

AI Incident-Response Field Guide

A print-ready command aid for Property & Casualty insurers

Use this guide to activate a coordinated response, determine severity, contain harmful or uncontrolled behavior, preserve AI-specific evidence, protect consumers, and control recovery.

ACTIVATE WHEN	FIRST PRIORITY
An AI model, generative system, or agent may have caused or amplified error, unauthorized action, unfair outcome, security or privacy exposure, noncompliance, operational disruption, or a near miss.	Protect consumers and essential service. Stop uncontrolled action, preserve evidence, establish command, and record decisions. Do not wait for complete certainty before safe containment.

Incident activation record

Field	Record
Incident ID / opened	
System, version, environment	
Business process / jurisdictions	
Observed event and source	
Potentially affected population	
Provisional severity	
Incident commander / next review	

Version 1.0 | September 2026

First 30 Minutes

Record facts, assumptions, unknowns, owners, and decision times in one authoritative incident record.

Done	Action	Owner / time
[]	Protect consumers; stop harmful or uncontrolled actions.	
[]	Preserve the alert, complaint, output, transaction, prompt, context, logs, and version state.	
[]	Route consequential work to qualified people or a tested safe fallback.	
[]	Name the incident commander and business, technical, legal/compliance, privacy/security, and communications leads.	
[]	Set severity, secure channels, briefing cadence, and next decision time.	
[]	Identify notification clocks without delaying containment.	

Immediate command decisions

Decision	Record
Safe containment	
Essential-service fallback	
Executive escalation	
Legal / regulatory review	
Evidence preservation owner	
Affected-population owner	
Next command briefing	

Severity Matrix

Use the highest credible impact. Reassess when scope, evidence, autonomy, reversibility, affected population, jurisdiction, or obligations change.

Level	Target	Definition	Signals
SEV-1 Critical	Activate immediately	Actual or imminent widespread consumer harm, unlawful or unsafe automated action, material financial exposure, active data compromise, loss of control over an agent, or a regulator-notifiable event requiring urgent executive direction.	Systemic wrongful denials; discriminatory eligibility action at scale; autonomous payments outside authority; exposed protected information; inability to stop a production agent.
SEV-2 High	Activate within 1 hour	Material but containable impact, credible adverse consumer outcome, significant model or workflow failure, repeated unauthorized action, or an incident likely to require legal, compliance, regulator, or executive involvement.	Incorrect settlement recommendations in production; material drift; repeated fabricated claim communications; vendor model failure affecting a book of business.
SEV-3 Moderate	Activate same business day	Limited impact with effective containment, no known material rights impact, and a credible path to correction through established operational controls.	Localized classification errors; incorrect internal summaries caught before use; isolated access or configuration issue with no confirmed adverse decision.
SEV-4 Low / Near miss	Log within 1 business day	No confirmed harmful outcome; issue was prevented, detected in testing, or remained within low-impact internal activity. Track for patterns and control improvement.	Blocked prompt-injection attempt; pre-production evaluation failure; harmless output defect; alert that proves non-actionable after review.

Impact assessment

Dimension	Question	Rating / evidence
Consumer or rights impact	Did the event affect eligibility, price, coverage, benefits, payment, investigation, appeal, communication, privacy, accessibility, or fair treatment?	
Scale and propagation	How many consumers, claims, policies, transactions, users, jurisdictions, or downstream systems are affected or plausibly exposed?	
Autonomy and control	Did a model or agent act, communicate, write, transact, or call tools beyond approved authority, and can it be stopped reliably?	
Legal or regulatory exposure	Could the event violate insurance, privacy, cybersecurity, consumer-protection, recordkeeping, discrimination, or market-conduct obligations?	
Data and security	Was protected, proprietary, privileged, or regulated data exposed, altered, inferred, exfiltrated, or used outside approved purpose?	
Financial and operational impact	What payments, reserves, premiums, expenses, deadlines, service levels, or business processes are affected?	
Detectability and reversibility	Can affected records and decisions be identified, explained, corrected, reprocessed, recalled, or reversed before further harm?	
Evidence integrity	Are prompts, outputs, model versions, configurations, tool calls, logs, approvals, and source data available and trustworthy?	

Response Lifecycle

Phases overlap. Preserve one chronology and return every material decision to the incident command record.

