ODA3 INSTITUTE

WHP — ADOPTION / PRACTITIONER FIELD GUIDE

AI Incident Response Field Guide

Using UAIF™ and AI-IRF™ for Real-World AI Incidents

Document ID: ODA3-2026-08-WHP-COM-013
Classification: Public — GEL v1.0

ODA3 Institute

AI Incident Response Field Guide

Using UAIF™ and AI-IRF™ for Real-World AI Incidents

Document ID: ODA3-2026-08-WHP-COM-013
Document family: WHP - Adoption / Practitioner Field Guide
Publication state: FINAL
Classification: Public - GEL v1.0
Primary audience: SOC/CSIRT leaders, incident responders, CISOs, Security Architects, AI Security teams, AI Governance Leads
Secondary audience: Legal, Privacy, Compliance, Safety, Model/Platform Engineering, Risk Committees
Publisher: ODA3 Institute
Legal entity: ODA3 Pvt Ltd
Date: 12 August 2026 (original FINAL release); reconciled 14 August 2026 (UAIF™/AI-IRF™ Final Publication reconciliation)

Production footer source line: ODA3 sources: TCR-STD-002 | TCR-HAI-001 (79 controls / 19 domains) | AIIS v1.0.0 | current UAIF JSON package | docs.oda3.org

---

Executive Summary

AI incidents rarely arrive as cleanly classified security tickets. A responder may first see an anomalous model output, an unexpected tool call, a retrieval result containing restricted information, a vendor model behaviour change, a physical-AI safety event, or an alert whose relationship to AI is not yet established.

The operational problem is therefore twofold:

What do we know about the event, and how should that knowledge be structured?

What response actions are justified now, even while classification remains incomplete?

The GAISSF Ecosystem separates these concerns deliberately.

UAIF™ — Unified AI Incident Framework provides the structured incident record: identity, causality, severity, harm, incident-versus-vulnerability distinction, generative/agentic/control-plane characteristics, and regulatory/cross-sector metadata.

AI-IRF™ — AI Incident Response Framework provides the response discipline: preparation, detection, classification support, containment, recovery, notification, and learning across AI-specific incidents.

The relationship is not a one-way handoff:

observe → preserve → classify → respond → discover new evidence → reclassify where justified → adjust response → recover → learn

Classification can continue while containment is already under way. Response can generate evidence that changes the classification. Recovery can reveal control failures that need to be fed

back into GAISSF™ governance and, where physical-AI systems are involved, PAI-SF™ controls and safety assurance.

This field guide operationalizes that feedback loop.

It does **not** replace the controlled UAIF or AI-IRF technical specifications, an organization's existing incident-response plan, sector-specific safety procedures, legal analysis, regulatory notification rules, or applicable emergency procedures.

Methodology Note

This guide is a practitioner-oriented implementation document derived from the controlled/current ODA3 source set for UAIF™ and AI-IRF™, together with the established GAISSF Ecosystem architecture.

Public framework-state check - 14 August 2026: ODA3 Institute's current Framework Comparison page identifies **UAIF™ v1.0 as Final Publication** and **AI-IRF™ v1.0 as Final Publication**. A final practitioner guide does not change either underlying framework state, and Final Publication status does not by itself establish completed sector calibration, production-scale validation, legal recognition, or empirical reliability claims.

ODA3 Institute was founded in March 2026. This guide therefore contains no claim of long-term client outcome data, proprietary incident telemetry, incident-frequency statistics, sector-wide effectiveness data, or mature certification-program operating history.

Examples are hypothetical unless explicitly identified otherwise.

The guide distinguishes:

observed facts from causal hypotheses;

provider representations from enterprise-observed evidence;

classification from legal determination;

containment from proof of root cause;

recovery from proof of safety or effectiveness;

evidence presence from evidence sufficiency;

framework correspondence from regulatory equivalence.

AI-IRF™ v1.0 is verified in the corrected technical source and current AIIS v1.0.0 machine-readable companion as a **19-domain, 79-control** framework. AIIS defines structured control records, playbook definitions and severity structures for that inventory. Its optional conformance_test structure is reserved for a future AIIS release and is not populated in v1.0.0. Organizations should therefore treat automation as implementation support, not as a substitute for human incident judgment or independent evidence review.

PART I — THE OPERATING MODEL

1. Why AI Incident Response Needs a Different Evidence Discipline

Traditional cyber incident response already provides mature practices for triage, containment, eradication, recovery, communications, and post-incident learning. AI systems do not invalidate those practices.

They do, however, introduce additional ambiguity.

An unexpected AI outcome may arise from:

malicious input;

indirect prompt injection;

retrieval poisoning;

model or configuration change;

data-quality failure;

authorization failure;

tool-use error;

agent-planning behaviour;

unsafe human workflow;

third-party provider change;

integration failure;

sensor or actuator fault;

environmental conditions;

conventional infrastructure compromise;

a combination of several causes.

The first responder may not know which explanation is correct.

The response process must therefore preserve **actionable uncertainty** rather than force an early narrative.

Key Principle

Do not wait for perfect classification before taking justified protective action. Do not convert protective action into proof of causality.

A responder may isolate an agent, disable a tool, restrict a retrieval source, suspend a model version, place a robot in a safe state, or revoke credentials because the observed condition justifies containment. None of those actions, by themselves, prove the hypothesized root cause.

---

2. The UAIF™ + AI-IRF™ Feedback Loop

The operating model used throughout this guide is:

Operational question	Primary structure	Practitioner outcome
What was observed?	UAIF™	Evidence-bounded incident record
What is known vs hypothesized?	UAIF™	Explicit uncertainty
How severe might this be?	UAIF™ + local policy	Triage input, not legal conclusion
What must be protected immediately?	AI-IRF™ + existing IR/safety process	Containment decision
What evidence could response destroy?	UAIF™ + AI-IRF™	Preservation plan
What response actions are justified?	AI-IRF™	Action plan
What did response reveal?	UAIF™	Updated classification
Can the system return to service?	AI-IRF™ + applicable engineering/safety authority	Recovery decision
What control changes follow?	GAISSF™ / PAI-SF™ where applicable	Governance feedback

The loop

Detection / report
↓
Preserve initial evidence
↓
Create or update UAIF-compatible incident record
↓
Take proportionate AI-IRF response actions
↓
Capture new evidence generated by investigation and containment
↓
Re-evaluate classification, severity, scope, causality and escalation
↓
Adjust containment / investigation / communications
↓
Recover under explicit exit criteria
↓
Feed lessons into GAISSF™ and PAI-SF™ where applicable

This is deliberately iterative.

---

3. What Each Framework Owns

GAISSF™

GAISSF™ remains relevant before, during, and after an incident because governance, control ownership, evidence expectations, supply-chain management, logging, oversight, and corrective action do not stop when an incident begins.

UAIF™

UAIF™ structures the incident record. It is particularly useful for keeping separate:

identity and workflow state;

observed manifestation and causal hypothesis;

severity and harm;

realized incident versus vulnerability/exposure;

generative, agentic, RAG, tool-use and control-plane characteristics;

regulatory and cross-sector routing metadata.

UAIF™ v1.0 is a **Final Publication**. TCR-STD-002 states that it is not a production regulatory-submission instrument, certification scheme, or legal determination of reporting obligations. Regulatory-routing fields provide technical assistance only; applicable legal and sectoral duties require the appropriate legal/regulatory determination.

UAIF™ records are composable rather than monolithic. A record carries a single base **profile** — Core, Enterprise, or Regulatory — that sets its baseline field requirements, plus zero or more additive **extensions** such as SOC_SIEM or AI_Security that layer in extra fields for a specific operational context. Practitioners should confirm which profile and extensions their organization has adopted before treating a field as required or absent.

AI-IRF™

AI-IRF™ structures AI-specific response capability. The verified current inventory contains **79 controls across 19 domains**. The corrected technical source defines the AI-adapted incident-response lifecycle as:

Phase 0 Continuous Discovery → Phase 1 Preparation → Phase 2 Detection & Triage → Phase 3 Machine-Speed Containment → Phase 4 Eradication → Phase 5 Recovery → Phase 6 Post-Incident & Regulatory.

This field guide may use shorter practitioner verbs such as *observe, preserve, classify, respond, recover,* and *learn* to make operational work readable. Those verbs are orientation language only; they do not replace, rename, or add to the controlled AI-IRF lifecycle.

The framework can be integrated with existing SOC/CSIRT processes. It does not require replacement of an organization's SIEM, SOAR, case-management system, ticketing system, crisis-management process, or general cyber incident-response framework.

