localization. Enhanced human review during transition. Corrected erroneous responses and provided remediation to affected customers.

Comparative Example for Step 7 Verify in Financial Services and Agentic AI	
Financial Services: Structuring Detection	**Agentic AI: Customer Service Agent - International Expansion**
Immediate Validation: Test new detection algorithms on historical data containing known structuring patterns. Verify algorithms detect previously missed cases. Run simulation with test accounts exhibiting structuring behavior across multiple banks in a controlled environment. Confirm detection without excessive false positives. Staged rollout with tracked detection rates and false positive rates during pilot.	**Immediate Validation:** Test retrained model on held-out dataset with German address formats and product terminology. Measure fallback rates and quality scores across markets. A/B testing: deploy corrected system to 5% of German market traffic initially, compare against baseline system performance. Week 1: 5% traffic, Week 2: 25% traffic, Week 4: 50% traffic, Week 6: 100% traffic if metrics improve. Monitor market-specific performance metrics daily during rollout with defined rollback criteria if performance degrades.
Sustained Monitoring: Track detection rates for structured transaction patterns monthly. Monitor false positive rates by customer segment and geographic region. Quarterly review of structuring techniques identified, assessing whether new variations emerge. Annual assessment of cross-bank information sharing effectiveness. Continue analyst training on evolving structuring methods.	**Sustained Monitoring:** Weekly market-specific performance analysis for first 3 months post-deployment comparing German market to US market baseline. Monthly operational reviews examining fallback rates, quality scores, customer satisfaction across all international markets. Continuous automated monitoring with alerts for cross-market performance disparities exceeding 2 percentage points. Quarterly market expansion readiness assessments for future geographic launches.
Metrics Tracking: Before Correction: Structuring pattern detection rate: 2% of examined networks, Mean Time to Detect: 90 days from first transaction, Cross-bank coordination: detected in 1% of shared cases. **After Correction:** Detection rate >3%, Mean Time to Detect <85 days, Cross-bank coordination detected >2% of cases.	**Metrics Tracking: Before Correction:** Fallback rate disparity: 13 percentage points (15% German market vs 2% US market), Quality score gap: 0.9 points (3.2 vs 4.1), Customer satisfaction gap: 1.2 points, Complaint rate: 3x higher in German market. **After Correction:** Fallback rate disparity <2 percentage points across markets, Quality score gap <0.2 points, Customer satisfaction gap <0.3 points, Complaint rates equalized across markets, Market expansion timeline back on track.
Closure: Incident closed after 6-month monitoring period confirms sustained improvement. Detection capability meets targets. False positive rate acceptable with ongoing optimization. Documentation updated with new detection procedures. Lessons shared. Correction deemed effective.	**Closure:** .Incident closed after 6-month monitoring period demonstrates sustained improvement across international markets. Market-specific performance within acceptable ranges. Automated monitoring confirms ongoing quality consistency. Documentation updated with international market readiness requirements and localization best practices. Lessons incorporated into market expansion playbooks. Correction deemed effective with continued market-specific monitoring as standard practice for all geographic expansions. German market expansion achieves business targets with 8-week delay from original timeline.

Preparedness Recommendations

Organizations should establish verification capabilities before deploying AI systems. Effective verification requires infrastructure, processes, and metrics that cannot be created during incident response.

- **Testing infrastructure:** Implement staging environments where corrections can be tested before production deployment. Establish rollback procedures for rapid reversion if verification reveals problems. Create test datasets representing diverse operational conditions.
- **Reliability metrics defined and automated:** Select metrics appropriate for system types and operational contexts. Implement automated collection and calculation of reliability metrics. Define baseline measurement procedures. Establish statistical significance thresholds for assessing improvement. Document detailed metric definitions and calculation methods.
- **Regular metric review process:** Schedule periodic reviews of reliability metrics with appropriate stakeholders. Define escalation procedures when metrics indicate degradation. Establish accountability for acting on metric trends. Create processes for updating metrics as systems and contexts evolve.
- **Monitoring systems configured:** If necessary, extend detection systems established in Step 1 to track correction effectiveness. Configure alerts for performance degradation or unexpected behaviors following corrections. If relevant, implement focused monitoring for specific populations or business segmentations. Create dashboards visualizing reliability trends over time.
- **Integration with detection systems:** Ensure verification monitoring feeds back into detection systems for continuous improvement. Configure systems to detect similar failure patterns earlier. Update detection thresholds based on verified correction effectiveness. Close the loop between verification and detection.
- **Procedures for incident closure:** Define closure criteria requiring verification evidence. Establish approval processes for marking incidents resolved. Document evidence needed for closure, including validation testing results, monitoring period completion, and absence of recurrence. Create templates for closure documentation.

Without this preparedness infrastructure, organizations cannot determine whether corrections actually work, cannot demonstrate improvement to stakeholders, and cannot close the incident response loop systematically.

3.3 Integration with Existing Frameworks

This framework complements rather than replaces existing standards and frameworks. Organizations already invest in risk management, cybersecurity, and quality assurance processes. The AI incident response framework builds on these investments while extending capabilities for AI-specific characteristics.

This framework complements existing standards and frameworks rather than replacing them. The NIST AI Risk Management Framework[57] provides high-level governance guidance. The NIST Cybersecurity Framework[58] addresses security incident response. ISO standards cover

[57] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.
[58] National Institute of Standards and Technology. (2018). Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1. NIST.

information security and risk management.[59] [60] Where these frameworks identify what organizations should do, this framework provides operational detail for how to implement AI incident response capabilities, with particular attention to AI-specific characteristics like non-determinism, context-dependency, and system-of-systems interactions.

Organizations can integrate AI incident response within existing risk management and IT service management frameworks. However, AI incident response requires extensions to accommodate AI-specific characteristics. Traditional severity classifications need augmentation for factors such as model drift and performance degradation, privacy violations through data leakage, and context-dependent failures. Root cause analysis requires expertise in AI system architectures and behaviors. Verification must account for non-deterministic performance.

Emerging regulatory requirements and sector-specific standards shape implementation. The EU AI Act[61] establishes incident reporting obligations for high-risk AI systems. This framework supports regulatory compliance through standardized reporting structures, systematic incident handling, and clear delineation of responsibilities across developers and deployers. Sector-specific standards in financial services, healthcare, and defense take precedence where they specify more stringent requirements. Organizations subject to multiple regulatory requirements benefit from standardized internal processes that can feed multiple external reporting obligations.

How much to customize

Organizations can adapt response procedures to specific contexts, maintain sector-specific requirements appropriate to their risks, and customize the seven-step loop for their operational needs. They should simultaneously contribute to collective ecosystem capabilities through standardized elements enabling coordination, information sharing, and field-wide learning.

This balance between customization for local needs and standardization for ecosystem benefits represents a core challenge in building mature AI incident response capabilities. Organizations should resist both the temptation to customize everything (losing ecosystem benefits) and the temptation to standardize everything (losing contextual appropriateness). The framework presented in this white paper provides common structure while preserving necessary flexibility

[59] International Organization for Standardization. (2023). ISO/IEC 23894:2023 - Artificial intelligence — Guidance on risk management. ISO.
[60] International Organization for Standardization. (2023). ISO/IEC 42001:2023 - Artificial intelligence — Management system. ISO.
[61] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Integration Strategy

Organizations should integrate AI incident response with existing processes rather than creating isolated systems. Practical integration includes:

- **Leverage Existing Infrastructure:** Use established incident tracking systems, escalation procedures, communication protocols, and evidence preservation processes. Extend these systems for AI-specific data rather than building parallel infrastructure.
- **Extend Severity Classifications:** Augment existing severity frameworks with AI-specific factors, including output quality inconsistencies and systematic error patterns, privacy violation considerations, context-dependent failure assessment, and harm accumulation across interactions. Existing severity levels remain valid but require additional criteria for AI incidents.
- **Augment Root Cause Analysis:** Incorporate methodologies for non-deterministic behavior analysis, multi-level investigation across components through system-of-systems, human factors analysis specific to AI system interactions, and analysis in the context of other reported incidents. Traditional root cause analysis techniques apply but need extension for AI characteristics.
- **Adapt Verification Approaches:** Modify verification to assess distributional shifts rather than pass/fail outcomes, implement staged rollouts with demographic or population-based monitoring, track reliability metrics appropriate for AI systems, and verify correction effectiveness statistically over time.

Section 4: The Ecosystem: Key Stakeholders and Their Roles

Individual organizations can implement the seven-step framework presented in Section 3 and achieve meaningful improvements in their AI incident response capabilities. Organizations with mature processes can systematically detect incidents, respond effectively, and learn from failures to improve reliability over time. However, the full potential of AI incident response emerges when multiple stakeholders coordinate within an interconnected ecosystem.

Ecosystem coordination enables capabilities that individual organizations cannot achieve alone. Pattern recognition across incidents reveals systemic issues invisible within any single organization. Shared learning about failure modes accelerates reliability improvement across the field. Standardized reporting structures enable both computational analysis and regulatory oversight while protecting proprietary information. Independent verification provides credibility for incident disclosures that self-reporting cannot match.

