ODA3 INSTITUTE

ADOPTION GUIDE

When AI Breaks

The ODA3 Guide to AI Incident Response

Doc ID: ODA3-2026-08-WHP-COM-011
Classification: Public – GEL v1.0 (L1 Public Use)

Revision: FINAL R2 (corrected)

Executive Summary

AI incidents can be difficult to recognise because a detection signal is not automatically an incident, an incorrect output is not automatically malicious, and response may need to begin before causality, severity or classification is settled. This guide provides a short operational orientation for teams that need to recognise potentially material AI events, preserve evidence and uncertainty, determine what is known, and begin proportionate response.

UAIF™ structures incident identity, causality, harm, severity, confidence and the evidence/uncertainty record. AI-IRF™ structures preparation and response. GAISSF™ provides governance and control context. Where physical or cyber-physical capabilities and consequences place physical-AI security requirements within scope, PAI-SF™ provides the authoritative detailed physical-AI security requirements and associated containment, evidence-preservation and recovery considerations. Where PAI-SF™ applies, it supersedes and expands upon GAISSF™ Domain D9 (Physical AI Safety), which is retained in the current GAISSF™ release for continuity. These functions interact: protective action can begin while incident determination or classification remains provisional.

METHODOLOGY NOTE

This guide is introductory and operationally illustrative. ODA3 Institute was founded in March 2026. The guide contains no client outcome data, proprietary telemetry, incident-frequency statistics, or implied deployment history. Examples are hypothetical and are used to explain incident-response reasoning rather than to establish universal thresholds, legal conclusions, severity levels or regulatory duties.

BOUNDARY

This guide does not replace an organisation's existing SOC/CSIRT process. It is intended to fit around and enrich existing incident-management practices with AI-specific evidence, uncertainty, classification and response considerations. It does not require a new SIEM, SOAR, case-management platform or proprietary ODA3 tooling.

1. Event, Detection and Incident Are Different Questions

For purposes of this introductory guide, an event is treated as an observable occurrence or signal brought to attention for evaluation. Authoritative UAIF™ terminology and classification criteria govern the structured event/incident record. A detection is the mechanism or record that brought the occurrence to review. Whether the event should be treated as an AI incident depends on the available facts, potential or actual consequence, organisational policy, applicable obligations and the need for governed response. This guide does not define a universal materiality threshold.

KEY PRINCIPLE

Detection is not classification. Preserve uncertainty: record what is confirmed, what is inferred, what remains unknown, and what evidence is missing before converting a signal into a conclusion.

Questions to ask immediately

- What signal triggered review, from which source, and at what time?
- Which system, model, agent, tool, data source, provider or physical component may be affected?
- What is confirmed by evidence right now?
- What are working hypotheses rather than established facts?
- What potential consequence requires protective action even if incident determination is incomplete?
- What evidence could be lost, changed or made inaccessible by the next response action?

2. The ODA3 Incident Response Orientation

Orientation step	Operational question	Illustrative record
Recognise	What signal or behaviour requires review?	Alert, detection source, time, affected system

Orientation step	Operational question	Illustrative record
Preserve	What evidence or state could disappear or change?	Trace/log hold, versions, snapshot, custody record
Determine	What is confirmed, inferred, unknown or not established?	Local triage/case record; converts to a UAIF™-structured event/incident record once classification criteria are met
Classify	What identity, causality, harm, severity and confidence are supportable?	Rationale, confidence, missing evidence
Respond	What authority, tool, identity, model or workflow must be constrained?	Decision, approval, containment, verification
Recover	What state is safe to restore and what validation is required?	Recovery criteria, tests, approval
Notify	What contractual, legal, regulatory or stakeholder duties may apply?	Decision basis, content, timing, recipients
Learn	What changed and how will recurrence/effectiveness be reviewed?	Corrective action, retest, recurrence review

Ecosystem Relationship at a Glance

Operational question	Primary orientation
What happened, and what is known?	UAIF™ structured event/incident record
What is causal fact vs working hypothesis?	UAIF™ causality and evidence discipline
How severe might this be, and how confident are we?	UAIF™ severity indicators, confidence and missing evidence
What must we constrain or preserve now?	AI-IRF™ response informed by available facts
What state is safe to restore?	AI-IRF™ recovery and validation
What changed in our control baseline?	GAISSF™ governance/control feedback
Are physical effects or recovery involved?	PAI-SF™ physical containment, evidence and return-to-service considerations

RELATIONSHIP TO AI-IRF™

Recognise, Preserve, Determine and Classify are operational elaborations used by this guide to make early response work visible. They are not new AI-IRF™ lifecycle stages. AI-IRF™'s governed lifecycle remains preparation, detection/analysis, containment, recovery, notification where applicable, learning and post-incident validation.