PAI-SF™

Where AI can produce physical consequences, PAI-SF™ provides the deeper physical-AI security control set. GAISSF™ remains applicable according to determined scope, including the current D9 physical-AI controls where applicable. PAI-SF™ is used for more detailed physical-AI security assurance; it is not a replacement for sector-specific functional-safety, medical-device, automotive, aviation, industrial, or other safety engineering requirements.

3A. Establish the Operational State Without Forcing Classification

An anomalous AI-related signal does not automatically become a confirmed AI incident. During early triage, responders should preserve the distinction between:

Signal / event under triage — an observation requiring assessment.

Suspected AI incident — available evidence justifies incident-response handling, but material classification or causality remains provisional.

Managed incident — the organization has invoked its applicable incident-management process based on its own criteria.

These are **local practitioner workflow states**, not UAIF™ enums. TCR-STD-002 places identity and workflow state within **L0 — Record Identity & Workflow**, which also carries the record_type field, but does not establish a workflow_state field matching these three labels. Teams should therefore use their existing intake/triage/investigation workflow until evidence supports the applicable record_type determination in L0. **L4 — Incident– Vulnerability Relationship** is a separate, later layer that records linkage and exposure context (such as linked_vulnerability_id and exposure_duration_hours) once a record already exists; it does not itself decide whether a matter is an Incident or a Vulnerability. Any UAIF payload must follow the applicable controlled profile/schema rather than encoding these three labels as UAIF values.

A team may take protective action while the matter remains under triage. The decision to contain is evidence of a risk-management decision; it is not proof that the event was caused by AI, malicious activity, or a specific control failure.

AI Involvement Test

Before describing an event as an **AI incident**, ask how AI is actually involved:

Affected asset: Was an AI model, AI application, agent, retrieval system, model gateway, evaluation system, or AI-supporting component itself affected?

Attack surface: Was an AI-specific interface or mechanism materially used, such as prompting, retrieval, tool use, model supply chain, or agent control?

Causal mechanism: Is there evidence that AI behaviour materially caused or contributed to the observed outcome?

Enabling mechanism: Did AI materially increase scale, speed, autonomy, or capability even if the underlying incident is conventional?

Evidence source: Is AI telemetry or AI-generated content relevant only as evidence rather than as the cause?

Incidental presence: Is AI merely present in the affected system while the established mechanism is conventional infrastructure, identity, application, data, mechanical, or human-process failure?

More than one relationship may apply.

Do not infer AI causality merely because an AI-enabled system is involved. If the relationship remains unknown, record that uncertainty and continue investigation.

3B. The AI-IRF™ Lifecycle in Field Use

The corrected AI-IRF™ technical source defines seven controlled phases. This field guide does not replace those phases. It translates them into responder questions, evidence expectations, and decision points.

Phase 0 - Continuous Discovery

Operational objective: know which AI systems, agents, model endpoints, MCP/tool services, retrieval systems, and material dependencies exist before an incident requires them to be found under pressure.

A field responder should be able to answer:

Which production AI instances and endpoints are known?

Which are sanctioned, experimental, shadow, personal, or third-party?

Which models/providers/versions are in use?

Which systems can act, write, send, purchase, change configuration, or influence physical processes?

Which identities, service accounts, tool permissions, model gateways, vector stores, and data sources are material?

Which systems share the same provider/model/retrieval dependency?

Where is authoritative ownership recorded?

Evidence that may support discovery readiness: asset inventories, AI/application manifests, service catalogs, endpoint/model gateway records, provider inventories, CMDB records, architecture diagrams, service-principal inventories, data-flow maps, tool registries, model release records, and dependency records.

Common failure mode: treating inventory completeness as a binary claim. Discovery can be partial. Record coverage boundaries and unknown estate rather than describing the inventory as complete without evidence.

Incident implication: previously unknown AI use can widen the blast radius. Discovery findings should update both the incident scope and the governance baseline.

Phase 1 - Preparation

Preparation converts incident-response intent into executable authority.

Before an incident, teams should establish:

who can disable, isolate, restrict, roll back, or place a system into a safe state;

which actions require dual control or specialist approval;

what evidence should be captured before destructive containment;

how provider escalation occurs;

how regulated-sector safety/clinical/legal functions enter the response;

which fallback/manual modes exist;

which playbooks exist and how they are versioned/tested;

how recovery authority differs from containment authority.

The corrected AI-IRF source emphasizes pre-approved playbooks, human-in-the-loop authority, signed/verified artifacts, segmentation, exercises, and rollback readiness. COM-013 uses those concepts as preparation prompts without importing every historical quantitative target as a universal requirement.

Preparation artifact set: - AI incident authority matrix; - provider escalation directory; - playbook register; - evidence-preservation map; - system-specific emergency-action card; - rollback/fallback record; - sector safety/regulatory contacts; - tabletop exercise record; - recovery-authorization record.

Phase 2 - Detection & Triage

Detection is not limited to a SOC alert. Relevant signals can come from:

users;

model/application evaluation;

business process anomaly;

fraud or abuse operations;

privacy complaints;

provider notices;

model gateway / API telemetry;

retrieval/data pipeline monitoring;

agent/tool-call monitoring;

safety interlocks;

OT/physical-system alarms;

red-team or evaluation activity;

third-party intelligence.

The first triage task is not “pick a root cause.” It is:

preserve the observation;

determine immediate safety/impact concerns;

establish whether AI is affected, causal, enabling, evidentiary, or merely present;

identify systems/dependencies;

preserve volatile evidence;

decide whether proportionate containment is justified;

establish the next decision checkpoint.

Triage failure mode: treating high detection confidence as high causal confidence. A detector can be confident that an anomaly exists while the cause remains unknown.

Phase 3 - Machine-Speed Containment

AI-IRF’s technical source includes rapid and automated containment patterns. COM-013 does not assume that automation is appropriate for every organization or incident.

A field implementation should distinguish:

Low-consequence automated actions - rate limiting; - temporary session isolation; - read-only mode; - disabling nonessential autonomous actions; - quarantine of a suspicious retrieval object where safe and reversible.

Consequential actions requiring explicit authority - credential revocation; - production rollback; - deletion/reindexing; - account quarantine; - stopping critical service; - emergency stop or actuation isolation; - blocking a provider integration used by multiple business processes.

The implementation question is:

Which actions can be safely automated, which require human approval, and which require joint security/safety/business authority?

Every automatic action should remain traceable to the triggering evidence, rule/logic, timestamp, system, and resulting state.

Phase 4 - Eradication

Containment limits harm. Eradication addresses the condition that must be removed or neutralized before normal operation can resume.

Depending on the incident, eradication may include:

removing poisoned or untrusted retrieval content;

rotating compromised credentials;

removing an unauthorized connector or tool permission;

rebuilding an affected index or deployment;

replacing a compromised artifact from a trusted source;

removing malicious instructions or configuration;

correcting an authorization-policy defect;

eliminating a compromised model/plugin/package from the release path;

removing persistence in conventional infrastructure;

correcting unsafe agent workflow or approval logic.

Evidence discipline: eradication is not established merely because the observable symptom disappeared. Preserve evidence showing what changed and why the team believes the relevant condition was removed.

Rebuild boundary: rebuilding can destroy evidence. Preserve the necessary pre-change state first where safety permits.

Phase 5 - Recovery

Recovery is a controlled return to an acceptable operating state.

A good recovery decision is based on:

defined remediation;

validation evidence;

known residual uncertainty;

monitoring;

rollback;

authorized scope;

business/safety conditions;

explicit return-to-service authority.

Recovery may be staged:

isolated validation;

limited/canary use;

increased monitoring;

bounded user population;

progressive restoration;

normal operations after evidence-supported review.

For physical-AI and regulated systems, the authorized sector process controls return to service.

Phase 6 - Post-Incident & Regulatory

Post-incident work should preserve the distinction between:

technical findings;

regulatory-routing indicators;

legal determinations;

external notifications;

governance learning.

The technical source includes regulatory decision-tree concepts and evidence packaging. COM-013 does not convert technical severity or routing logic into legal-reporting rules.

Outputs can include:

final incident/classification record;

evidence index;

decision history;

notification record where applicable;

unresolved-issue register;

corrective-action plan;

provider/vendor findings;

GAISSF™ control feedback;

PAI-SF™ feedback where applicable;

playbook changes;

exercise/test updates.

What spans all seven phases

Across every phase, preserve:

identity → evidence → uncertainty → decision authority → action → resulting state → new evidence → next decision

That record is more durable than a single static incident label.