This section explains why ecosystem coordination multiplies the benefits of individual incident response efforts (4.1), distinguishes incident types requiring different stakeholder involvement (4.2), and details the roles and capabilities of six stakeholder categories across the incident response process (4.3). Organizations can begin implementing incident response independently while simultaneously contributing to and benefiting from ecosystem development.

4.1 Why Ecosystem Coordination Multiplies Benefits

Individual organizations face natural constraints that limit what they can see and learn from AI incidents. Developers possess deep technical knowledge but cannot observe how their models perform across diverse deployment contexts. Deployers understand operational environments but may lack expertise for model-level root cause analysis. Users experience system performance in specific contexts but may not recognize AI involvement in decisions. Oversight bodies can mandate reporting but depend on others for incident visibility. Each stakeholder sees only incidents affecting their own users or systems, and no single entity holds complete authority or incident volume needed to identify patterns reliably.

These limitations are inherent characteristics of how AI systems are developed, deployed, and used. AI systems cross organizational boundaries by design. Foundation models serve thousands of deployers. Each deployer configures and customizes systems for specific contexts. Users interact with resulting systems in ways reflecting their unique needs and circumstances. This distributed structure creates blind spots where no single organization can see the complete picture.

What Ecosystem Coordination Enables

When different stakeholders play coordinated roles with clear responsibilities and communication channels, the ecosystem achieves capabilities beyond what any individual organization can accomplish. In particular, an ecosystem can enable:

- **Pattern recognition across incidents.**Individual organizations see incidents affecting only their users or systems. Aggregating incidents across organizations reveals patterns invisible to any single entity. Recurring failure modes in specific operational contexts, cascading failures propagating through interconnected systems, and sophisticated attacks distributed across multiple targets become visible only through cross-organizational analysis.

- **Shared learning about failure modes.** As researchers, developers, and deployers discover new ways AI systems can fail or cause harm, sharing this knowledge accelerates improvement across the field. Organizations benefit from lessons learned elsewhere without having to experience every possible failure mode themselves. This collective intelligence reduces duplicated effort and speeds reliability improvement.
- **Systematic improvement in AI system reliability and incident reduction across the field**. Individual organizational improvements remain isolated without mechanisms for sharing insights. Ecosystem coordination enables field-wide reliability improvements by spreading effective mitigation strategies, successful investigation approaches, and practical preparedness measures across organizations.

Building What Doesn't Yet Exist

This ecosystem does not exist in a mature form. While individual components operate in various sectors and contexts, the coordinated structure with clear roles, established communication channels, and standardized reporting, enabling pattern recognition, requires deliberate development. Building this ecosystem is essential for achieving the full benefits of systematic AI incident response. The following sections describe the stakeholders who should participate and the roles they should play.

The Ecosystem Requires Deliberate Development

This coordinated AI incident response ecosystem does not exist in a mature form. While individual components operate in various sectors and contexts, the coordinated structure with clear roles, established communication channels, and standardized reporting, enabling pattern recognition, requires deliberate development. Building this ecosystem is essential for achieving the full benefits of systematic AI incident response. Organizations can begin implementing incident response independently while simultaneously contributing to and benefiting from ecosystem development.

4.2 Distinguishing Incident Types

Different types of AI incidents require different response approaches. Understanding these distinctions clarifies which stakeholders should lead response efforts and how organizational functions should coordinate. Three primary distinctions shape incident response implementation.

4.2.1 Bad Actors vs. Security Threats vs. Unintentional Harm

Table 10: High-level summary of some common incident types

Common Incident Types			
Incident Type	**Characteristics**	**Response Focus**	**Examples**
Bad Actors	Intentional exploitation	Security, forensics, law enforcement	Prompt injection, data poisoning
Security Threats	Vulnerabilities (may not be exploited)	Containment, patching, hardening	Model inversion, adversarial attacks
Unintentional Harm	No malicious actors	System improvement, testing expansion	Harmful advice, performance degradation

Bad actors intentionally exploit system vulnerabilities or manipulate AI systems for malicious purposes. These incidents require security-focused incident response, forensic investigation capabilities, and potential law enforcement coordination. Response focuses on containment of the threat, evidence preservation for investigation, and attribution where possible. Examples include prompt injection attacks designed to bypass guardrails, data poisoning campaigns targeting training processes, coordinated jailbreaking efforts to extract harmful content, and deliberate manipulation of AI systems for fraud or abuse.

Security threats involve system **vulnerabilities** that could be exploited, whether or not active exploitation has occurred. These incidents require rapid containment, security-focused root cause analysis, and coordination between security teams and AI system owners. Response focuses on patch development, system hardening to prevent exploitation, and vulnerability disclosure procedures that balance transparency with security. Examples include model inversion vulnerabilities that could leak training data, adversarial attack susceptibilities that could compromise decision integrity, authentication bypass possibilities, and data exfiltration risks from deployed systems.

Incidents from **unintentional harm and failures** occur without malicious actors or deliberate exploitation. These incidents require system improvement, expanded testing coverage, and refinement of operational procedures. Response focuses on understanding the failure mechanism, improving system reliability, and preventing recurrence through better design, testing, or deployment practices. Examples include harmful advice, performance degradation in production environments, unexpected behaviors in edge cases the system was not designed to handle, and failures arising from component interactions in complex workflows.

Security and Safety Can Overlap

Many real-world incidents involve both security and safety dimensions simultaneously. A compromised account used to generate harmful content creates both a security incident (unauthorized access) and a safety incident (harmful output). A model vulnerability that enables training data extraction affects both security (exploitable weakness) and safety (privacy violation and information disclosure). This overlap requires coordinated response across organizational functions that traditionally operate separately, with distinct expertise and procedures.

4.2.2 Trusted vs. Untrusted Users

An additional operational distinction shapes incident response implementation: whether incidents involve trusted or untrusted users. AI systems deployed for employees or verified customers face different threat models than systems deployed for the general public. Organizations design systems with different security controls, monitoring thresholds, and response procedures depending on who uses them. The same technical failure may warrant different responses depending on whether it occurred with trusted users in normal operation or untrusted users attempting exploitation. Detection approaches, investigation procedures, and response strategies should account for these different threat environments.

Trusted users include authenticated employees, verified customers, authorized partners, and other actors operating within environments designed for cooperative use. Systems deployed for trusted users make different security tradeoffs, typically prioritizing usability and functionality over restrictive controls. Organizations know the identity of trusted users and can apply consequences for misuse through employment relationships, customer agreements, or partnership contracts.

Detection approaches for incidents involving trusted users often rely on anomaly detection based on known usage patterns. Organizations track normal behavior for authenticated users and flag deviations. Investigation can directly contact users to understand context. Response may involve retraining, clearer guidance, or interface improvements rather than assuming malicious intent.

Untrusted users include external actors, unauthenticated users, insider threats, and potential adversaries without established relationships to the deploying organization. These users may include attackers actively trying to exploit vulnerabilities, competitors seeking to understand system capabilities, or researchers testing system boundaries. Systems deployed in environments with untrusted users require more restrictive controls, stricter monitoring, activity analysis for known bad-actor behaviors, and additional focus on security.

Detection approaches assume potential malicious intent, using stricter thresholds for anomalous behavior and monitoring for known attack patterns. Investigation cannot rely on user cooperation. Response should assume ongoing adversarial activity and design mitigations that function even against determined attackers.

The Blurry Boundary

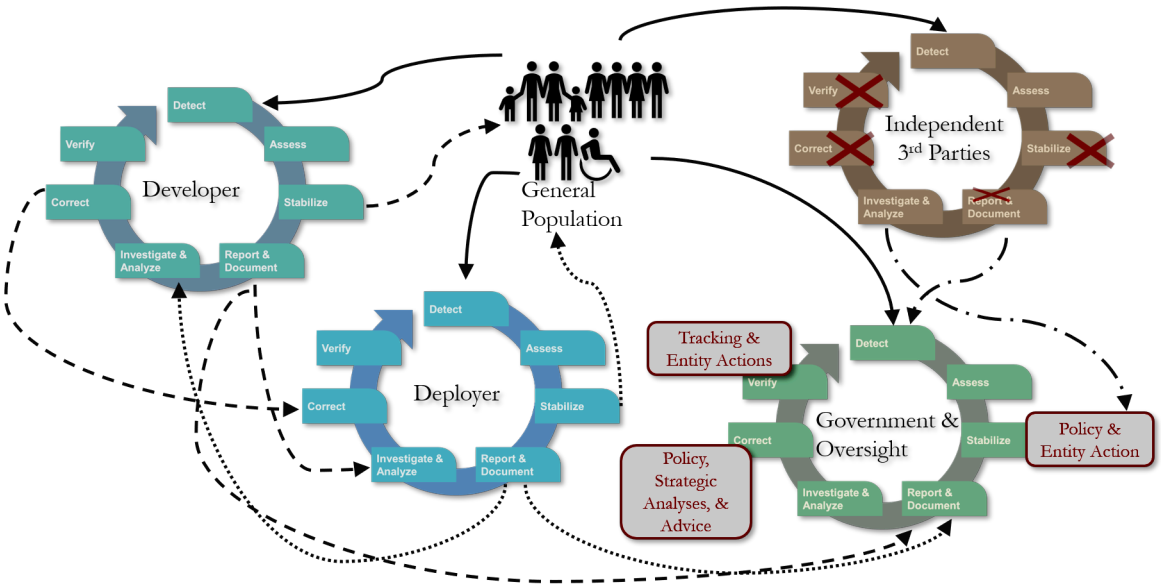
The **distinction between trusted and untrusted users** becomes complicated when bad actors compromise legitimate accounts. Malicious activity appears to originate from trusted sources with valid authentication. This creates situations analogous to "money mules" in financial crime, where legitimate accounts serve illegitimate purposes. Organizations should implement behavioral analysis even for authenticated users, recognizing that credentials alone do not guarantee trustworthy intent.