Integration with an Existing SOC/CSIRT

- Use the organisation's existing incident identifier, severity process, case-management platform and escalation route wherever practical.
- Add AI-specific fields rather than creating a parallel case unless separation is operationally necessary.
- Existing SIEM/SOAR, cloud logs, observability, version control, ticketing and evidence repositories can supply or preserve relevant evidence; no particular technology is mandated.
- Early triage information remains in the organisation's local SOC/case-management workflow as an observation or event-under-triage; it is not treated as a formal UAIF™ record_type until UAIF™'s own classification criteria for a structured event/incident record are met.

- UAIF™ can structure the AI-specific event/incident record while existing SOC/CSIRT governance remains in force.
- AI-IRF™ response actions should be integrated with existing containment, recovery, notification and crisis-management authorities.

3. Incident Declaration, Severity and Escalation

A short guide should not invent a universal declaration threshold or severity scale. Organisations should retain their existing declaration and severity governance, but ensure AI-specific factors are considered.

Declaration factors to consider

- confirmed or plausibly consequential unauthorised action, access or tool use
- potentially material harm to customers, patients, employees, the public, critical operations or physical assets
- potential compromise of model, retrieval source, agent, provider dependency or control boundary
- physical effect, safety concern or loss of safe-state confidence
- potential legal, contractual, regulatory or insurance notification duty
- uncertainty significant enough that governed incident ownership and evidence preservation are warranted

Severity indicators to consider

- scope: affected systems, users, data subjects, assets or dependent services
- consequence: confirmed and potential financial, operational, safety, rights, privacy or service impact
- reversibility: ease or difficulty of correcting outputs, actions, data exposure or physical effects, including whether onward propagation has occurred
- progression: whether impact is continuing, contained, cascading or dependent on a third party
- criticality: importance of the affected service, system or physical function
- confidence: strength and completeness of evidence supporting the current assessment

> **CONFIDENCE DISCIPLINE**
> Use the confidence terminology defined by the authoritative UAIF™ instrument. This guide does not create a new confidence scale. Record the evidence basis, missing evidence and uncertainty that support the current confidence statement.

Escalation triggers to predefine

- potential physical safety impact or loss of safe-state confidence
- potentially material customer, patient, employee, public or critical-service consequence
- possible notification or reporting duty
- containment action beyond the current responder's authority
- multi-system or cascading impact
- third-party/provider dependency that limits evidence access or containment
- material change in classification, consequence or confidence
- unresolved uncertainty that affects a time-critical response decision

> **NOTABLY ABSENT**
> This guide does not prescribe Low/Medium/High severity levels, a declaration rule, a mandatory escalation cadence, or a fixed executive-update interval. Those thresholds and authorities belong in the organisation's own governed incident-response plan.

4. Roles and Decision Authority

AI incident response is cross-functional. The functions below may be separate roles in a large enterprise or combined in a small team. What matters is that authority, accountability and handoffs are defined before an incident.

Function	Typical responsibility
Incident lead	Coordinates response, priorities, decisions, escalation and status
Technical analysis	Preserves technical evidence, analyses system behaviour and validates containment
System owner	Provides architecture/context and implements approved containment/recovery actions
Evidence custodian	Tracks preservation, integrity, access and chain-of-custody where required; this function may sit outside the SOC and may require pre-authorised log/evidence holds
Governance / risk	Connects incident facts to control baselines, risk ownership and assurance consequences
Legal / compliance	Determines applicable legal, regulatory, contractual and privilege considerations
Communications	Coordinates authorised internal/external messaging without speculation beyond confirmed facts
Safety / OT / clinical	Leads domain-specific safety actions where physical or clinical consequence is possible
Provider liaison	Requests provider evidence, notices, support and change information where dependencies exist

SMALL-TEAM ADAPTATION

Small teams may combine functions, but should still record who owns incident coordination, technical analysis, evidence preservation, legal/compliance consultation where needed, and return-to-service authority. Resource constraint does not make missing evidence or unassigned risk disappear.