3C. Regulated-Sector Adoption Boundary - Read Before Operational Use

Regulated and safety-relevant organizations should establish this boundary before applying the operational workflow. ODA3 incident frameworks add AI-specific security, evidence and classification discipline; they do not displace the authority that governs clinical safety, product safety, airworthiness, roadworthiness, machinery safety, critical-infrastructure operations, defence authorization or legal reporting.

Context	ODA3 contribution	External / sector authority remains controlling	Notably Absent
Healthcare decision support	UAIF™ structures facts, uncertainty and incident classification; AI-IRF™ structures response	Clinical governance, privacy, professional obligations and applicable health regulation	Does not establish clinical effectiveness or medical judgment
Medical devices / robotics	PAI-SF™ adds physical-AI security considerations; UAIF/AI-IRF support incident evidence and response	Medical-device risk management, clinical validation, quality systems and regulator authorization	Does not establish device approval or safe clinical use
Automotive / ADAS	PAI-SF™ supports perception, autonomy, actuation and cyber-physical incident reasoning	Vehicle safety engineering, SOTIF/functional-safety processes, type approval and roadworthiness	Does not establish roadworthiness or vehicle approval
Aviation / drones	PAI-SF™ supports navigation, autonomy, command/control and physical-AI incident evidence	Aviation regulation, airworthiness, operational approval and authorized contingency procedures	Does not establish airworthiness or flight authorization
Manufacturing / OT / robotics	PAI-SF™ supports motion, actuation, autonomy and AI/OT evidence interfaces	Machinery/robotics safety, functional safety, OT operating procedures and plant authority	Does not replace machinery or functional-safety processes
Energy / critical infrastructure	UAIF/AI-IRF support incident reasoning, evidence and coordinated response; PAI-SF may apply to physical AI	Utility/OT safety, reliability, emergency management and sector regulation	Does not establish reliability, grid safety or operational authorization
Defence / national-security contexts	UAIF may preserve classification, provenance and uncertainty where authorized	Defence authority, classified-system rules, mission authorization and applicable law	Does not authorize targeting, weapons use or operational deployment

Conflict rule: where a sector-specific safety, clinical, operational or regulatory process conflicts with a generic sequence in this guide, the applicable authorized sector process governs.

The more detailed implementation treatment remains in the sector/safety section later in this guide.

PART II — PREPARE BEFORE THE INCIDENT

4. Minimum Incident-Response Readiness

An organization should not wait for an AI incident to determine:

which AI systems exist;

who owns them;

which third parties operate material components;

where logs are retained;

who can disable a model, agent, tool, retrieval source, or autonomous function;

which evidence may be ephemeral;

who can authorize emergency containment;

who decides whether a system can return to service;

when Legal, Privacy, Safety, Compliance, Communications, or executive leadership must be engaged.

Minimum readiness record

For each in-scope AI system, maintain enough information to answer:

System / service name and owner.

Business process supported.

Model/provider and deployment model.

Data sources and retrieval sources.

Tool/API permissions.

Agent autonomy and approval gates.

Identity and secrets dependencies.

Logging locations and retention constraints.

Critical third-party dependencies.

Safety/physical consequence potential.

Emergency containment options.

Recovery authority.

Applicable legal/regulatory review routes.

Evidence custodian(s).

This is not a universal mandatory template. It is a minimum operational orientation for avoiding basic discovery delays during response.

---

5. Roles and Decision Rights

A mature response does not require a large dedicated AI incident team, but it does require explicit decision rights.

Role	Typical incident responsibility
Incident commander / response lead	Coordinates response, decisions, timeline and escalation
SOC / CSIRT	Detection, triage, evidence capture, containment coordination

Role	Typical incident responsibility
AI / model owner	Model behaviour, versions, evaluation history, provider interface
Platform / application engineering	Runtime, integrations, deployment and rollback
Identity / cloud security	Credentials, service principals, access and cloud evidence
Data / RAG owner	Retrieval sources, data lineage, permissions and poisoning investigation
Safety / OT / physical-system owner	Safe state, physical isolation, return-to-service constraints
Legal / Compliance	Legal characterization, privilege, notification obligations
Privacy	Personal-data impact and privacy notification analysis
Vendor / procurement owner	Provider escalation, contractual evidence and change records
Communications	Internal/external messaging where required
Executive / risk authority	Material-risk decisions and business acceptance

Small-Team Adaptation

One person may hold several roles.

Resource constraint does not make missing evidence, unresolved authority, or unassigned risk disappear.

Where roles are combined, record who made the material decision and what evidence supported it.

---

6. Evidence Preservation Plan

Before an incident, identify evidence that may disappear through normal operations or containment.

Examples include:

prompts and system prompts;

retrieved context;

model outputs;

tool-call requests and responses;

agent plans and intermediate steps where available;

model/version identifiers;

deployment/configuration snapshots;

identity and authorization events;

API gateway records;

safety-controller events;

sensor and actuator telemetry;

model-provider status/change notices;

vector-store changes;

document-ingestion records;

CI/CD and model-release records;

human approval/override records;

network and cloud logs;

case/ticket history.

Evidence Quality

Evidence should be evaluated for:

relevance → authenticity → traceability → completeness → currency → linkage to the incident scope

A screenshot may be useful evidence. It is not automatically sufficient evidence.

A vendor statement may be useful evidence. It should remain identifiable as a vendor representation unless independently corroborated.

6.1 Tiered Observability Baseline

The evidence classes in this guide are **not** a requirement to collect every possible artifact for every AI system. Logging should be proportionate to system consequence, autonomy, data sensitivity, provider dependency, investigation need, privacy constraints, retention obligations and cost.

The following three-tier pattern is an **illustrative implementation baseline**, not a GAISSF/UAIF/AI-IRF normative tier or certification level.

Illustrative tier	Typical system characteristics	Minimum incident-observability baseline to consider	Additional evidence when justified
Tier 1 - Bounded / internal	Internal assistant or RAG; limited autonomy; no direct external action; bounded repositories	Timestamp; system/application ID; model/provider/version where available; user/session or workload identity as permitted; policy/authorization decision; high-level request/output event; relevant deployment/configuration version	Retrieval document/source IDs; prompt-template version; DLP/security decision; provider change notice
Tier 2 - Consequential / sensitive	Decision support, regulated/sensitive data, external-facing use, material business consequence, meaningful human review expected	Tier 1 plus prompt/template/configuration version; retrieval/source IDs and authorization result; output/action record; human approval/override and rationale where applicable; model/application change record; relevant identity/API/cloud events	Evaluation result; provider incident/change evidence; data-ingestion provenance; decision trace sufficient to reconstruct the consequential action
Tier 3 - Autonomous / physical / high-consequence	Agents with consequential tools; autonomous transactions; physical AI; safety-relevant or critical operations	Tier 2 plus complete consequential tool/action trace; before/after state; approval-gate evidence; sensor/actuator or safety-controller snapshot where applicable; autonomy/safe-state mode; release/update provenance; rollback/fallback event; joint security/safety decision record	High-fidelity execution trace; environment/ODD state; hardware/firmware/controller version; redundant telemetry; evidence needed by the controlling safety/sector process

Tiering rules

Internal-facing does not mean low consequence. An internal RAG may require Tier 2 evidence when it can expose highly sensitive data or materially affect decisions.

Tier upward for consequence, not marketing labels. A nominal “copilot” that can execute payments or production changes should be treated according to its actual authority.

Provider opacity does not remove local evidence obligations. Capture what the enterprise can observe and identify the provider-side unknown explicitly.

Do not collect restricted data merely because it may be useful later. Privacy, employment, secrecy, privilege, sector and cross-border constraints remain applicable.

Retention is separate from capture. A system may capture an event without retaining full content indefinitely.

System-specific readiness cards should record the chosen baseline and any known gaps.

The objective is not maximum telemetry. It is **decision-sufficient, proportionate and defensible evidence**.

6A. Evidence Custody and the Incident Evidence Pack

AI incident evidence often spans several systems that use different clocks, retention policies, identities, owners, and formats. A defensible evidence pack therefore needs more than a folder of screenshots.

Evidence index

For each material artifact, record where practicable:

evidence identifier;

source system;

source owner;

collection time;

relevant event time;

collector;

collection method;

original/native format;

integrity/hash information where appropriate;

access restrictions;

relationship to the incident question;

whether the artifact is original evidence, derived evidence, or a provider representation;

retention/hold requirements under local process.

Original vs derived evidence

Examples:

Original / source evidence - native logs; - raw provider notice; - captured model/API response; - configuration export; - signed release artifact; - safety-controller event; - source document retrieved by RAG.

Derived evidence - analyst timeline; - screenshot; - parsed log extract; - dashboard export; - correlation graph; - severity calculation; - summary prepared for leadership.