Compromised accounts create complex detection challenges. Simple authentication checks pass normally. Unusual behavior may reflect legitimate changes in usage patterns rather than compromise. Examination should balance security concerns with user privacy and relationship maintenance. Response procedures should distinguish between users acting maliciously, users whose accounts have been compromised without their knowledge, and users making honest mistakes.

4.3 Key Stakeholders and Their Roles

The six-stakeholder ecosystem enables AI incident response capabilities beyond what individual organizations can achieve alone. Organizations can implement the seven-step framework independently and respond effectively to incidents affecting their systems. However, ecosystem coordination multiplies these benefits through pattern identification across organizations, shared learning about failure modes, and field-wide reliability improvement. This coordinated structure does not yet exist in mature form and requires deliberate development.

Table 11 provides a comprehensive reference showing what each stakeholder can contribute at each step of the incident response process. This table helps organizations to quickly identify which stakeholders to engage for different types of incidents and understand natural coordination points where multiple stakeholders should work together.



Notes for interpreting the figure

- X marks indicate steps where that entity does NOT participate
- Red boxes highlight specialized functions unique to each stakeholder's role
- Arrows and connecting lines show information flow and coordination, feedback between stakeholders
- Developer & Deployer: Implement all 7 steps of the AI incident response loops
- Independent 3rd Parties: Focus on Detect, Assess, Investigate, and Document (but not Report)
- Government & Oversight: Implement all 7 steps, but focus on aggregating patterns across organizations, policy development, tracking, and requirements enforcement
- General Population: Reports incidents and experiences system behaviors

Figure 5: The AI Incident Response Ecosystem showing stakeholder coordination across the seven-step response process.

Table 11: Stakeholder capabilities across the seven-step incident response process. Coordination between stakeholders enables a comprehensive response that is impossible for any single entity.

Stakeholder Capabilities at Each Step in the Incident Response Process							
Stakeholder	**1. Detect**	**2. Assess**	**3. Stabilize**	**4. Report**	**5. Investigate**	**6. Correct**	**7. Verify**
Developers	Internal testing, continuous monitoring, security research	Technical severity, exploitability, cross-deployment impact	Rollback versions, emergency patches, kill switches	Technical specifications, root causes, corrective actions	Deep technical root cause analysis of model internals	Retrain models, redesign architectures, implement guardrails	Technical validation across diverse scenarios
Deployers	User feedback, operational monitoring, compliance audits	Business impact, regulatory requirements, affected populations	Manual processes, restrict access, additional oversight	Operational impacts, business consequences, notifications	System-level analysis, human factors, operational context	Harm mitigation, configuration adjustments, procedure updates	Operational effectiveness, user feedback, business metrics
Users	Direct experience with failures, reporting through channels	---	---	---	Provide context during investigations	---	Validate corrections to address experienced harms
Government & Oversight Bodies	Cross-organizational pattern recognition	Sector-wide impact, regulatory priority	Emergency regulatory action (rare, systemic only)	Public reports on trends and systemic issues	Cross-organizational root causes, systemic factors	Policy changes, regulatory requirements, enforcement	Compliance monitoring, policy effectiveness assessment
Independent Third Parties	Aggregate public reports, voluntary disclosures	Classify using consistent taxonomies	---	Maintain public databases, publish analyses	Cross-organizational pattern research	---	---
Assurance & Audit Orgs,	Discover unreported incidents through audits	Verify severity classifications	---	Audit documentation completeness	Independent review of root cause analyses	Confirm implementation	**Core function:** Independent validation of effectiveness

3.1 AI Developers

AI developers include organizations and teams that create AI models and systems across all applications. This includes foundation model creators, developers building traditional machine learning models for classification and prediction, computer vision system developers, recommendation system developers, and specialists in reinforcement learning and other AI approaches. System integrators assemble AI components into complete solutions. All share deep technical involvement in creating AI systems, whether building from scratch or adapting existing models.

- **Unique capabilities**
  - Deep understanding of system internals including model architectures, training processes, and implementation details. Access to training data, model weights, and system implementation code
  - Ability to perform technical interventions including model retraining, fine-tuning for specific contexts, architectural modifications, and implementation of technical guardrails
  - Technical expertise enabling component-level and model-level root cause analysis
- **What they typically cannot do**
  - Directly observe deployment contexts showing how users actually employ systems in practice
  - Access visibility into actual usage patterns in production environments
  - Understand operational environment factors contributing to incidents without deployer collaboration
  - Directly mitigate harm to end users without working through deployer relationships
  - See system-level and system-of-systems interactions that occur during deployment

Activities Across the Seven-Step Loop

Developers participate in all seven steps of the incident response process, with particular strength in technical analysis and model-level corrections. Their involvement focuses on component-level and model-level issues, detecting problems through internal testing and continuous monitoring, then investigating root causes through deep technical analysis of training data, model architectures, and system interactions. Corrections typically involve model retraining, architectural modifications, or implementation of technical guardrails. Technical validation across diverse scenarios ensures that fixes address root causes without introducing new failure modes. This technical focus complements deployers' operational expertise, creating natural coordination points throughout the response process where model-level insights must integrate with deployment context understanding.

Developer-Deployer Coordination is Essential

Developers need deployer insight into operational incidents that reveal failure modes invisible in developmental testing. Deployers need developer expertise for technical corrections addressing root causes. This natural interdependency creates a coordination requirement. Many incidents require both operational adjustments (deployer responsibility) and technical corrections (developer responsibility) for effective resolution.

4.3.2 Deployers (Organizations Using AI Systems)

AI deployers include enterprises deploying AI for business operations, healthcare organizations using diagnostic AI systems, financial institutions using AI for risk assessment and fraud detection, government agencies using AI for public services, educational institutions using AI for student support, and any organization integrating AI into operational processes. Deployers configure, customize, and integrate AI systems for specific use cases.

- **Unique capabilities**
  - Understanding of operational context including how AI integrates with existing systems, processes, and workflows
  - Knowledge of who uses systems and for what purposes, what operational constraints apply, and what consequences incidents create
  - Authority over operational procedures, system configurations, and deployment choices
  - Visibility into real-world consequences of AI decisions and behaviors
  - Ability to observe system-level and system-of-systems interactions that occur in production
- **What they typically cannot do**
  - Perform deep technical debugging of model internals without developer support
  - Access training data, training processes, and technical documentation necessary for component-level root cause analysis
  - Retrain models, modify model architectures, or implement certain technical guardrails requiring changes to model behavior rather than deployment configuration
  - See how design decisions made during model development affect system behavior in ways requiring architectural changes
  - Distinguish whether an incident requires model-level correction or deployment-level adjustment without developer collaboration

Activities Across the Seven-Step Loop

Deployers engage across all seven steps with a focus on operational response and business impact mitigation. Their proximity to users enables rapid detection through feedback channels and operational monitoring, while their authority over system configurations allows immediate stabilization through manual processes, access restrictions, or additional oversight. Assessment activities evaluate business impact, regulatory requirements, and affected population characteristics. System-level investigation analyzes how failures propagate through workflows, examines human factors and operational context, and identifies patterns across multiple incidents in their environment. Corrections address both system improvements and harm mitigation for affected users, while verification confirms operational effectiveness through user feedback and business metrics. This operational focus complements developers' technical expertise, with coordination essential for distinguishing deployment-level issues from model-level problems requiring architectural changes.

Configuration and Customization

Many deployers fine-tune models for specific use cases, configure systems for operational environments, customize applications for business needs, and integrate AI with existing systems. These activities make deployers partially responsible for system behavior. Distinguishing developer responsibilities from deployer responsibilities requires understanding what was configured during deployment versus what is inherent in the model.

Multiple Deployers Per Developer

One developer's foundation model may serve thousands of deployers, each with different operational contexts, use cases, user populations, and performance requirements. The same model may work well for some deployers while failing for others due to differences in deployment context. This creates coordination challenges at scale. Developers cannot customize for every deployer. Deployers should configure appropriately for their contexts.