5. Preserve Evidence Before Response Destroys It

Restarting a service, rotating credentials, changing prompts or policies, disabling an agent, updating a model, wiping a device or moving a physical system can alter or destroy evidence. Preservation should therefore be considered before or alongside containment, while safety and operational continuity remain paramount.

- volatile traces and telemetry where available, relevant and proportionate
- clock, timestamp and identifier consistency across relevant systems
- model, prompt, retrieval, tool, policy and configuration versions
- log-retention holds and integrity verification where applicable
- third-party notices, support records and evidence requests
- restricted, privileged, personal or sensitive materials with appropriate access controls
- documented reasons where relevant evidence cannot lawfully, safely or technically be retained

KEY PRINCIPLE

Where preservation of evidence conflicts with an immediate action necessary to protect human safety or prevent material physical harm, the required safety action takes priority. Any resulting evidence loss or alteration should be documented where practicable.

PRIVACY AND PRIVILEGE CONSIDERATIONS

Preserved prompts, outputs and retrieval context frequently contain personal data, patient information or attorney-client privileged material. Apply the organisation's existing data-masking, access-control and legal-

hold practices when preserving this evidence, and route material that may be privileged through legal/compliance before broader internal sharing. This guide does not treat unfiltered duplication of prompts or context windows into general-purpose ticketing or case-management systems as a safe default preservation method.

Minimum evidence preparedness

To make preservation possible, the AI-IRF™ preparation phase should determine, for each system or risk tier, what versions, traces, configuration records, provider records and other evidence should be available; how long they should be retained; and who may access or place holds on them. These decisions should reflect architecture, investigative value, privacy, security, contractual, sector and legal constraints.

Chain of custody where needed

Where evidence may support regulatory, legal, insurance, disciplinary or high-assurance review, record who collected or accessed it, when, from where, what was done to it, and how integrity was checked. Not every internal AI event requires formal forensic custody; apply the level of control appropriate to the case and applicable obligations.

NOTABLY ABSENT

This guide does not require universal prompt logging, universal output logging, indefinite telemetry retention or a mandatory forensic toolset. Logging and preservation decisions must account for architecture, investigative value, privacy, security, contractual, sector and legal constraints.

6. Prompt Injection / Agent Tool Abuse

An AI assistant receives an instruction designed to override intended behaviour and attempts to invoke a tool outside the expected workflow.

Immediate orientation

- Preserve the triggering input, relevant system/policy versions, tool-call request and authorisation decision.
- Do not assume malicious success merely because an injection attempt was detected.
- Constrain affected tools, identities, sessions or workflows where proportionate.
- Where credentials are implicated, distinguish exposure or theft from in-memory/authorised-session abuse before deciding whether rotation, session revocation or another containment action is proportionate.
- Determine whether unauthorised action executed, partially executed, failed or was blocked, and what remains unknown.

NOTABLY ABSENT

Does not establish that every prompt-injection attempt is an incident or that tool invocation proves compromise.

7. Sensitive-Data Disclosure

An AI system produces output that appears to contain sensitive or restricted information.

Immediate orientation

- Preserve the output and relevant retrieval/context evidence without unnecessarily replicating sensitive content.
- Determine whether the information is authentic, synthetic, previously public, misclassified or actually disclosed from a protected source.
- Assess affected data, recipients, onward exposure and relevant access paths.
- Contain retrieval, sharing or access paths where necessary while preserving investigative evidence.

NOTABLY ABSENT

Does not determine breach status, privacy-law notification duties or the sensitivity classification of the data.

8. RAG / Knowledge-Base Poisoning

A retrieval-augmented system begins surfacing misleading or manipulated material after a source, connector or index change.