Derived evidence can be useful. Preserve the relationship back to the source where possible.

AI-generated evidence summaries

Teams may use AI to summarize logs, cluster incidents, translate provider messages, or draft timelines. If AI assists evidence analysis:

preserve the underlying source;

record the tool/model used where material;

do not treat the generated summary as proof;

validate high-consequence conclusions;

avoid exposing restricted incident data to unauthorized AI services.

Clock and sequence problems

AI incidents frequently cross: - application logs; - provider timestamps; - cloud audit records; - identity platforms; - SIEM; - data pipelines; - device/robot clocks; - safety controllers.

Record known clock offsets or uncertainty. Do not overstate sequence when time synchronization is uncertain.

Privacy and minimization

Evidence preservation does not automatically authorize unrestricted collection. Apply the organization’s legal, privacy, employment, data-retention, cross-border, privilege and investigation rules.

Preserve enough to answer the incident question without turning the evidence pack into an uncontrolled secondary data repository.

6B. Evidence Sufficiency Questions

Before relying on an artifact for a material decision, ask:

Relevance: Does it address the actual incident question?

Authenticity: Do we know where it came from and whether it is genuine?

Traceability: Can another reviewer connect it to the system, time, decision or event?

Completeness: Is critical context missing?

Currency: Does it represent the relevant version/time period?

Scope linkage: Does it relate to the systems/users/processes actually in scope?

Independence: Is it enterprise-observed, provider-supplied, or independently reproduced?

Reproducibility: Can the finding be safely repeated, and if not, is the limitation recorded?

Decision fitness: Is this artifact sufficient for this decision, or only one input?

No universal hierarchy makes a screenshot always weak or an immutable log always strong. Evidence quality is decision- and context-dependent.

PART III — THE FIRST RESPONSE WINDOW

7. The First 15 Minutes

The exact sequence depends on system context and safety. The following is an orientation, not a universal SLA.

0–5 minutes: Stabilize and preserve

Establish an incident/case identifier.

Record who detected or reported the condition.

Capture the initial observation in the reporter's terms.

Preserve volatile evidence where doing so does not increase safety risk.

Identify immediate human-safety or physical-consequence concerns.

Determine whether emergency isolation or safe-state action is required.

5–10 minutes: Bound the system

Identify affected AI system, model/provider, version and environment.

Identify exposed data, tools, credentials and downstream systems.

Determine whether the event appears isolated or potentially shared across systems.

Identify third-party dependencies.

Record what is still unknown.

10–15 minutes: Initiate classification and response

Create/update the UAIF-compatible record.

Separate facts from hypotheses.

Establish an initial severity/impact view under the applicable UAIF/local method.

Select proportionate containment.

Assign investigation owners.

Start Legal/Privacy/Safety/Compliance routing where applicable.

Notably Absent

Completing this checklist does not establish root cause, legal reportability, materiality, control failure, conformance, or system safety.

---

8. The First Hour

By the end of the first operational hour, where circumstances permit, the team should aim to have:

an incident owner;

an initial evidence inventory;

an explicit list of unknowns;

a bounded blast-radius hypothesis;

an initial containment state;

provider/vendor escalation initiated where needed;

affected model/system versions recorded;

relevant identities and credentials reviewed;

a classification record with provisional status clearly visible;

safety, privacy, legal and regulatory review routes engaged where applicable;

a next-decision time.

First-Hour Brief

A useful first-hour leadership brief answers:

What happened?
Only what is observed.

What is affected?
Systems, data, users, processes, physical assets.

What have we done?
Containment and preservation actions.

What do we think may have happened?
Working hypotheses, clearly labelled.

What do we not know?
Material uncertainties.

What is the next decision?
What must be decided, by whom, and by when.

8A. Four Operational Decision Checkpoints

The feedback loop should remain iterative, but material decisions should be traceable. The following checkpoints provide an implementation pattern; they are not universal SLAs and do not create new UAIF™ or AI-IRF™ lifecycle stages.

Checkpoint	Decision	Minimum decision record
CP1 — Initial Triage	Continue monitoring, invoke incident handling, or take immediate protective action	Observation, affected system, immediate safety/impact concern, known/unknown facts, initial evidence preserved, decision authority

Checkpoint	Decision	Minimum decision record
CP2 — Containment Review	Maintain, expand, reduce, or reverse containment	Current scope, evidence gained/lost, working causal view, containment effectiveness, downstream impact, unresolved uncertainty
CP3 — Recovery Authorization	Permit bounded return to service or continue containment	Remediation evidence, validation/test results, residual uncertainty, monitoring, rollback, required safety/legal/privacy conditions, authorizer
CP4 — Closure / Learning	Close operational response while preserving follow-up obligations	Final known facts, unresolved questions, classification history, realized harm, residual risk owner, corrective actions, governance feedback

A checkpoint can be revisited whenever new evidence materially changes the decision basis.

PART IV — CLASSIFY WHILE RESPONDING

9. Maintain a Living UAIF Record

The incident record should evolve as evidence evolves.

A useful operating discipline is:

Confirmed

Supported by evidence currently available.

Working hypothesis

Plausible explanation under investigation.

Provider representation

Statement supplied by an external provider or supplier.

Unknown

Material fact not yet established.

Not observed

Condition specifically checked for and not found within the stated scope.

These are practitioner labels for evidence discipline, not asserted UAIF controlled values. Implementers should map only to fields and vocabularies actually defined by the applicable UAIF profile/schema.

---

10. Reclassification Triggers

Revisit classification when:

new systems are found affected;

realized harm differs from initial assumptions;

provider evidence changes the causal hypothesis;

a vulnerability becomes a realized incident;

an apparent AI issue is traced to conventional infrastructure;

an apparent benign failure shows adversarial indicators;

a physical consequence occurs;

a downstream party reports impact;

new jurisdiction/sector facts change routing;

containment reveals a broader shared dependency.

Every material change should preserve:

previous classification;

new classification;

evidence that prompted the change;

reviewer/decision authority;

time of change;

resulting response change.

---

11. Severity Without False Precision

UAIF provides structured severity outputs, but a numerical score should not become a substitute for judgment.

Ask:

What harm is realized?

What harm remains potential?

How many systems/users/processes are affected?

Is the condition ongoing?

Can the system act autonomously?

Is there a physical consequence?

Is sensitive data involved?

Is there evidence of malicious activity?

Are critical services affected?

Is the causal mechanism understood?

How confident are we in the underlying evidence?

Confidence Discipline

Do not invent a parallel ODA3 confidence scale in this guide. Use the authoritative UAIF terminology and local incident-management controls.

PART V — CONTAINMENT PATTERNS

12. Containment Is a Risk Decision

Containment should reduce expected harm without unnecessarily destroying evidence or creating a larger operational/safety hazard.

Potential actions include:

disable an AI feature;

revoke a tool permission;

restrict agent autonomy;

block an external connector;

quarantine a retrieval source;

remove a document from an ingestion pipeline;

rotate credentials;

pin or roll back a model/application version;

switch to a non-AI/manual fallback;

suspend automated decisions;

disable outbound actions;

isolate a robot or autonomous system;

place a physical system into an engineered safe state;

restrict user population;

disable a provider integration.

The correct action depends on context.

Reversibility and Evidence Impact

Where safety and expected harm permit, prefer containment that is reversible and preserves the ability to investigate. Examples may include narrowing a permission, disabling a connector, pinning a version, isolating a retrieval source, or moving to a controlled fallback.

Some actions may materially alter or destroy evidence: rebuilding an index, wiping a workload, resetting state, overwriting configuration, replacing an artifact, clearing caches, or physically resetting equipment.

Before a potentially destructive action, where circumstances permit, record:

the observed condition requiring action;

why delay would create unacceptable risk;

evidence expected to be altered or lost;

preservation completed before action;

alternatives considered;

decision authority;

rollback or reconstruction options.

Human safety, engineered safe-state requirements, and applicable emergency procedures take precedence over evidence-preservation convenience.

---

13. Generative-AI Containment

For prompt injection, unsafe output, disclosure, or retrieval-related incidents, consider:

preserve prompt/context/output before modification;

identify direct and indirect prompt sources;

identify retrieved documents;

isolate poisoned/untrusted retrieval sources;

inspect tool calls and authorization;

review system-prompt/configuration changes;

restrict high-risk actions while investigation continues;

preserve provider/version information;

test whether the behaviour reproduces in a safe environment.

Do Not Assume

A malicious prompt followed by an unsafe output does not, by itself, establish that the model is the root cause. Application controls, retrieval configuration, tool permissions, identity, data boundaries and provider behaviour may all matter.