4.3.3 Users

Users encompass anyone who interacts with or is affected by AI systems, whether external or internal to the deploying organization. External users include customers, service recipients, and members of the general public affected by AI decisions. Internal users include employees using enterprise AI for business operations, decision support, data analysis, or other organizational functions. Both groups experience AI system performance in authentic operational contexts and provide essential feedback about real-world behavior, though their relationships to the deploying organization and available reporting channels may differ significantly.

- **Unique capabilities**
  - Direct experience with system failures, unexpected behaviors, harmful outputs, or discriminatory decisions
  - Ability to report through feedback mechanisms, help desks, complaint processes, or regulatory channels
  - Real-world context about how AI systems perform in authentic operational environments
  - Diverse perspectives reflecting different populations, use cases, and operational environments
- **What they typically cannot do**
  - Participate in Assess, Stabilize, Report, or Correct steps requiring organizational authority, technical expertise, or access to system internals

Activities Across the Seven-Step Loop

Users contribute primarily through detection, investigation context, and validation of corrections. While they do not perform technical response activities, they have direct experience with the impact of AI incidents. Users typically do not participate in Assess, Stabilize, Report, or Correct steps, as these require organizational authority, technical expertise, or access to system internals that users do not possess. However, their contributions to Detect, Investigate, and Verify steps are essential for effective incident response.

Challenges

Users may not recognize when AI is involved in decisions, limiting their ability to report AI-related issues. They may lack the technical knowledge needed to describe problems effectively. Barriers to reporting include unclear channels, time requirements, and concerns about consequences. Available recourse when harmed by AI systems often remains unknown to affected individuals.

Despite these challenges, users provide irreplaceable detection signals and validation that corrections work in practice. Effective incident response depends on treating users as essential stakeholders whose experiences and feedback drive improvement.

4.3.4 Government and Oversight Bodies

Government and oversight bodies include regulatory agencies with authority over specific sectors, consumer protection agencies, civil rights enforcement bodies, data protection authorities, and sector-specific regulators across different jurisdictions. Additionally, they occupy a position no other stakeholder can fill: cross-organizational visibility combined with enforcement authority. This combination enables pattern recognition at scale while providing mechanisms to drive systemic change.

- **Unique capabilities**
  - Authority to set requirements and enforce compliance through regulatory action
  - Ability to mandate incident reporting across many organizations, enabling access to incident data at scale
  - Capacity to conduct cross-organizational pattern analysis revealing systemic issues
  - Authority to take regulatory action directing specific entities to correct problems
- **What they typically cannot do**
  - Stabilize individual operational incidents
  - Access technical details necessary for deep component-level analysis without requiring disclosure
  - Implement corrections directly (may direct regulated entities to take action)
  - Verify technical correction effectiveness without independent evaluation capabilities

Activities Across the Seven-Step Loop

Government and oversight bodies operate at a different level than individual organizations. While they may respond to egregious individual AI incidents, they more typically run their AI incident response process on aggregated incident patterns identified across multiple organizations. Their activities fall into three complementary areas: aggregating and analyzing incident data, setting requirements for organizational response, and responding through policy and enforcement.

- **Aggregating and Analyzing:** Oversight bodies could collect incident data across organizations within their jurisdiction. They can identify cross-organizational patterns revealing systemic issues, detect sector-wide trends requiring policy attention, track industry-wide metrics, and analyze the effectiveness of regulatory requirements.
- **Setting Requirements:** Oversight bodies could recommend or mandate specific incident response processes appropriate to sector risks. They can require reporting of certain incident types meeting defined thresholds, set AI standards that AI systems must meet, and provide compliance guidance.
- **Responding Through Policy and Enforcement:** Oversight bodies could develop policy based on AI incident patterns seen across organizations. They could take regulatory action directed at specific entities violating requirements, issue strategic guidance for sectors based on lessons learned, publicly report on incident trends, and coordinate with other jurisdictions on shared challenges.

Current State of Government Involvement

Government involvement in systematic AI incident response remains limited globally. Most regulatory activity occurs through enforcement actions rather than systematic incident collection and analysis. **The infrastructure for cross-organizational pattern recognition by**

government bodies largely does not yet exist. This is beginning to change as new regulatory frameworks take effect. However, even with new regulatory frameworks, institutions may not have the infrastructure to perform cross-organizational pattern recognition or participate in the AI incident response loop in a manner that mitigates future incidents or improves AI reliability.

The European Union's AI Act, which enters into force in stages through 2027[62], represents the most developed government incident reporting system globally. Starting in August 2026,[63] national market surveillance authorities in EU member states will receive reports of serious incidents from providers of high-risk AI systems. These reports should be submitted within 2 to 15 days depending on incident severity. Serious incidents include those causing death or serious health harm, serious and irreversible disruption to critical infrastructure, infringements of fundamental rights protections, or serious harm to property or the environment. This system will provide the first large-scale example of government bodies systematically receiving and analyzing AI incident data across organizations and sectors.

In other jurisdictions, regulatory agencies respond to AI incidents primarily through case-by-case enforcement rather than systematic collection. Consumer protection authorities use existing statutory powers to take action against companies whose AI systems harm consumers through deceptive practices, unfair treatment, or discriminatory outcomes. Civil rights enforcement bodies apply anti-discrimination laws when AI systems used in employment, housing, or other consequential decisions create discriminatory effects. Financial regulators address AI-related issues in their sectors, and healthcare regulators oversee AI in medical applications. However, these enforcement activities generally occur reactively in response to complaints rather than proactively based on pattern analysis across systematically collected incident data.

Some jurisdictions have introduced targeted legislative[64] requirements for specific AI applications. Requirements for algorithmic impact assessments, bias audits for hiring tools, and transparency obligations for automated decision-making systems create compliance frameworks that regulators can enforce. These requirements often include documentation and reporting obligations that could support more systematic incident tracking, though comprehensive incident reporting infrastructure remains underdeveloped outside the EU framework.[65]

Dependency on Standardized Reporting

Effective oversight depends on access to incident data collected in forms enabling pattern recognition and trend analysis. Standardized incident reporting structures facilitate this function. Without structured, analyzable data, oversight bodies cannot detect patterns, cannot identify systemic issues requiring attention, and cannot develop evidence-based policy. The design of reporting requirements profoundly affects oversight effectiveness.

---

[62] European Commission Draft Guidance (September 26, 2025): Article 73

[63] IAPP Timeline Resource: High-risk AI incident reporting by August 2, 2025 [refers to preparatory obligations]; practical implementation of high-risk AI requirements by February 2, 2026; entry into application August 2, 2026, https://iapp.org/resources/article/eu-ai-act-timeline/

[64] States Ring in the New Year with Proposed AI Legislation by: Adam S. Forman, Greta Ravitsky, Elizabeth S. Torkelsen, Jennifer Stefanick Barna Epstein Becker & Green, P.C. - Workforce Bulletin Tuesday, January 21, 2025 https://natlawreview.com/article/states-ring-new-year-proposed-ai-legislation

[65] EU AI Act Article 73 on serious incident reporting, https://artificialintelligenceact.eu/article/73/

4.3.5 Independent Third-Party Organizations

Independent third-party organizations operate outside both government and industry structures to support the AI incident response ecosystem.[66] These include incident collection and monitoring organizations such as the AI Incident Database (AIID)[67] [68] and the OECD AI Incidents and Hazards Monitor (AIM),[69] specialized databases tracking specific incident types including deepfake incidents and legal AI issues, academic research groups analyzing AI failures and incident patterns, professional organizations such as IEEE that convene practitioners and develop guidance, and standards development organizations such as ETSI and the OECD that create technical standards and reporting frameworks for AI systems.

- **Unique capabilities**
  - Operation outside both government and industry structures, enabling flexibility and freedom from direct conflicts of interest
  - Ability to move quickly, where government agencies may face bureaucratic constraints
  - Capacity to maintain transparency where companies face competitive pressures limiting disclosure
  - Ability to aggregate data across organizations, sectors, and geographies, enabling pattern recognition at scale
  - Provide public resources, including databases, research findings, and standards that benefit the entire field
- **What they typically cannot do**
  - Perform stabilization, correction, or verification steps
  - Access proprietary technical details necessary for deep analysis without voluntary disclosure
  - Compel incident reporting from organizations
  - Directly enforce corrections or verify that organizations implement effective responses

Activities Across the Seven-Step Loop

Independent organizations typically do not perform stabilization, correction, or verification steps because they are not in operational roles with authority over deployed systems. Their contributions focus on transparency, pattern recognition, and standards development rather than operational response. They perform three complementary functions that support the broader ecosystem.

First, they collect and aggregate incident information from diverse sources. The OECD's AIM tracks incidents in real time through automated analysis of over news sources worldwide. The OECD Expert Group on AI Incidents has developed standardized definitions for AI incidents

---

[66] Note that this role differs from trusted intermediaries that handle sensitive proprietary data under protected sharing agreements.