Immediate orientation

- Preserve source versions, ingestion/change history, retrieval results and index/configuration state.
- Treat source compromise, indirect prompt injection, retrieval failure, model behaviour and ordinary content error as competing hypotheses until evidence supports a conclusion.
- Constrain suspect sources or connectors where proportionate.
- Review downstream decisions or outputs only to the extent supported by evidence.

NOTABLY ABSENT

Does not establish intentional poisoning, exploit success or downstream harm merely from incorrect retrieval.

9. Model or Agent Behaviour Change

A deployed model or agent begins behaving differently following a provider update, configuration change, prompt revision or tool modification.

Immediate orientation

- Preserve before/after versions and the change timeline.
- Distinguish expected change, drift, configuration error, provider-side change and adversarial manipulation.
- Identify affected workflows and whether rollback or capability restriction is warranted.
- Record what can and cannot be independently verified about third-party changes.

NOTABLY ABSENT

Does not treat every behavioural change as a security incident or provider fault.

10. AI Supply-Chain / Provider Incident

An external model, library, agent component, dataset or hosted AI service is suspected of introducing harmful behaviour, integrity failure or unavailable evidence.

Immediate orientation

- Preserve provider notices, versions, identifiers/hashes where available, dependency records and internal observations.
- Request relevant provider evidence through established contractual/support channels and preserve the request/response trail.
- Separate provider representations from independently verified evidence and document visibility limitations.
- Determine blast radius across all systems using the affected dependency; consider temporary restriction, substitution or rollback based on consequence and evidence.
- If the provider cannot support necessary evidence access, version identification or pinning, rollback, or another containment dependency, containment may require constraining or disabling the dependent capability. Record the limitation as a supply-chain finding for procurement and GAISSF™ control review.

NOTABLY ABSENT

Does not establish provider culpability or contractual breach, and does not define a universal minimum vendor attestation.

11. Harmful or Materially Incorrect Output

An AI output may have influenced a consequential business, customer, clinical, employment or public-sector decision.

Immediate orientation

- Preserve the output, context, human review, downstream decision and relevant system version.
- Determine whether the AI output was relied upon, overridden, ignored or independently corroborated.
- Do not equate incorrectness with maliciousness or incident severity.
- Consider whether bounded review of similar outputs is warranted based on evidence rather than assumption.

NOTABLY ABSENT

Does not determine legal harm, discrimination, professional negligence or regulatory breach.

12. Unauthorised Autonomous Action

An agent or autonomous workflow performs, or attempts to perform, an action outside its intended authority.

Immediate orientation

- Preserve action requests, authorisation decisions, tool/API records, identity context and resulting state changes.
- Constrain the relevant authority path where immediate recurrence is plausible.
- Determine whether the action executed, partially executed, failed or was blocked.
- Review delegated authority, approval boundaries, blast radius and recovery requirements.

NOTABLY ABSENT

Does not establish that every blocked unauthorised attempt is a material incident.

13. Physical-AI Incident

An AI-enabled device, robot, vehicle or cyber-physical system behaves unexpectedly and may create physical consequences.

Immediate orientation

- Prioritise human safety and safe-state actions where required; cyber containment alone may be insufficient.
- Do not apply default IT-style containment actions — such as network isolation or power-cycling — to a cyber-physical system without safety-engineering authorisation. Severing network or telemetry links intended to isolate a compromised component can itself disable safety-critical monitoring or trigger uncontrolled physical momentum; the safety-engineering function should confirm the containment action itself will not create a hazard before it is taken.
- Preserve perishable physical evidence such as sensor state, actuator position, environmental conditions and relevant telemetry where safe and feasible.
- Record human override, kinetic-command interruption, physical recovery and functional-safety interfaces where relevant.

- Before return to service, determine validation tied to the identified failure or root-cause hypothesis, which may include physical inspection, safe-state verification or hardware-in-the-loop testing.
- GAISSF™ Domain D9 is retained in the current release for continuity; where PAI-SF™ applies to the system, it supersedes and expands upon D9 as the authoritative source for detailed physical-AI security requirements.

> **NOTABLY ABSENT**
>
> Does not replace emergency procedures, functional safety, medical-device, machinery, vehicle, aviation or other sector-specific incident obligations, and does not establish that GAISSF™ Domain D9 alone satisfies PAI-SF™'s expanded physical-AI requirements where PAI-SF™ applies.