---

14. Agentic-System Containment

Agentic incidents can propagate faster because the system may call tools, create artifacts, alter data, send messages, make purchases, change infrastructure, or trigger other agents.

Priorities may include:

Stop or restrict high-consequence actions.

Revoke or narrow credentials.

Disable affected tool integrations.

Preserve tool-call history.

Identify chained/sub-agent actions.

Identify external systems modified.

Record human approvals/overrides.

Prevent automatic retry loops.

Establish transaction rollback options.

Re-evaluate blast radius.

A stopped agent is not necessarily a contained incident if previously issued actions remain active downstream.

---

15. RAG / Knowledge-System Containment

Investigate:

retrieval boundary enforcement;

document permissions;

ingestion history;

poisoning or malicious embedded instructions;

chunking/indexing changes;

cross-tenant/cross-user retrieval;

vector-store access controls;

cached context;

downstream output distribution.

Containment may involve quarantining sources, rebuilding indexes, restricting retrieval scope, or disabling specific connectors.

Do not call an event RAG leakage as a root cause until the retrieval mechanism is established.

---

16. Third-Party Model / SaaS Containment

Most enterprises cannot inspect provider model weights or internal training pipelines.

That limitation should be recorded, not disguised.

Separate:

Enterprise-observed evidence - requests sent; - outputs received; - timestamps; - versions exposed by API; - local configuration; - local evaluations; - application logs; - user impact.

Provider representations - incident notices; - status-page statements; - model-change notices; - support responses; - SOC reports; - model/system cards; - contractual attestations.

Unknown provider-internal facts - exact internal change; - training-data issue; - internal compromise; - model-weight change; - hidden routing; - infrastructure failure.

Containment can still proceed locally by disabling integration, pinning a version where supported, applying output restrictions, switching provider/model, increasing human review, or moving to fallback workflows.

16A. Third-Party Evidence Package

When the provider controls the model or service internals, the enterprise incident team should create a provider evidence package rather than collapsing everything into “waiting for vendor.”

A useful package includes:

Enterprise observations

exact request/output examples;

timestamps;

exposed model/version where available;

local configuration;

affected applications;

local safety/business impact;

reproducibility results;

known unaffected controls.

Provider evidence requested

incident/change reference;

affected service/model/version;

confirmed timeline;

scope;

provider-confirmed cause if established;

mitigation;

rollback/version options;

customer actions;

evidence-preservation confirmation;

future change-notification commitment where relevant.

Contract / assurance context

service description;

security/assurance commitments;

SOC/independent reports available to the customer;

incident notification clauses;

change-management commitments;

data-handling/location commitments;

support/escalation route.

Unresolved provider boundary

If the provider cannot disclose internal evidence, record:

what was requested;

what was supplied;

what remains unknown;

whether local testing reproduced the condition;

whether the residual uncertainty affects recovery/closure;

who accepted the residual uncertainty.

A closed provider ticket is not proof that enterprise causal questions are resolved.

16B. Closing With Provider Unknowns

A SaaS/provider-dependent incident may reach operational closure before the provider supplies a complete post-incident report.

Before closing with material provider-side unknowns, record:

what enterprise facts are sufficient to close the immediate response;

what provider questions remain open;

whether a separate vendor/dependency action remains active;

the owner of that dependency;

the expected provider follow-up channel;

conditions that would reopen or amend the incident;

whether a later provider report requires classification, severity, notification, corrective-action or governance reassessment.

A late provider report should not overwrite the historical incident record. Append the new evidence, identify when it became available, and document any resulting change in conclusion.

---

17. Physical-AI Containment

Human safety and engineered safe-state requirements take priority over preservation convenience.

Potential actions include:

emergency stop;

isolate actuation;

reduce operating envelope;

move to manual control;

disable autonomous mode;

geofence;

suspend operation;

preserve safety-controller records;

preserve sensor and actuator telemetry;

preserve model/software/firmware versions;

inspect recent updates;

protect the physical scene/equipment where relevant.

Do not reproduce an unsafe condition merely to obtain cleaner evidence.

Return to service may require specialist engineering, safety validation, simulation, hardware-in-the-loop testing, regulatory review, or other sector-specific assurance beyond this guide.

17A. Sector and Safety Boundary Matrix

The table below is an **interface guide**, not a compliance crosswalk. It shows which external authority or safety discipline must continue to govern alongside UAIF™, AI-IRF™, GAISSF™ and PAI-SF™.

Sector / system context	ODA3 incident-response contribution	External regime / authority that remains controlling	Evidence or recovery emphasis	Notably Absent
Healthcare decision support	UAIF structures incident facts/classification; AI-IRF structures response	Clinical governance, privacy/data protection, applicable healthcare and professional obligations	Input/output/version, clinician review, downstream decision, correction and notification record	Does not establish clinical effectiveness, negligence, standard of care, or reportability
Medical devices / medical robotics	PAI-SF adds physical-AI security controls; UAIF/AI-IRF support incident classification/response	Medical-device risk management, clinical validation, quality management, product safety and regulator authorization	Sensor/actuator state, autonomy mode, override, device/model version, patient-safety action, service/validation evidence	Does not establish device approval, clinical validation, treatment safety, or regulator clearance
Automotive / mobility / ADAS	PAI-SF supports perception, autonomy, actuation, safe-state and recovery evidence	Vehicle safety engineering, SOTIF/functional-safety processes, type approval, roadworthiness, transport/deployment rules	ODD, sensor integrity, steering/braking/throttle commands, fallback/takeover, fleet update and recovery evidence	Does not establish roadworthiness, type approval, functional safety or autonomous-driving authorization
Aviation / drones / uncrewed systems	PAI-SF supports navigation, autonomy, payload/command, safe-state and incident reconstruction	Aviation regulation, airworthiness, operational approval, pilot/operator qualification and mission authorization	GNSS/navigation, command/payload state, geofence, link loss, return-to-home/landing, override and maintenance evidence	Does not establish airworthiness, aviation approval or mission authorization
Manufacturing / robotics / OT	PAI-SF supports motion/actuation, autonomy boundaries, safety-interface evidence and physical recovery	Machinery/robotics safety, occupational safety, functional safety, plant/OT procedures and local regulation	Interlocks, motion bounds, PLC/controller integration, manual takeover, degraded mode, maintenance and HIL/site validation	Does not replace machinery safety, functional safety or industrial certification
Energy / critical infrastructure / smart infrastructure	UAIF/AI-IRF support incident reasoning; PAI-SF applies where AI materially affects physical processes	Utility/OT safety, emergency management, resilience, critical-infrastructure and sector regulatory requirements	Sensor/command state, operator action, physical command, degraded mode, rollback, service-continuity and public-safety evidence	Does not establish reliability, sector compliance, emergency authorization or public-safety acceptance
Defence / national-security systems	UAIF can preserve classification, uncertainty and provenance; AI-IRF can structure response	Applicable defence authority, classified-system authorization, operational security, mission rules, export/control and international/national legal obligations	Classified evidence handling, provenance, command authority, human authorization, mission consequence, supply chain and compartmented access	Does not authorize targeting, weapons use, autonomous lethal action, classified-system accreditation or mission approval

Boundary rule for regulated sectors

Where a sector-specific safety, clinical, operational or regulatory process requires an action that differs from this guide's generic sequence, the applicable authorized sector process governs.

For example:

a clinician or device-safety authority may suspend a medical device before a complete cyber classification exists;

an autonomous vehicle may require a safe fallback rather than an immediate stop if stopping itself creates greater hazard;

an aircraft/uncrewed system may require an approved contingency procedure rather than ad hoc isolation;

an industrial control process may require controlled shutdown or manual operation instead of immediate power removal;

critical infrastructure may require service-continuity and public-safety decisions alongside cyber containment;

defence incidents may require classified handling and operational-security controls that limit normal evidence sharing.

ODA3 incident records support these decisions; they do not supersede the authorized decision-maker.

17B. Playbook Design Pattern

The corrected AI-IRF source includes an adaptive playbook library for priority incident patterns such as agent privilege abuse, MCP compromise, model extraction, poisoned RAG, shadow-AI discovery, and physical-safety triggers. COM-013 uses that structure as a design reference, not as a claim that every organization must implement the same automation or thresholds.

A field-ready playbook should answer:

Trigger

What observation or combination of observations starts the playbook?

Preconditions

What must be known before an automatic or manual branch executes?

Immediate evidence capture

Which volatile evidence must be preserved before action?

Branch logic

What changes when confidence, scope, harm, autonomy, or physical consequence changes?