[67] The AI Incident Database (AIID), operated as an independent third-party initiative, serves as a pioneering model for cross-organizational incident collection. By establishing the first comprehensive repository of AI incidents from diverse sources, AIID demonstrates the feasibility of aggregated incident tracking while maintaining contributor confidentiality. Its operational precedent addresses the collective action problem inherent in incident sharing: organizations hesitate to report in isolation, but AIID's existing infrastructure and dataset lower barriers to participation and enable pattern recognition that no single organization can achieve alone. AIID's work has influence the OECD's work on defining AI incidents See https://incidentdatabase.ai.

[68] McGregor, S. (2021). "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database." Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 15458-15463.

[69] OECD collects incidents through its AI Incident Monitor (AIM), https://oecd.ai/en/incidents-methodology

and related terminology to foster international interoperability. Other databases such as the AI Incident Database and sector-specific repositories document incidents from news media, research papers, voluntary organizational submissions, and public disclosures. These databases enable cross-organizational pattern analysis without requiring regulatory authority.

Second, standards development organizations convene stakeholders to create technical frameworks. The OECD is developing a common reporting framework for AI incidents to enable consistent and interoperable reporting globally.[70] ETSI conducts working groups developing technical standards for AI systems. Professional organizations create guidance documents and facilitate information sharing among practitioners.

Third, researchers analyze aggregated incident patterns, develop taxonomies and classification schemes, and publish findings that inform policy discussions with empirical evidence. Together, these activities help identify emerging trends, new incident types, and systemic issues while building the knowledge base needed for effective AI governance.

Limitations

Independent organizations rely on publicly available information, voluntary submissions, or volunteer participation, limiting their access and involvement in incident response. They may lack access to proprietary technical details necessary for deep analysis. They cannot compel incident reporting from organizations. They cannot directly enforce corrections or verify that organizations implement effective responses. Additionally, while they can document and investigate AI incidents, they are typically not reporting new incidents.

Despite these limitations, independent third parties play crucial roles in aggregate analysis, transparency, research advancing the field, and standards development based on lessons learned.

Future Potential: ISACs for AI

Information Sharing and Analysis Centers (ISACs) operate successfully in sectors facing analogous challenges: financial services (FS-ISAC),[71] healthcare (H-ISAC),[72] aviation (A-ISAC),[73] automotive (Auto-ISAC),[74] and energy (E-ISAC).[75] These member-based professional organizations actively collect, analyze, and share incident data among members while protecting proprietary information and competitive sensitivities.

ISACs provide trusted venues where member organizations can report incidents confidentially, often anonymously. Pattern recognition occurs across member data without requiring public disclosure of individual incidents. Members receive real-time threat intelligence based on patterns detected across the membership, enabling collective defense while protecting competitive information. This model solves a critical challenge: organizations gain the benefits of shared intelligence without the risks of public disclosure.

Similar structures could develop in the AI sector, complementing existing independent collectors like the AI Incident Database and OECD's AI Incidents and Hazards Monitor. AI-focused

---

[70] https://oecd.ai/en/site/incidents
[71] https://www.fsisac.com/
[72] https://health-isac.org/
[73] https://www.a-isac.com/
[74] https://automotiveisac.com/
[75] https://www.eisac.com/s/

ISAC[76]-like centers could provide member-only incident sharing with more complete technical details than public databases allow, aggregated public analysis that informs the broader ecosystem without exposing individual organizations, real-time alerts about emerging threats based on member reports, and sector-specific coordination tailored to particular AI applications or industries. This approach would build on the proven financial services model while addressing AI-specific coordination needs.

4.3.6 Assurance Organizations and Auditors

Assurance organizations and auditors include internal audit teams[77] within organizations deploying AI, external auditing firms providing independent verification, specialized third-party assurance providers focusing on AI systems, AI audit specialists with technical expertise, and compliance verification organizations validating regulatory adherence.[78]

Assurance organizations do not respond to incidents directly but provide independent verification that strengthens the reliability of incident response across the ecosystem. Their audits serve multiple audiences: they give developers and deployers confidence that their processes function as intended, provide regulators with verified evidence for compliance and enforcement decisions, and offer stakeholders credible assurance that incident disclosures are trustworthy. By validating both preparedness infrastructure before incidents occur and response effectiveness after incidents are resolved, assurance functions create accountability mechanisms that bridge operational entities and oversight bodies. As AI assurance evolves toward fiduciary accountability under proposals like O'Reilly and Strauss's framework (discussed below), this verification role becomes integral to corporate governance rather than an optional technical review

- **Unique capabilities**
  - Independent verification and validation separate from operational responsibilities
  - Audit expertise and methodologies proven in other domains
  - Objectivity through professional standards and independence requirements
  - Credibility with regulators and stakeholders through established reputations
  - Expertise in evidence gathering, assessment procedures, and verification approaches
- **What they typically cannot do**
  - Perform stabilization (not in operational roles)
  - Access systems and data without organizational cooperation
  - Compel corrective actions (can only recommend and verify)
  - Conduct real-time monitoring (typically perform point-in-time audits)

Activities Across the Seven-Step Loop

Assurance organizations operate selectively across the incident response process, concentrating where independent verification adds value. Through scheduled audits, they may discover unreported incidents and validate that severity classifications follow documented frameworks. Documentation reviews assess completeness and accuracy of incident reports, while independent analysis examines whether root cause investigations employed rigorous methodologies.

---

[76] U.S. Department of Homeland Security. (2015). "Information Sharing and Analysis Centers: Best Practices Guide." DHS National Protection and Programs Directorate.
[77] Raji, I.D., et al. (2020). "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44.
[78] Lam, M.K., et al. (2024). "A Framework for Assurance Audits of Algorithmic Systems." Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 1124-1137.

Confirmation that documented corrective actions were actually implemented in operational systems represents a critical checkpoint, though the primary assurance function centers on verification. Independent validation of correction effectiveness provides third-party confirmation that reliability improvements are accurately measured and reported. Notably, assurance organizations do not perform stabilization, as this requires operational authority they do not possess. Their role remains focused on verification rather than execution, creating accountability bridges between operational entities and oversight bodies.

Critical Role

As AI systems become embedded in regulated and high-stakes contexts, assurance functions provide verification that:

- Incident disclosures are complete and accurate
- Model behaviors meet established standards
- Mitigation measures are effective
- Incident response processes function as documented
- Organizations meet reliability, safety, and fairness standards

The role of AI assurance is expanding beyond technical verification. Tim O'Reilly and Ilan Strauss[79] [80] propose that AI systems, including their incidents, dependencies, and risks, should be disclosed as part of company financial and operational reporting. This perspective reframes AI assurance not merely as a technical or ethical responsibility but as a matter of fiduciary accountability. If AI systems contribute to productivity, liability, or reputational risk, their performance and incidents should be subject to the same verification, audit, and disclosure standards as other material business operations.

Integrating AI incident data into financial reporting would align incentives across regulators, auditors, and boards. AI reliability and transparency would become components of corporate governance rather than after-the-fact remediation. Assurance organizations would provide the verification function necessary for boards and shareholders to rely on AI-related disclosures.

Relationship to Other Stakeholders

Typically, assurance organizations and auditors do not have their own independent response loop processes. Instead, they support AI incident response for other entities.

- **With Developers and Deployers:** Assurance organizations audit incident response processes and capabilities, validate accuracy of incident reports, verify effectiveness of corrections, and provide attestation for compliance with requirements.
- **With Government and Oversight:** Assurance organizations provide independent verification for regulatory compliance, supply evidence for enforcement actions, support policy development with audited data, and enable oversight bodies to rely on verified information rather than unaudited organizational claims.

---

[79] Ilan Strauss, Isobel Moure, Tim O'Reilly and Sruly Rosenblat. "Real-World Gaps in AI Governance Research: AI Safety and Reliability in Everyday Deployments." SSRC AI Disclosures Project Working Paper Series (SSRC AI WP 2025-04), Social Science Research Council, April 2025. https://www.ssrc.org/publications/real-world-gaps-in-ai-governance-research/

[80] Strauss, Ilan, O'Reilly, Tim, Rosenblat, Sruly, and Isobel Moure. "Governing AI Through SEC Disclosure." Working Paper. Social Science Research Council, October 2025. https://www.ssrc.org/publications/governing-ai-through-sec-disclosure-materiality-standards-and-incident-reportinglessons-from-cybersecurity/

- **With Independent Third Parties:** Assurance organizations may use aggregated incident data for benchmarking and validation, contribute to standards development based on audit experience, and inform best practices with findings from verification activities.

Additionally, assurance organizations can help organizations develop severity frameworks before incidents occur, assist with developing investigation capabilities, verify preparedness infrastructure before systems are deployed, and conduct readiness assessments identifying gaps in incident response capabilities.

The addition of assurance and audit functions to the AI incident response ecosystem reflects the maturing recognition that AI systems require the same governance, verification, and accountability structures as other critical business operations.

4.3.7 Ecosystem Coordination

Figure 5 illustrates how stakeholders interact within the incident response ecosystem. The diagram shows which entities can perform each step of the seven-step process, the information flows between stakeholders, and the distinct loops that characterize operational response (developers and deployers), oversight activities (assurance organizations and government bodies), and ecosystem learning (independent third parties aggregating and analyzing patterns).