Phase	Objective and required actions	Lead / status
P1 Detect and intake	Objective: Create one authoritative incident record and protect the signal from being lost or informally handled. - Record who detected the issue, when, where, and how. - Capture the system, use case, version, environment, workflow, and affected business process. - Preserve the original alert, complaint, output, transaction, or observation. - Assign a provisional severity and response owner. - Open a decision and evidence log.	Incident coordinator ☐ Started ☐ Complete
P2 Triage and activate	Objective: Confirm scope, establish command, and mobilize the right insurance, technical, legal, and control functions. - Validate that the event is real and identify credible worst-case impact. - Apply the severity matrix and document assumptions and unknowns. - Name the incident commander, technical lead, business lead, legal/compliance lead, and communications lead. - Set meeting cadence, secure channels, and executive escalation. - Identify notification clocks without delaying containment.	Incident commander ☐ Started ☐ Complete
P3 Contain and preserve	Objective: Stop additional harm while preserving evidence and essential service. - Disable or narrow affected actions, credentials, integrations, models, prompts, routes, or data access. - Route consequential decisions to qualified people or a safe fallback. - Quarantine affected versions and prevent silent redeployment. - Preserve logs, prompts, outputs, tool traces, data snapshots, approvals, and configuration state. - Document every containment decision, owner, time, and tradeoff.	Technical and business leads ☐ Started ☐ Complete
P4 Assess impact and cause	Objective: Determine what happened, why controls failed, who or what was affected, and what must be corrected. - Reconstruct the timeline from source evidence. - Identify affected consumers, decisions, records, transactions, and jurisdictions. - Test for differential impact, error patterns, security exploitation, and downstream propagation. - Separate initiating cause, contributing conditions, failed controls, and detection gaps. - Maintain legal privilege and investigation independence where directed.	Investigation lead ☐ Started ☐ Complete
P5 Notify and remediate	Objective: Meet obligations and restore affected people, records, and decisions fairly and consistently. - Decide regulator, consumer, partner, board, insurer, reinsurer, vendor, and law-enforcement notifications. - Correct or re-adjudicate affected decisions using qualified human review. - Issue accurate consumer communications with appeal and contact paths. - Refund, repay, restore, suppress, or correct records as appropriate. - Track remediation completion and exceptions at the individual and aggregate level.	Legal/compliance and business ☐ Started ☐ Complete
P6 Recover and validate	Objective: Return to service only after evidence shows the risk is controlled and the operating design is supportable. - Define recovery criteria and obtain documented approval. - Correct the model, prompt, data, permissions, workflow, or vendor dependency. - Run regression, security, fairness, performance, and scenario tests proportionate to impact. - Use staged deployment, enhanced monitoring, rollback, and manual fallback. - Confirm no unresolved affected population or downstream record remains.	System owner and independent control ☐ Started ☐ Complete
P7 Close and learn	Objective: Close only when remediation, evidence, ownership, and systemic improvement are complete. - Approve the final incident report, impact statement, and root-cause analysis. - Assign corrective actions with owners, dates, and validation evidence. - Update inventory, risk tier, monitoring, tests, procedures, contracts, and training. - Share lessons across similar systems and vendors. - Track recurrence indicators and report material themes to governance and the board.	Accountable executive ☐ Started ☐ Complete

Containment and AI Evidence

Preserve before destructive troubleshooting when feasible. Technical containment must be enforced through permissions, credentials, routes, configurations, queues, and integrations - not only through an instruction to the model.

Containment checklist

Done	Control	Verification / owner
☐	Disable or narrow affected action, model route, prompt, retrieval source, integration, or data access.	
☐	Revoke agent credentials and stop queued, scheduled, retried, and downstream actions.	
☐	Activate qualified-human or approved manual fallback.	
☐	Quarantine affected versions and prevent silent redeployment.	
☐	Verify containment stopped the behavior and document service tradeoffs.	

Evidence checklist

Family	Preserve
Identity / versions	Model, agent, prompt, code, rules, configuration, environment, vendor, release, and timestamps.
Inputs / outputs	Prompts, attachments, retrieved context, outputs, scores, citations, edits, and sessions.
Actions / tools	Tool calls, API activity, credentials, permissions, writes, messages, payments, retries, downstream events.
Data	Sources, features, labels, corpus, transformations, lineage, access, quality, and affected records.
Human control	Evidence shown, approvals, overrides, rationale, queue, workload, stop and escalation actions.
Assurance / change	Monitoring, alerts, tests, drift, tickets, approvals, prior incidents, contracts, and vendor notices.

Impact, Cause, and Remediation

A technical fix is incomplete if affected decisions, consumers, records, payments, communications, deadlines, or rights remain uncorrected.

Investigation record

Question	Evidence / conclusion
What happened, when, and through which workflow?	
Which consumers, claims, policies, transactions, jurisdictions, and downstream systems are affected?	
What direct, differential, financial, operational, security, privacy, and rights impact occurred?	
What initiated the event?	
Which conditions and failed or absent controls allowed it to matter?	
Why was it detected when it was, and which earlier signals existed?	
What confidence and limitations apply to the scope and conclusions?	