14. Multi-System / Cascading AI Incident

A shared model, provider, dataset, identity path or agent component creates potentially harmful effects across more than one AI system or dependent business service.

Immediate orientation

- Identify the common dependency and maintain a parent incident view with linked system-specific cases where useful.
- Do not assume identical impact across all dependent systems; verify scope and consequence separately.
- Prioritise containment at the shared dependency where feasible while preserving system-specific evidence.
- Track provider, business-service and regulatory implications across affected environments without collapsing them into a single unsupported conclusion.

> **NOTABLY ABSENT**
>
> Does not establish that every shared dependency failure creates the same severity or notification duty across affected systems.

15. UAIF™ and AI-IRF™: Complementary, Not Sequential

UAIF™ and AI-IRF™ address different operational needs. UAIF™ helps structure what happened, incident identity, causality, harm, severity, confidence, contextual modifiers, available evidence and missing evidence. AI-IRF™ governs preparation and response. A team may therefore begin containment, safety action or notification analysis while UAIF™ classification remains provisional. The ecosystem relationship table appears near the start of this guide so teams can use that architecture during initial response.

16. The First Incident Record: Populate What Matters First

UAIF™ contains a broader structured field set. Under time pressure, the first record should prioritise information needed to identify the case, preserve uncertainty, support immediate response and show what evidence is missing. Additional fields can be completed as analysis develops.

- affected system, components and relevant dependencies
- initiating event and detection source
- confirmed facts and working hypotheses
- suspected attack or failure mechanism, if supportable
- affected AI-system layer
- confirmed and potential impact
- causality status
- severity indicators and contextual modifiers
- confidence, available evidence and missing evidence
- physical-world capability or effect where relevant

- links to related cases, provider notices or recent changes
- provider case ID / support ticket reference where applicable

> **EVIDENCE DISCIPLINE**
>
> A contemporaneous incident record should distinguish confirmed facts from working hypotheses and unresolved questions. Where a proposition has not been established by available evidence, record that boundary rather than converting absence of evidence into a conclusion.

17. Communication During Active Response

Operational communication should support coordination without outrunning the evidence. Define who may brief leadership, affected business owners, providers, customers or external parties; maintain an authoritative incident record; and distinguish confirmed facts from provisional analysis in every status update.

- designate an internal coordination channel and an accountable incident lead
- record material decisions, approvals and changes in incident status
- brief leadership when organisational escalation criteria are met rather than on an arbitrary universal cadence
- coordinate provider communications through a defined liaison where third-party evidence or containment is required
- avoid speculation beyond confirmed facts; preserve material communications where they form part of the decision record

18. Notification Is a Decision Process

Notification analysis may involve contractual, legal, regulatory, customer, safety, insurance, partner or internal governance duties. The relevant obligation depends on jurisdiction, entity, regulated activity, incident facts, affected data or systems, consequence and timing. Preserve the decision basis, content, timing and recipients where notification is made or considered.

> **NOTABLY ABSENT**
>
> This guide does not prescribe notification thresholds or deadlines. Those duties must be determined from the applicable legal, regulatory, contractual and sector-specific requirements.

19. Recovery Is Not Just Restarting

Recovery should establish what state is safe and trustworthy enough to restore, what validation is required, who approves return to service, and what residual uncertainty remains. For Physical AI, recovery may require physical inspection, safe-state verification, environmental checks, sensor/actuator validation or hardware-in-the-loop revalidation tied to the identified failure or root-cause hypothesis.

20. Post-Incident Review and Learning

The post-incident review should reconstruct the response without pretending uncertainty has disappeared. A useful review covers:

- timeline: what happened, when and according to which evidence
- detection: how the event was detected and what delayed recognition
- classification: how facts, hypotheses, severity indicators and confidence evolved
- containment: what was constrained, who authorised it and whether it worked
- evidence: what was preserved, what was lost or inaccessible, and why
- notification: what duties were considered and on what basis
- recovery: what validation supported return to service and what residual uncertainty remained

- control feedback: what GAISSF™, PAI-SF™, runbook, provider, logging, training or governance changes are required

METRICS WITH CARE

Measures such as time-to-detect, time-to-contain, evidence completeness or recurrence can support internal learning, but this guide does not define universal targets or benchmarks. Metrics should be interpreted in context; faster is not automatically better if evidence quality, safety or decision quality deteriorate.