Containment action

What is the least-destructive action that sufficiently reduces expected harm?

Human approval

Which actions require: - analyst approval; - incident commander approval; - system owner approval; - Legal/Privacy approval; - safety-engineering approval; - two-person control?

Escalation path
Who owns the next decision?

Recovery gate
What must be demonstrated before the playbook can transition from containment to recovery?

Closure / learning
What evidence and governance updates are required before the playbook is considered complete?

Playbook anti-patterns
Avoid: - hard-coded legal notification solely from a technical severity score; - deleting evidence as the first containment action where safer reversible options exist; - auto-revoking critical infrastructure or physical-system authority without safety/operations review; - assuming the same playbook fits SaaS, self-hosted, agentic and physical-AI systems; - treating a successful automation action as proof that the causal hypothesis was correct; - allowing playbook logic to drift without version/change control.

17C. Six Source-Derived Priority Playbook Archetypes

The AI-IRF technical source identifies six priority playbook archetypes. The table below translates them into field-guide questions while avoiding universalization of historical quantitative thresholds.

Archetype	Typical trigger	Immediate objective	Human decision point	Evidence to preserve	Recovery focus
Agent privilege abuse	Tool action outside intended permission or anomalous write/export	Stop consequential agent actions and bound downstream changes	Credential revocation, production delete/financial actions, broad account quarantine	Agent plan, tool calls, identity/authorization, approvals, before/after state	Permission-scope correction, downstream repair, approval-gate validation
MCP / tool-server compromise	Unauthenticated/ tampered request, malicious skill/tool payload, abnormal A2A/tool traffic	Isolate broker/server and prevent propagation	Destructive remediation and trust-domain credential action	Endpoint state, tool/skill package, identities, connected agents, network/API records	Trusted rebuild, identity rotation, re-attestation/integration validation where supported
Model extraction / abuse	Sustained unusual query similarity/volume or evidence of extraction behavior	Rate-limit/isolate abusive sessions and preserve query evidence	Credential revocation, legal hold, customer/account action	Queries, outputs, identities, rate-limit data, session correlation	Abuse controls, access policy, monitoring, customer/legal follow-up
Poisoned RAG	Provenance failure, suspicious source, index-integrity anomaly	Quarantine source and prevent further retrieval	Reindex/rebuild and source reinstatement	Source, ingestion, chunks, index/version, retrieval context, output	Trusted-source validation, controlled reindex, retrieval tests
Shadow AI discovery	Unregistered endpoint or unsanctioned AI with material data access	Bound data exposure and identify owner/use	Allowlist/block decision, exposure notification	Endpoint, identity, data flow, configuration, user/business owner	Registration, approved controls, migration/blocking, data-exposure follow-up
Physical-safety trigger	Cyber anomaly plus unsafe command/safety signal	Protect people and force safe/degraded state under authorized procedure	Any override of the safe state or consequential return-to-service action	Sensor/actuator/safety-controller state, commands, autonomy mode, version, environment	Joint safety/engineering validation and controlled return to service

PART VI — INVESTIGATION

18. Build the Timeline

Create a timeline that distinguishes:

first known precursor;

first observed anomalous behaviour;

first user/customer impact;

first detection;

first containment;

provider notifications;

configuration/model changes;

credential events;

retrieval/index changes;

physical/safety events;

investigation findings;

classification changes;

recovery decisions.

Record the source of each timeline item.

---

19. Causal Analysis Without Premature Closure

A useful causal record separates:

Observed manifestation — what happened.

Immediate mechanism — how the observed behaviour occurred, if established.

Contributing conditions — controls, configuration, workflow, dependencies.

Root-cause hypothesis — proposed deeper cause.

Root cause — only when evidence supports the conclusion.

Unknowns — unresolved causal questions.

Temporal sequence alone does not establish causality.

---

20. Blast Radius

For AI incidents, blast radius may extend through shared:

foundation models;

provider endpoints;

retrieval sources;

prompts/templates;

agents;

tools;

identity/service accounts;

model gateways;

data pipelines;

evaluation infrastructure;

safety controllers;

deployment artifacts.

When multiple systems share a dependency, determine whether to use:

one incident with multiple affected systems;

linked incident records;

a parent/common-cause record;

another structure supported by local/UAIF implementation.

The objective is traceability, not forcing every event into one record.

PART VII — COMMUNICATIONS, LEGAL AND REGULATORY ROUTING

21. Technical Routing Is Not Legal Determination

UAIF regulatory metadata and AI-IRF notification structures can help route a matter for review.

They do not decide:

whether an incident is legally reportable;

which regulator has jurisdiction;

whether a statutory threshold is met;

whether privilege applies;

whether negligence or liability exists;

whether a contractual notification is triggered.

Those decisions remain with authorized legal/compliance functions.

---

22. Internal Communications Discipline

Leadership communications should distinguish:

confirmed facts;

current operational impact;

containment status;

hypotheses;

unknowns;

decision requests;

next update time.

Avoid turning an early hypothesis into an organizational statement of fact.

---

23. Provider Communications

When escalating to a provider, request specific evidence where contractually and technically available:

incident/change reference;

affected service/model/version;

time window;

known impact;

provider-confirmed root cause, if established;

mitigations;

customer action required;

rollback/version options;

preservation of relevant provider evidence;

future change-notification commitments.

Do not represent the absence of provider evidence as proof that no provider-side issue occurred.

24A. Governance of Severity Calibration

UAIF severity weights referenced by the current technical source remain provisional pending sector validation. Organizations should therefore avoid hard-wiring legal, regulatory or safety decisions to the default score alone.

A practical enterprise implementation can maintain three distinct values:

UAIF technical severity output - calculated using the applicable UAIF profile/version.

Evidence confidence / trust assessment - a local statement of how well the underlying facts support the current classification.

Local escalation decision - the organization’s operational, sector, safety, contractual or legal routing decision.

When sector-validated UAIF profiles or weights become available, adoption should be controlled through normal change governance:

record profile/version and effective date;

test the new calibration against historical/sample incidents;

document differences from the prior profile;

obtain the appropriate risk/sector approval;

update automation and routing rules;

preserve the profile used for each historical incident;

avoid silently recalculating closed incidents unless the organization has a defined reason and audit trail.

This guide does not prescribe a universal confidence score.

PART VIII — RECOVERY AND RETURN TO SERVICE

24. Recovery Is More Than Restarting

Recovery should answer:

What condition are we recovering from?

What has changed?

What evidence supports the remediation?

What residual uncertainty remains?

What tests were performed?

What monitoring is increased?

What rollback remains available?

Who authorizes return to service?

What downstream parties need notice?

What conditions would trigger re-containment?

---

25. Recovery Evidence

Depending on incident type:

patched/configured version record;

credential rotation evidence;

tool-permission changes;

retrieval-source remediation;

clean index/rebuild evidence;

evaluation/test results;

adversarial test results;

provider remediation notice;

model/version pinning;

safety validation;

simulation/HIL evidence;

monitoring changes;

updated runbook;

approval record.

Evidence of remediation does not automatically establish long-term effectiveness.

---

26. Physical-AI Return to Service

For physical-AI systems, recovery authority may sit outside the SOC.

The decision may require:

engineering owner;

safety function;

operations;

manufacturer/vendor;

clinical/medical-device authority;

automotive/aviation/industrial safety authority;

regulatory approval where applicable.

AI-IRF/UAIF records can support the evidence trail. They do not replace the applicable return-to-service regime.

PART IX — LEARNING AND GOVERNANCE FEEDBACK

27. Post-Incident Review

Ask:

What was observed?

What was initially believed?

What changed our understanding?

Which evidence was missing?

Which containment actions worked?

Which actions destroyed or degraded evidence?

Which dependencies were not known beforehand?

Did provider evidence arrive in time?

Were decision rights clear?

Were safety/legal/privacy routes engaged appropriately?

Did classification changes remain traceable?

What should change in GAISSF™ controls?

What should change in PAI-SF™ controls where applicable?

What should change in UAIF/AI-IRF implementation?

What remains unresolved?

27A. Closure With Unresolved Questions

Operational closure does not require manufactured certainty.

An incident may be closed while causal, provider-internal, attribution, or long-term-effectiveness questions remain unresolved, provided the organization records:

what is confirmed;

what remains unknown;

why further investigation is not currently justified or feasible;

residual risk and its accountable owner;

open corrective or monitoring actions;

any external dependency still awaiting evidence;

conditions that would reopen the incident;

closure authority and date.

Closure is an operational state, not proof of safety, absence of future harm, complete root-cause knowledge, or control effectiveness.