Consumer remediation

Population / record	Correction	Owner / due	Status / evidence

Notification and Communications

Apply current law, contract, company policy, privilege, and jurisdiction-specific requirements. Document both notification and non-notification decisions.

Audience	Assessment question	Owner / deadline	Decision / basis
Insurance regulator	Does the event implicate reporting, examination cooperation, market-conduct, unfair-practice, AI-governance, or data obligations?		
Consumer / claimant	Is notice needed to explain, correct, repay, re-adjudicate, protect rights, or provide appeal?		
Privacy / cyber authority	Did regulated data or a covered cybersecurity event occur?		
Board / executive	Is impact material, systemic, strategic, reputation-sensitive, or outside risk appetite?		
Vendor / partner / reinsurer	Do contracts, shared service, claims, recovery, or evidence rights require action?		
Law enforcement / other	Is fraud, criminal activity, public safety, or another reporting channel implicated?		

Communication approval record

Message / audience	Facts and uncertainty	Approvals	Release / follow-up

Recovery and Restart Gate

Recovery is an accountable risk decision. Do not restart only because a patch is available.

Done	Required evidence	Approver / conditions
☐	Initiating cause and material contributing conditions are understood.	
☐	Affected populations and downstream records are identified or conservatively bounded.	
☐	Corrective changes passed reproduction, regression, boundary, security, fairness, and scenario tests proportionate to impact.	
☐	Permissions, authority, data, human oversight, monitoring, escalation, and fallback match the approved design.	
☐	Staged deployment includes enhanced monitoring, stop criteria, rollback, manual fallback, and accountable owners.	
☐	Consumer remediation and notification work can continue without being obscured by restart.	
☐	Required business, technical, independent-control, legal/compliance, privacy/security, and executive approvals are complete.	

Restart decision

Element	Record
Approved scope	
Residual risk / conditions	
Approval time	
Monitoring window	
Stop criteria	
Next review	

Closure, Corrective Action, and Learning

Closure requires more than restored service. Confirm remediation, evidence, notification decisions, ownership, verification, and recurrence monitoring.

Closure criteria

Done	Criterion	Evidence / owner
☐	Final severity, timeline, impact, affected population, and root cause are approved.	
☐	Consumer, financial, record, and operational remediation is complete or governed through an approved exception.	
☐	Notification and communication decisions are complete and retained.	
☐	Corrective actions have owners, dates, validation methods, and governance visibility.	
☐	Related systems, agents, vendors, data, and workflows were assessed for the same failure.	
☐	Inventory, risk tier, tests, monitoring, controls, contracts, procedures, and training are updated.	
☐	Recurrence indicators and reporting duration are defined.	

Corrective action tracker

Action	Owner / due	Validation evidence	Status

Insurance Incident Scenarios

Use these as exercise starters; adapt to the actual system, authority, population, jurisdiction, and insurer obligations.

Domain	Incident	Immediate response pattern
Claims	Hallucinated coverage explanation sent to claimants	Stop outbound automation; preserve prompt/output/tool traces; identify recipients; qualified adjuster review; correct communications; assess unfair claims and complaint exposure.
Claims	Payment agent exceeds settlement authority	Disable credentials and payment route; reconcile transactions; recover or correct payments; inspect authority boundaries and approval controls.
Underwriting	Eligibility model produces unexplained adverse pattern	Suspend affected use; preserve population/version/data; test disparity and accuracy; re-underwrite affected applicants; legal and regulator assessment.
Pricing	Deployment uses wrong rating factor or version	Stop quotes/bindings as needed; identify policies and jurisdictions; calculate premium impact; correct/re-rate/refund; filing and regulator review.
Fraud	Prompt injection alters referral-agent behavior	Isolate input path and credentials; preserve adversarial content; inspect tool calls; review referred or suppressed claims; security testing before restart.
Servicing	Assistant exposes policyholder information	Stop channel; preserve session/access logs; privacy and security assessment; identify data and recipients; notification decision and access-control correction.
Vendor	Third-party model changes without notice	Freeze or revert version; invoke contract and evidence rights; validate performance and affected decisions; reassess vendor and change controls.

Exercise injects

Add incomplete logs, a vendor delay, a vulnerable claimant population, a cross-jurisdiction notification question, a second downstream action, executive pressure to restore service, and evidence that the first containment did not stop queued actions.

Official foundations

NIST SP 800-61 Rev. 3 - Incident Response Recommendations and Considerations for Cybersecurity Risk Management

NIST AI Risk Management Framework

NIST AI 600-1 - Generative AI Profile

NAIC Model Bulletin on the Use of Artificial Intelligence Systems by Insurers

Confirm current legal, regulatory, contractual, and company requirements. This field guide supports response discipline and does not replace legal advice or established emergency procedures.