As described in Section 4.1, coordinated ecosystem action unlocks capabilities impossible for individual organizations: pattern recognition revealing systemic issues, shared learning accelerating field-wide reliability improvement, and standardized reporting enabling both computational analysis and regulatory oversight. Independent verification provides credibility for incident disclosures. Together, these coordination mechanisms transform incident response from reactive firefighting within organizations into proactive, systematic reliability improvement across the entire AI ecosystem.

Section 5. Building the Ecosystem: Recommendations for Action

This framework and the discussed ecosystem require coordinated action across multiple stakeholders. Organizations deploying or developing AI systems should begin implementing systematic incident response processes. Assurance organizations should develop verification capabilities for AI-specific risks. Standards bodies, regulators, and professional organizations should build the infrastructure enabling ecosystem-wide learning. While individual organizations can implement incident response independently, the full benefits (pattern identification, shared learning, and field-wide reliability improvement) depend on ecosystem-level coordination.

5.1 For AI Deployers

Organizations deploying AI systems are primarily responsible for operational incident response and mitigating user harm. Deployers should build incident response capabilities proactively, before incidents occur, rather than developing processes during active crises.

Immediate Priorities

Deployers should focus initial efforts on three foundational capabilities:

- **Detection infrastructure**: Implement monitoring of AI system outputs and performance metrics, create accessible user feedback channels, and designate personnel responsible for receiving and triaging incident reports
- **Severity assessment frameworks**: Adapt established severity classification systems such as MIL-STD-882E to operational context, document escalation criteria accounting for AI-specific factors like hallucinations and accumulated harm, and train personnel on applying assessment criteria consistently
- **Incident documentation**: Create standardized incident report templates, establish secure storage for incident records, and define notification procedures for internal stakeholders and external parties

Integration with Existing Processes

Deployers with mature IT or security incident response should extend existing capabilities rather than building parallel systems:

- Adapt severity criteria to include AI-specific harms such as hallucinations, quality degradation, and privacy violations
- Augment root cause analysis methodologies to account for non-deterministic behavior and context-dependent failures
- Modify verification processes to assess distributional shifts rather than deterministic pass/fail testing
- Bridge organizational divides between security teams, engineering teams, and operational units

Preparedness Infrastructure

Beyond reactive response, deployers should invest in preparedness:

- Document likely failure modes before deployment and develop stabilization procedures for each
- Establish multidisciplinary investigation teams including data science, domain expertise, operational knowledge, and human factors specialists

- Create backup systems or manual processes enabling rapid failover during incidents
- Develop harm notification and remediation procedures for affected users
- Track reliability metrics demonstrating system improvement over time

Ecosystem Participation

Deployers should contribute to collective learning while protecting competitive interests:

- Participate in sector-specific information sharing arrangements when available
- Consider voluntary incident reporting to independent databases supporting research and standards development
- Engage with standards bodies developing incident reporting frameworks
- Share lessons learned through professional organizations while protecting proprietary details

5.2 For AI Developers

AI developers possess unique technical capabilities essential for understanding and correcting model-level failures. However, developers typically lack visibility into how systems perform in operational contexts, creating natural dependencies with deployers.

Technical Response Capabilities

Developers should build infrastructure enabling rapid technical response:

- **Model monitoring and rollback**: Maintain version control for models and ability to rapidly roll back to previous versions, implement automated monitoring for model behavior changes, and create emergency patch processes
- **Root cause analysis for models**: Develop methodologies for investigating non-reproducible failures, build tools for analyzing training data contributions to failures, and create processes for identifying architectural vulnerabilities
- **Testing and validation**: Establish staging environments for testing corrections before deployment, implement A/B testing capabilities for comparing corrected versus original models, and develop metrics for assessing distributional shifts in model behavior

Coordination with Deployers

Effective incident response requires information flow between developers and deployers:

- Create channels for deployers to report operational incidents with sufficient technical detail
- Provide deployers with documentation about known limitations and failure modes
- Establish service level agreements for technical support during incidents
- Share information about model updates and their potential operational impacts

Vulnerability Disclosure

Developers should implement coordinated disclosure processes for security vulnerabilities:

- Create clear channels for security researchers to report vulnerabilities
- Establish timelines for developing and distributing patches
- Coordinate with affected deployers before public disclosure
- Document lessons learned for improving future model development

Proactive Reliability Investment

Beyond reactive incident response, developers should invest in reducing incident frequency:

- Conduct regular output quality audits across different input types and use cases
- Test for robustness against adversarial inputs and edge cases
- Implement guardrails limiting harmful outputs
- Monitor deployed systems for performance degradation
- Incorporate lessons from reported incidents into development processes

5.3 For Assurance and Audit Organizations

As AI systems become embedded in regulated and high-stakes contexts, assurance functions provide independent verification that incident response processes function effectively and produce trustworthy outputs. This represents an emerging market opportunity as regulatory requirements and corporate governance expectations expand.

Core Verification Capabilities

Assurance organizations should develop expertise in auditing AI incident response:

- **Preparedness assessment**: Verify that severity frameworks are appropriate for operational contexts, confirm investigation teams possess necessary expertise, validate that stabilization procedures are tested and documented, and assess whether monitoring infrastructure can detect relevant failures
- **Process audit**: Confirm organizations follow documented incident response procedures, verify incident reports are complete and accurate, review root cause analyses for methodological rigor, and validate that corrective actions were actually implemented
- **Effectiveness verification**: Independently test whether corrections achieved intended effects, confirm reliability metrics accurately reflect system improvements, and provide third-party validation for stakeholders and regulators

Integration with Corporate Governance

Following proposals by O'Reilly and Strauss, AI assurance should evolving from optional technical review toward fiduciary accountability:

- AI systems and their incidents should be disclosed in company financial and operational reporting when they contribute to productivity, liability, or reputational risk
- Boards and executives require independent verification of AI system reliability and incident response effectiveness
- Audit firms should develop AI assurance practices comparable to financial audit capabilities
- Internal audit functions should expand scope to include AI risk management and incident response

Developing Audit Programs

Assurance organizations should create structured audit programs covering:

- Adequacy of incident detection mechanisms, including both automated monitoring and user feedback channels
- Appropriateness of severity classification frameworks for the operational environment
- Completeness and accuracy of incident documentation

- Rigor of root cause investigation methodologies
- Implementation of documented corrective actions
- Effectiveness of verification testing and ongoing monitoring

Regulatory and Standards Engagement
Assurance organizations should participate in developing audit standards:

- Contribute to standards bodies defining AI assurance requirements
- Develop consensus methodologies for AI system audits
- Share best practices through professional organizations
- Provide regulators with insights about audit feasibility and effectiveness

5.4 For Standards Bodies

Standards organizations provide the technical frameworks enabling interoperability and consistency across jurisdictions and sectors.

ETSI (European Telecommunications Standards Institute):

- Continue developing common reporting standards for AI incidents
- Harmonize with OECD, ISO, and NIST efforts to ensure international interoperability
- Engage AI developers and deployers in standards development processes
- Pilot standards with early adopters to refine practical implementation

ISO (International Standards Organization):

- Develop AI incident response standards building on ISO 27001 (Information Security Management) and ISO 31000 (Risk Management)
- Coordinate with NIST, ETSI, and OECD to ensure consistency across frameworks
- Create sector-specific guidance adapting general frameworks for financial services, healthcare, defense, and other regulated domains

NIST (National Institute of Standards and Technology):

- Expand AI Risk Management Framework guidance to include detailed incident response specifications
- Develop technical standards for AI system monitoring and logging
- Create guidance on severity classification for AI incidents
- Publish reference implementations and case studies

IEEE (Institute of Electrical and Electronics Engineers):

- Develop best practices and implementation guidance for practitioners
- Create training and certification programs for AI incident responders
- Facilitate practitioner communities for sharing lessons learned

5.5 For Regulators

Regulatory bodies have unique authority to mandate incident response processes, establish reporting requirements, and create incentives for systematic approaches.

Establish Clear Requirements:

- Define mandatory incident reporting for high-risk AI systems based on sector, application, and potential impact
- Specify minimum incident response capabilities appropriate to system risk levels
- Set reporting timelines and formats aligned with international standards
- Provide safe harbor provisions for good-faith reporting to encourage transparency rather than concealment

Enable Information Sharing:

- Create legal frameworks permitting confidential incident sharing among organizations
- Establish or designate entities with aggregation and analysis capabilities
- Publish cross-organizational pattern analyses informing sector-wide learning while protecting individual organizational identities
- Protect proprietary information while enabling collective intelligence

Provide Guidance and Resources:

- Publish incident response frameworks and templates organizations can adapt to specific contexts
- Offer implementation guidance tailored to different sectors and organization sizes
- Create maturity assessment tools organizations can use for self-evaluation
- Fund research on incident patterns, root causes, and effective response practices

Coordinate Across Jurisdictions:

- Harmonize reporting requirements internationally to reduce compliance burden on global organizations
- Share incident data across regulatory bodies through secure channels
- Avoid conflicting requirements creating impossible compliance situations
- Enable global AI deployments with consistent incident response expectations

Build Analytical Capacity:

- Develop capabilities for pattern recognition across incident reports from multiple organizations
- Invest in technical expertise enabling analysis of AI-specific failure modes
- Create systems for tracking sector-wide reliability trends over time
- Use aggregated incident data to inform evidence-based policy development

5.6 For Professional Organizations

Professional organizations can facilitate information sharing, build practitioner communities, and develop consensus practices that complement formal regulatory requirements.