---

28. Metrics — Use With Care

Potential metrics:

time to detect;

time to establish incident ownership;

time to first containment;

time to preserve critical evidence;

time to identify affected systems;

time to provider escalation;

time to classification review;

time to safe recovery;

percentage of incidents with complete decision trail;

percentage with unresolved evidence gaps at closure;

repeat incident rate for the same mechanism.

Faster is not automatically better if evidence quality, safety, or decision quality deteriorates.

Metrics should be interpreted in context.

28A. Tabletop and Technical Exercise Programme

A field guide becomes useful only if teams can execute it under pressure.

Exercises should test decision-making and evidence flow, not merely whether a meeting was held.

Exercise types

Discussion tabletop - roles; - escalation; - legal/privacy/safety routing; - provider coordination; - recovery authority.

Technical evidence drill - collect prompt/context/output/tool-call evidence; - identify versions/configuration; - export identity/cloud/provider logs; - build timeline; - preserve provenance.

Containment drill - disable tool; - narrow credentials; - quarantine retrieval source; - roll back model/app version; - switch to manual/fallback mode.

Provider escalation drill - test support route; - request evidence; - evaluate provider response; - document unresolved boundary.

Physical-AI joint drill - safe state; - safety/security incident command; - telemetry preservation; - return-to-service authority.

What to measure

Avoid turning every drill into a speed competition. Record:

evidence gaps;

ownership gaps;

access/permission gaps;

unclear decision authority;

destructive steps that would erase evidence;

missing provider contact;

recovery criteria that cannot be demonstrated;

unknown shared dependencies;

documentation that did not match actual architecture.

Exercise closure

Every exercise should produce: - findings; - owners; - due dates; - revised playbooks; - revised evidence map; - revised authority matrix; - any needed GAISSF/PAI-SF control feedback.

PART X — WORKED INCIDENTS

29. Scenario 1 — Direct Prompt Injection Against an Enterprise Agent

Hypothetical scenario: A user supplies instructions intended to override an enterprise agent's policy and induce an unauthorized external action.

Initial observation: The agent attempted a tool call outside the expected task.

Unknowns: Whether the tool call was actually executed; whether authorization controls blocked it; whether the behaviour is reproducible; whether system prompt/configuration changed.

UAIF discipline: Record the observed action separately from the causal hypothesis prompt injection. Do not assume successful compromise because adversarial text was present.

Immediate AI-IRF actions: - preserve prompt/context/tool-call history; - restrict affected tool permissions if justified; - verify whether downstream actions completed; - preserve authorization logs; - identify other sessions exposed to the same prompt pattern; - test safely.

Recovery criteria: Unauthorized actions are bounded; control changes are tested; affected credentials/tools are reviewed; monitoring is increased.

Notably Absent: No conclusion that the underlying model is compromised or that prompt injection was the sole cause without supporting evidence.

---

30. Scenario 2 — Indirect Prompt Injection Through RAG Content

A malicious instruction is embedded in a document later retrieved by an internal assistant.

Preserve: source document, ingestion record, chunks, retrieval result, prompt/context, output, tool calls.

Contain: quarantine source; restrict affected connector/index; disable high-consequence tools if needed.

Classify: distinguish poisoned content, retrieval behaviour, model behaviour, application policy and authorization.

Investigate: Was the instruction treated as trusted context? Could other documents contain similar content? Was the document authorized for ingestion?

Recover: clean/rebuild affected index where justified; update content trust controls; test retrieval and tool-use boundaries.

Notably Absent: Document presence does not establish that the malicious content caused the downstream behaviour.

---

31. Scenario 3 — Sensitive Data Disclosure

An internal assistant returns restricted information to an unauthorized user.

Immediate priorities: - preserve query, retrieval context and output; - identify the data source; - stop further disclosure where justified; - determine access-control path; - assess whether output was copied/exported; - engage Privacy/Legal where applicable.

Possible causes: retrieval ACL failure, application authorization error, cached context, cross-session contamination, model behaviour, data-source permissions.

Notably Absent: An incorrect output is not automatically a model-security failure; a privacy incident determination requires the applicable privacy process.

---

32. Scenario 4 — Autonomous Agent Performs an Unauthorized Action

An agent changes a production configuration without expected human approval.

Contain: disable autonomous action or affected tool; revoke/narrow credentials; verify downstream changes.

Preserve: agent plan, tool calls, identity events, approval workflow, configuration before/after, downstream effects.

UAIF: distinguish observed unauthorized change from why the approval gate failed.

AI-IRF: investigate and recover while classification continues.

Notably Absent: Human-in-the-loop presence in design documentation does not prove meaningful human authorization operated in this event.

---

33. Scenario 5 — Provider Model Behaviour Changes After an Update

Several applications begin producing materially different outputs after a provider update.

Confirmed: behaviour changed within a known period; provider states an update occurred.

Unknown: whether update caused the behaviour; what internal changes occurred; whether all applications share the mechanism.

Response: - preserve enterprise-observed before/after evidence; - identify affected applications; - run bounded local evaluations; - consider version pin/fallback where available; - request provider evidence; - assess business/safety consequences.

Notably Absent: Temporal alignment and a provider update notice do not establish causation.

---

34. Scenario 6 — Model Supply-Chain Compromise

A model artifact or AI component is suspected of unauthorized modification.

Priorities: stop propagation; preserve artifact/hash/provenance; identify deployments; isolate build/release path; review signing and repository access; determine downstream use.

UAIF: distinguish suspected compromise from confirmed malicious modification.

AI-IRF: coordinate containment across model registry, CI/CD, runtime and affected products.

Notably Absent: A hash mismatch may establish artifact difference; it does not by itself establish malicious intent.

35. Scenario 7 — Clinical Decision-Support Error

An AI system produces an incorrect clinical recommendation; a clinician reviews it before treatment.

Preserve: input, output, version, clinical workflow, human review record, relevant integration data.

Safety: apply the authorized healthcare, clinical-safety or medical-device safety process first where applicable. If the AI is decision support only and has no physical command path, do not assume PAI-SF scope solely because the setting is healthcare.

UAIF: separate incorrect output from realized patient harm.

AI-IRF: contain or suspend affected function where justified; investigate data/model/integration causes.

Notably Absent: No conclusion on clinical negligence, medical-device reportability, patient harm, causality, or adequacy of human oversight.

36. Scenario 8 — Physical-AI Malfunction

An autonomous inspection robot deviates from its expected path and triggers an emergency stop.

Safety action: maintain safe state.

Preserve: sensor data, actuator commands, safety-controller records, model/software versions, environment, maintenance state, human observations.

Hypotheses: perception, localization, model/configuration, sensor, environment, mechanical, control system.

Recovery: do not return to service solely because software was restarted or the event cannot be reproduced. Return-to-service authority must remain with the applicable engineering/safety/operations function and, where relevant, the product or sector authorization process.

Notably Absent: Presence of AI does not establish AI causality; this guide does not establish functional safety.

37. Scenario 9 — AI-Enabled Fraud / Social Engineering

An AI-enabled workflow is suspected of generating or amplifying fraudulent communications.

Preserve: content, account/session data, generation path where known, delivery channel, authentication events, downstream transactions.

Response: contain compromised accounts/channels; coordinate fraud and security teams; preserve transaction evidence.

Classification: distinguish AI involvement from conventional account compromise.

Notably Absent: Sophisticated or synthetic-looking content does not establish that AI was used.

38. Scenario 10 — Cascading Shared-Dependency Incident

Five applications using the same foundation-model endpoint exhibit abnormal behaviour.

Priorities: identify common dependencies; preserve application-specific evidence; determine whether blast radius is shared; coordinate one enterprise response without collapsing distinct impacts.

Record design: parent/common-cause relationship or linked records may be appropriate.

Notably Absent: Shared dependency does not prove shared cause until evidence supports it.

38A. Incident Pattern Quick-Reference Matrix

This matrix is a response orientation aid. It does not replace system-specific runbooks, UAIF™ controlled classification, AI-IRF™ controls, or sector-specific response requirements.