21. Test the Capability Before an Incident

Tabletop exercises can test whether roles, evidence access, provider contacts, containment authority, notification analysis and physical-safety interfaces actually work together. Use realistic scenarios and preserve exercise findings as implementation evidence.

- exercise at least one digital AI event and, where applicable, one physical-AI or provider-dependency event
- include the functions that would actually make containment, legal/compliance consultation, business and recovery decisions
- test access to logs, versions, provider support channels and evidence repositories rather than assuming they are available
- record unresolved questions and update procedures, contacts, tooling or training based on findings

NOTABLY ABSENT

This guide does not prescribe a quarterly exercise cadence, mandatory training hours or a universal exercise format. Frequency and depth should reflect system criticality, change rate, incident exposure and organisational obligations.

22. A 15-Minute First-Response Checklist

The time bands below are an illustrative orientation for initial response, not mandatory response times, service levels, escalation deadlines or evidence-preservation thresholds. Safety, system criticality, organisational procedures and applicable obligations may require actions to occur earlier, later or concurrently.

0–5 minutes	5–10 minutes	10–15 minutes
Protect people and critical operations where needed. Record detection source/time. Identify affected system/session. Avoid unnecessary evidence destruction.	Preserve volatile evidence where proportionate. Record facts vs hypotheses. Identify immediate containment options. Check third-party/provider involvement.	Begin proportionate containment if needed. Assign incident owner/escalation. Start the UAIF™-structured record. Begin notification-duty analysis where potentially relevant.

23. First-Hour Extension

- confirm incident ownership, technical lead and required escalation functions
- stabilise evidence preservation and document any unavoidable evidence loss
- establish preliminary system/provider blast radius and linked incidents
- review severity indicators, confidence and unresolved questions
- confirm containment effectiveness and whether additional authority is required
- begin formal legal/regulatory/contractual notification analysis where potentially applicable
- establish recovery criteria and identify evidence or validation still required
- prepare a leadership status that separates confirmed facts, hypotheses, unknowns and next decisions

KEY PRINCIPLE

Respond to consequence and preserve evidence without pretending uncertainty has disappeared. Incident response and incident classification inform one another; neither should be forced into a rigid sequence.

NOTABLY ABSENT

No scenario in this guide reports an actual ODA3 client incident, proprietary telemetry, an incident-frequency statistic, universal severity threshold, legal conclusion, certification result or regulatory determination. No scenario establishes provider culpability, provider SLA or contractual breach, a model-performance warranty, safety certification, or regulatory violation. The examples are hypothetical and instructional.

Next Reading

- The 10-Minute Introduction to GAISSF™
- Applying GAISSF™: Worked Scenarios
- The First 30 Days with GAISSF™
- Classifying AI Incidents: A Practical Guide to UAIF™ — forthcoming
- AI Incident Response Field Guide: Using UAIF™ and AI-IRF™ — forthcoming
- Building an AI Assurance Evidence Pack — forthcoming

Revision Note (R2)

This FINAL R2 revision incorporates four corrections identified during post-publication editorial review. No architectural, framework-relationship or scope changes were made; all edits are additive clarifications within the existing document structure.

Section	Correction
8. RAG / Knowledge-Base Poisoning	Restored "model behaviour" to the competing-hypotheses list, alongside source compromise, indirect prompt injection, retrieval failure and ordinary content error.
2. Orientation table / Integration	Clarified that early triage information remains a local SOC/case-management observation and is not treated as a formal UAIF™ record_type until UAIF™'s own classification criteria are met.
5. Preserve Evidence	Added a Privacy and Privilege Considerations callout addressing masking, access control and legal-hold practice for preserved prompts, outputs and retrieval context.
13. Physical-AI Incident	Added an explicit prohibition on default IT-style containment (network isolation, power-cycling) for cyber-physical systems without safety-engineering authorisation.

Classification on the cover was also updated from "Public / Freely Distributable" to "Public – GEL v1.0 (L1 Public Use)" to align with current GAISSF Ecosystem License labelling used across the published document set.