Information Sharing and Analysis Centers (ISACs) for AI:

ISACs have proven effective for sharing cybersecurity information in financial services, healthcare, aviation, automotive, and energy. Similar structures could serve the AI sector:

- Establish member-based organizations for confidential incident sharing among competitors
- Provide anonymous reporting mechanisms protecting organizational identity
- Aggregate and analyze patterns across members to identify emerging threats
- Distribute threat intelligence and defensive guidance to membership
- Enable more complete incident information sharing than public disclosure allows

AI-focused ISAC-like centers would complement public databases like the AI Incident Database and OECD AI Incidents Monitor by providing trusted venues where organizations share sensitive technical details, competitive concerns do not prevent reporting, and members receive actionable intelligence about threats affecting their systems.

Training and Education:

- Develop incident response training programs addressing AI-specific characteristics including non-determinism, context-dependency, and emergent behaviors
- Provide toolkits and resources for organizations implementing the framework
- Build practitioner communities enabling ongoing knowledge exchange

Standards and Best Practices:

- Convene stakeholders from industry, government, academia, and civil society to develop consensus approaches
- Document case studies and lessons learned for broader dissemination
- Publish guidance materials incorporating field experience and evolving understanding

Section 6. Conclusion

Organizations deploying AI-enabled systems face an urgent need for systematic incident response processes. Ad hoc responses fail to build institutional knowledge, enable learning from failures, or support continuous reliability improvement. This white paper presents a comprehensive framework adapting proven reliability engineering practices from complex systems domains to AI-enabled systems. The framework is intentionally generalizable, enabling organizations to customize the seven-step process for their operational contexts while maintaining compatibility with broader ecosystem coordination. The paper provides tailored guidance for each stakeholder category (developers, deployers, users, oversight bodies, independent evaluators, and assurance organizations), clarifying distinct roles and responsibilities across the incident response ecosystem. This ecosystem approach enables pattern recognition, shared learning, and systematic reliability improvement that no individual organization can achieve alone.

6.1 The Framework

This framework treats AI-enabled systems as what they are: complex systems requiring systematic incident response processes. Decades of experience in aerospace, financial services, healthcare, and critical infrastructure demonstrate that complex systems demand preparation before incidents occur, systematic investigation when they happen, and continuous improvement based on lessons learned. AI systems benefit from these proven approaches while requiring adaptation for non-deterministic behavior, context-dependent failures, and adaptive characteristics.

The seven-step incident response process (detect, assess, stabilize, report, investigate, correct, verify) provides a complete cycle for responding to AI incidents. A distinguishing feature of this framework is its integration of response actions with Preparedness Recommendations. Organizations should make investments in infrastructure, training, procedures, and capabilities before deploying AI systems, not after incidents occur. This preparedness-focused approach moves organizations from reactive crisis management to systematic reliability improvement.

This framework complements existing AI incident and governance frameworks. It provides operational detail for implementing incident response capabilities, these standards require, while addressing AI-specific challenges, including non-deterministic behavior, context-dependent failures, and system-of-systems interactions.

6.2 The Ecosystem Requirement

Organizations can implement effective AI incident response independently using the seven-step framework. However, coordinated action across developers, deployers, users, oversight bodies, independent evaluators, and assurance organizations multiplies these benefits. Each stakeholder brings distinct capabilities that, when combined, enable comprehensive ecosystem-wide improvement. Coordination enables pattern recognition across incidents, shared learning about failure modes, and systematic reliability improvement across the field, capabilities that are difficult for individual organizations to achieve alone.

Building this ecosystem requires infrastructure that does not yet exist in mature form: standardized reporting structures enabling computational analysis, information sharing mechanisms balancing transparency with proprietary protection, and clear delineation of roles

and responsibilities. Financial crime enforcement demonstrates that such infrastructure can be built and provides proven models for cross-organizational incident collection while maintaining confidentiality.

6.3 The Transition to Systematic Response

Most organizations currently handle AI incidents through improvisation, developing responses during active crises rather than executing predetermined procedures. This reactive approach limits the ability to build institutional knowledge, learn systematically from failures, or track reliability improvement over time.

Implementing the framework presented here enables organizations to respond through established processes: detection mechanisms identify problems early, pre-planned procedures enable rapid harm containment, standardized investigation methodologies reveal root causes, documented correction approaches address underlying issues, and verification processes confirm whether improvements work. Incidents become opportunities for learning and reliability enhancement rather than isolated crises requiring improvisation.

The transition requires investment in preparedness infrastructure before deploying AI systems, personnel with appropriate multidisciplinary expertise, coordination across organizational boundaries, and participation in ecosystem information sharing. Organizations that already have incident management capabilities can adapt them for AI-specific characteristics. Benefits include reduced harm from AI incidents, improved system reliability and incident reduction through continuous learning, and regulatory compliance.

6.4 Moving Forward

Organizations can use this framework to assess current capabilities, identify gaps, and prioritize improvements. Beginning with detection mechanisms, severity assessment criteria, and incident documentation procedures provides a foundation for building out the complete seven-step process. Preparedness infrastructure developed before incidents occur enables more effective response than capabilities built during active crises.

As AI systems become more capable, autonomous, and integrated into critical functions, systematic incident response becomes increasingly important for continuous reliability improvement. Organizations that develop these capabilities will be better positioned to deploy AI securely, respond to incidents effectively, and show measurable improvement over time.

Acknowledgements

My participation in a panel discussion on AI incident response, moderated by Asha Hemrajani (at the Centre of Excellence for National Security, S. Rajaratnam School of International Studies), sparked the initial conceptualization of this framework and instigating this white paper.

My professional collaboration with MM at Reins AI, focused on technical consulting for reliability and repair of agentic AI systems, has been instrumental in developing these insights. Through my direct work with the AI Incident Database, alongside colleagues Dr. Sean McGregor, Dr. Patrick Hall, Dr. Danniel Atherton, Dr. Mia Hoffman, and Ren Bin Lee Dixon, I have analyzed and documented critical incident patterns. Additional perspectives were shaped by my professional experiences at FinCEN, working with Dr. Jill McCraken, Dr. Karen Carver, and Khaled Bitar, and at the Institute for Defense Analysis, alongside Dr. Michael Burlien and Dr. Joy Brathwaite.

References

[1] Meadows, D.H. (2008) Thinking in Systems: A Primer. Chelsea Green, White River Junction

[2] Russell, S. J., & Norvig, P. (2022). Artificial Intelligence: A Modern Approach

[3] Engineering a safer world: Systems thinking applied to safety (Engineering Systems). NG Leveson. Mit Press Cambridge, 2011. 3854, 2011

[4] National Institute of Standards and Technology. Guide for conducting risk assessments. NIST Special Publication 800-30 Revision 1, U.S. Department of Commerce, 2012. URL https:/nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-30r1.pdf.

[5] Cybersecurity & Infrastructure Security Agency (CISA), Incident Response Plan (IRP) Basics

[6] "IEEE Guide for General Principles of Reliability Analysis of Nuclear Power Generating Station Safety Systems," in ANSI/IEEE Std 352-1987 , vol., no., pp.1-118, 1987, doi: 10.1109/IEEESTD.1987.101069.

[8] Mia Hoffmann and Heather Frase, "Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework" (Center for Security and Emerging Technology, July 2023), https://doi.org/10.51593/20230022.

[9] Ren Bin Lee Dixon and Heather Frase, "AI Incidents: Key Components for a Mandatory Reporting Regime," (Center for Security and Emerging Technology, January 2025), https://doi.org/10.51593/20240023

[10] OECD (2025), "Towards a common reporting framework for AI incidents", OECD Artificial Intelligence Papers, No. 34, OECD Publishing, Paris, https://doi.org/10.1787/f326d4ac-en.

[11] OECD (2024), “Defining AI incidents and related terms”, OECD Artificial Intelligence Papers, No. 16, OECD Publishing, Paris, https://doi.org/10.1787/d1a8d965-en.

[12] Heather Frase and Owen Daniels, "Understanding AI Harms: An Overview," Center for Security and Emerging Technology, August 11, 2023, https://cset.georgetown.edu/article/understanding-ai-harms-an-overview/.

[15] Marvin Rausand and Arnljot Høyland. System Reliability Theory: Models, Statistical Methods and Applications. Wiley-Interscience, Hoboken, NJ, 2004.

[16] Hubbard, D.W., & Seiersen, R. (2023). How to Measure Anything in Cybersecurity Risk (2nd Edition). Wiley.

[17] Douglas W. Hubbard, How to Measure Anything: Finding the Value of "Intangibles" in Business, is: Hubbard, D. W. (2007). How to measure anything: Finding the value of "intangibles" in business. John Wiley & Sons.

[18] UK National Cyber Security Centre (NCSC), Incident Management https://www.ncsc.gov.uk/collection/incident-management

[19] Carnegie Mellon University, Incident Management https://www.cisa.gov/sites/default/files/c3vp/crr_resources_guides/CRR_Resource_Guide-IM.pdf

[20] Blanchard, B. S., & Fabrycky, W. J. (2011). Systems Engineering and Analysis (5th ed.)

[21] Erik Hollnagel, David D. Woods, Nancy Leveson, Resilience Engineering: Concepts and Precepts. Ashgate Publishing, Ltd., 2007. ISBN 978-0-754-68136-6.

[22] James McLinn. A short history of reliability. The Journal of Reliability Information, pages 8–15, 01 2011.

[23] Ebeling, C. E. (1997). An Introduction to Reliability and Maintainability Engineering

[25] McDermott, R. E., et al. (2009). The Basics of FMEA

[26] Stamatis, D. H. (2003). Failure Mode and Effect Analysis: FMEA from Theory to Execution

[27] Tripathi, J., Gomes, H., Botacin, M. (2025). "Towards Explainable Drift Detection and Early Retrain in ML-Based Malware Detection Pipelines." In: Egele, M., Moonsamy, V., Gruss, D., Carminati, M. (eds) Detection of Intrusions and Malware, and Vulnerability Assessment. DIMVA 2025. Lecture Notes in Computer Science, vol 15748. Springer, Cham

[28] Leest, J., Gerostathopoulos, I., and Raibulet, C. (2023). "Expert Monitoring: Human-Centered Concept Drift Detection in Machine Learning Operations." In Proceedings of the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results

[29] Senarath, Y., Mukhopadhyay, A., Vazirizade, S.M., Purohit, H., Nannapaneni, S., and Dubey, A. (2024). "Designing a Human-centered AI Tool for Proactive Incident Detection Using Crowdsourced Data Sources to Support Emergency Response." Digital Government: Research and Practice, Vol. 5, No. 1

[30] Y. Senarath, A. Mukhopadhyay, S. M. Vazirizade, H. Purohit, S. Nannapaneni and A. Dubey, "Practitioner-Centric Approach for Early Incident Detection Using Crowdsourced Data for Emergency Services," 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 2021, pp. 1318-1323, doi: 10.1109/ICDM51629.2021.00164.

[31] United States of America Department of Defense. Department of defense standard practice, system safety (mil-std-882e). Department of Defence, 2012. https://safety.army.mil/Portals/0/Documents/ON-DUTY/SYSTEMSAFETY/Standard/MIL-STD-882E-change-1.pdf

[32] H. J. Caldera and S. C. Wirasinghe. A universal severity classification for natural disasters. Natural Hazards, 111:1533–1573, 2021. doi: 10.1007/s11069-021-05106-9

[33] https://nij.ojp.gov/sites/g/files/xyckuh171/files/media/document/draft-failure-definitions-and-scoring-criteria.docx

[35] Boston, M. F., Frase, H., & Georgala, E. (2025). Reliability and Repair for Agentic Systems. Reins AI Technical White Paper v1.0. October 2025. Retrieved from www.reinsai.com/articles/reliability-and-repair-for-agentic-systems

[36] Hanmer, R.S. (2007). Patterns for Fault Tolerant Software. Wiley Software Patterns Series. John Wiley & Sons.

[37] Amazon Web Services. (2022). "REL05-BP01 Implement graceful degradation to transform applicable hard dependencies into soft dependencies." AWS Well-Architected Framework - Reliability Pillar.

[38] Edwards, Tamsyn & Lee, Paul. (2017). Towards Designing Graceful Degradation into Trajectory Based Operations: A Human-Machine System Integration Approach. 10.2514/6.2017-4487.

[39] Chipangila, B., Liswaniso, E., Mawila, A. et al. (2024). "Controlled vocabularies in digital libraries: challenges and solutions for increased discoverability of digital objects." International Journal on Digital Libraries, Vol. 25, pp. 139–155

[40] Akbari Gurabi, M., Nitz, L., Bregar, A., Popanda, J., Siemers, C., Matzutt, R., & Mandal, A. (2024). Requirements for Playbook-Assisted Cyber Incident Response, Reporting and Automation. Digital Threats: Research and Practice, 5(3), 1–11.

[41] Stamatis, D.H. (2003). Failure Mode and Effect Analysis: FMEA from Theory to Execution (2nd Edition). ASQ Quality Press.

[42] Leveson, Nancy & Daouk, Mirna & Dulac, Nicolas & Marais, Karen. (2003). Applying STAMP in accident analysis. Workshop Investigation Reporting Incidents Accidents (IRIA).

[43] Nancy G. Leveson, Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press, 2011. ISBN 978-0-262-01662-9.

[44] Vesely, W. E., et al. (2002). Fault Tree Handbook with Aerospace Applications Content

[45] Vesely, W. & Goldberg, F. & Roberts, N. & Haasl, D.. (1981). Fault Tree Handbook. 216.

[46] NASA Software Engineering Handbook, Section 8.07 - Software Fault Tree Analysis, Created by Haigh, Fred, last modified on Jun 30, 2023.

[47] Garcia-Martin, R., et al. (2024). "The impact of AI errors in a human-in-the-loop process." Cognitive Research: Principles and Implications, 9(1).

[48] Parasuraman, R., & Manzey, D.H. (2010). "Complacency and Bias in Human Use of Automation: An Attentional Integration." Human Factors, 52(3), 381-410.

[49] Singh, A., et al. (2025). "Exploring automation bias in human–AI collaboration: a review and implications for explainable AI." AI & Society.

[50] Dekker, S. (2012). Just Culture: Balancing Safety and Accountability (2nd Edition). CRC Press.

[51] International Organization for Standardization. (2015). ISO 9001:2015 Quality Management Systems - Requirements. ISO.

[52] Wilson, P.F., Dell, L.D., & Anderson, G.F. (2023). Root Cause Analysis: A Tool for Total Quality Management (3rd Edition). ASQ Quality Press.

[53] Garvin, D.A. (1993). "Building a Learning Organization." Harvard Business Review, 71(4), 78-91.

[54] Montgomery, D.C. (2019). Introduction to Statistical Quality Control (8th Edition). Wiley.

[55] Kim, G., Humble, J., Debois, P., Willis, J., & Forsgren, N. (2021). The DevOps Handbook (2nd Edition). IT Revolution Press.

[56] O'Connor, P., & Kleyner, A. (2012). Practical Reliability Engineering (5th Edition). Wiley.

[57] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.

[58] National Institute of Standards and Technology. (2018). Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1. NIST.

[59] International Organization for Standardization. (2023). ISO/IEC 23894:2023 - Artificial intelligence — Guidance on risk management. ISO.

[60] International Organization for Standardization. (2023). ISO/IEC 42001:2023 - Artificial intelligence — Management system. ISO.

[61] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

[62] European Commission Draft Guidance (September 26, 2025): Article 73

[63] IAPP Timeline Resource: High-risk AI incident reporting by August 2, 2025 [refers to preparatory obligations]; practical implementation of high-risk AI requirements by February 2, 2026; entry into application August 2, 2026, https://iapp.org/resources/article/eu-ai-act-timeline/

[64] States Ring in the New Year with Proposed AI Legislation by: Adam S. Forman, Greta Ravitsky, Elizabeth S. Torkelsen, Jennifer Stefanick Barna Epstein Becker & Green, P.C. - Workforce Bulletin Tuesday, January 21, 2025 https://natlawreview.com/article/states-ring-new-year-proposed-ai-legislation

[65] EU AI Act Article 73 on serious incident reporting, https://artificialintelligenceact.eu/article/73/

[68] McGregor, S. (2021). "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database." Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 15458-15463.

[76] U.S. Department of Homeland Security. (2015). "Information Sharing and Analysis Centers: Best Practices Guide." DHS National Protection and Programs Directorate.

[77] Raji, I.D., et al. (2020). "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44.

[78] Lam, M.K., et al. (2024). "A Framework for Assurance Audits of Algorithmic Systems." Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 1124-1137.

[79] Ilan Strauss, Isobel Moure, Tim O'Reilly and Sruly Rosenblat. "Real-World Gaps in AI Governance Research: AI Safety and Reliability in Everyday Deployments." SSRC AI Disclosures Project Working Paper Series (SSRC AI WP 2025-04), Social Science Research Council, April 2025. https://www.ssrc.org/publications/real-world-gaps-in-ai-governance-research/

[80] Strauss, Ilan, O'Reilly, Tim, Rosenblat, Sruly, and Isobel Moure. "Governing AI Through SEC Disclosure." Working Paper. Social Science Research Council, October 2025. https://www.ssrc.org/publications/governing-ai-through-sec-disclosure-materiality-standards-and-incident-reportinglessons-from-cybersecurity/