Pattern	Preserve first	Immediate containment focus	UAIF™ questions	AI-IRF™ response focus	Escalation	Recovery evidence	Notably Absent
Direct prompt injection / agent misuse	Prompt, context, tool calls, authorization, downstream actions	Restrict high-consequence tools/credentials	What action occurred? Was adversarial input causal? Was authorization bypassed?	Bound actions, identities, tools and affected sessions	SOC, app/model owner, IAM; Legal where impact warrants	Tool/permission tests, downstream verification, monitoring	Adversarial text does not prove successful compromise
Indirect prompt injection / RAG poisoning	Source document, ingestion, chunks, retrieval context, output, tool calls	Quarantine source/index/connector	Was malicious content retrieved? Did it influence behaviour?	Bound poisoned content and affected retrieval paths	SOC, data/RAG owner, app owner	Clean/rebuilt index where justified; retrieval/tool boundary tests	Document presence does not establish causation
Sensitive-data disclosure	Query, retrieval context, output, ACL/identity events, export evidence	Stop further disclosure; restrict affected path	What data was exposed, to whom, and through which mechanism?	Bound recipients, data sources and propagation	Privacy, Legal, data owner, SOC	Access-control/ remediation tests; distribution assessment	Incorrect output alone does not establish privacy reportability
Autonomous unauthorized action	Agent plan/state, tool calls, approvals, identity, before/after state	Stop actions; narrow credentials/tools; verify downstream effects	What executed? What remained attempted? Did approval operate?	Bound chained actions and rollback	SOC, platform, IAM, business owner	Rollback/repair evidence; approval-gate validation	Documented HITL does not prove meaningful authorization
Provider/model behaviour change	Before/after outputs, exposed version, configuration, provider notices	Pin/fallback/restrict use where available	Is change confirmed? Is provider causality established?	Bound dependent applications and business impact	Vendor owner, SOC, model owner, risk functions	Local evals, provider remediation evidence, fallback validation	Temporal alignment does not establish provider causality
Model/software supply chain	Artifact, hashes, provenance, signing, registry and CI/CD evidence	Stop propagation; isolate release path	Difference, vulnerability, or confirmed compromise?	Identify deployments and downstream exposure	SOC, DevSecOps, supply-chain/vendor owner	Trusted rebuild/redeployment and provenance evidence	Artifact difference does not establish malicious intent
Clinical / regulated decision support	Input/output/ version, workflow, human review, integration evidence	Apply applicable safety process; suspend affected function where justified	Incorrect output vs realized harm; mechanism known?	Preserve evidence and coordinate technical containment	Clinical/safety authority, Legal/Privacy, SOC	Applicable validation plus technical remediation evidence	Guide does not determine clinical negligence or reportability
Physical AI	Sensor/actuator/ safety-controller telemetry, versions, environment	Engineered safe state; isolate actuation/autonomy	Was AI causal, contributory, or incidental?	Protect people first; bound cyber/physical effects	Safety/ engineering/ operations plus SOC	Specialist validation, simulation/HIL where applicable, authorization	AI presence does not establish AI causality or functional safety
AI-enabled fraud/social engineering	Content, account/session, authentication, delivery, transaction evidence	Contain accounts/channels/transactions	Was AI actually involved? What harm occurred?	Coordinate cyber and fraud response	Fraud, SOC, Legal/Compliance as applicable	Account/channel remediation and transaction review	Synthetic-looking content does not prove AI use
Cascading shared dependency	Per-system evidence plus shared dependency/version /configuration	Bound common dependency without erasing system-specific impacts	One cause or correlated symptoms? One or linked records?	Coordinate enterprise-wide containment and reclassification	Incident command, platform/vendor owners, affected businesses	Dependency remediation plus per-system validation	Shared dependency does not prove shared cause

38B. How to Work a Scenario From Signal to Closure

The ten scenarios above are intentionally concise. The following pattern shows how a team can turn any one of them into a full operational record.

Step 1 - Capture the signal

Record the observation in the reporter's terms before translating it into a technical hypothesis.

Example: > "The agent submitted a payment that the user says they did not authorize."

Do not rewrite this immediately as: > “Prompt injection caused an unauthorized payment.”

The second sentence adds a causal conclusion that the signal does not support.

Step 2 - Identify the affected transaction / action

For agentic or decision systems, identify: - requested task; - actual action; - tool/API; - identity; - approval state; - downstream result; - reversibility.

Step 3 - Preserve the execution context

Where lawfully and technically available: - input; - system/configuration; - retrieved context; - model/provider/version; - agent plan/intermediate state; - tool calls; - authorization decisions; - timestamps; - output; - downstream state.

Step 4 - Establish the evidence states

Maintain at least: - confirmed facts; - working hypotheses; - provider representations; - unknowns; - specifically checked but not observed conditions.

These practitioner categories help investigation; do not invent UAIF enums.

Step 5 - Decide protective action

Ask: - What can happen if we do nothing? - What is the least destructive action that sufficiently reduces harm? - What evidence will the action change? - Who can authorize it? - Is there a safety/clinical/OT process that must govern?

Step 6 - Update classification as investigation develops

A causal hypothesis can strengthen, weaken or be replaced.

Examples: - suspected prompt injection → authorization defect; - suspected model issue → provider version change; - suspected AI malfunction → mechanical failure; - suspected isolated event → shared dependency incident.

Step 7 - Establish recovery criteria before restoring

Recovery should be defined before restoration begins: - what change is required; - what test demonstrates it; - what residual unknowns are acceptable; - what monitoring is required; - who authorizes.

Step 8 - Close without erasing uncertainty

The record should show: - final supported facts; - final supported causal conclusion, if any; - unresolved items; - realized harm; - residual risk; - corrective actions; - provider follow-up; - governance feedback.

38C. Common Incident-Response Anti-Patterns

Anti-pattern 1 - The instant root cause

The first plausible explanation becomes the official story.

Correction: preserve working-hypothesis status until evidence supports the causal conclusion.

Anti-pattern 2 - Contain first, preserve later

The team rebuilds, reindexes, wipes, rotates, or resets before capturing the state needed for investigation.

Correction: preserve first where safety and expected harm permit.

Anti-pattern 3 - “The vendor said so”

A provider support response is treated as independently verified root cause.

Correction: preserve provider representation separately from local evidence.

Anti-pattern 4 - The AI label swallows the cyber incident

A conventional credential, cloud, API, data, or endpoint compromise is described as an AI failure merely because an AI application is affected.

Correction: use the AI Involvement Test.

Anti-pattern 5 - Human-in-the-loop by diagram

Design documents show a human approval gate, so the incident record assumes meaningful human authorization occurred.

Correction: inspect the actual event: approval, timing, authority, rationale, override and system enforcement.

Anti-pattern 6 - Closure equals proof of safety

The ticket is closed and the system is therefore described as safe.

Correction: closure is operational; safety, effectiveness and sector authorization require their own evidence/authority.

Anti-pattern 7 - Severity equals legal reportability

A technical score or routing flag automatically triggers a legal statement.

Correction: use technical routing to engage the authorized legal/regulatory review.

Anti-pattern 8 - One shared provider means one incident

Multiple applications share a model provider, so all anomalies are collapsed into one root cause.

Correction: establish common evidence first; use linked records where that preserves system-specific facts.

Anti-pattern 9 - Playbook automation without escape conditions

Automated containment continues after context changes.

Correction: define stop, review, rollback and human-override conditions.

Anti-pattern 10 - Post-incident learning becomes generic training

The review produces “improve awareness” instead of specific control/evidence changes.

Correction: connect lessons to named owners, system changes, evidence gaps, playbook updates, GAISSF controls and PAI-SF controls where relevant.

38D. Enterprise SOC Mapping - NIST SP 800-61 Rev. 3

External reference note: NIST SP 800-61 Rev. 3 (April 2025) supersedes Rev. 2 and aligns incident response to NIST CSF 2.0. NIST describes **Govern, Identify and Protect** as broader preparation activities supporting incident response; **Detect, Respond and Recover** form the incident-response life cycle, with continuous improvement feeding the functions.

The mapping below is **orientation only**. It does not assert equivalence between AI-IRF™ and NIST SP 800-61r3.

AI-IRF™ controlled phase	Closest NIST SP 800-61r3 / CSF 2.0 relationship	Enterprise integration interpretation
Phase 0 - Continuous Discovery	Govern / Identify / Protect preparation; Detect also contributes new observations	Extend existing asset, dependency and exposure discovery to models, providers, RAG sources, agents, tools and physical-AI components
Phase 1 - Preparation	Govern / Identify / Protect preparation	Add AI-specific authority, evidence maps, provider routes, playbooks, fallback and safe-state planning to existing IR readiness
Phase 2 - Detection & Triage	Detect and Respond incident management/analysis	Keep existing alert intake and triage; add AI Involvement Test, UAIF-compatible evidence separation and AI-specific telemetry
Phase 3 - Machine-Speed Containment	Respond mitigation/incident management	Map AI-specific containment actions into existing SOAR/playbooks with explicit human/safety authority boundaries
Phase 4 - Eradication	Respond mitigation and transition toward recovery	Remove the supported causal condition while preserving evidence and avoiding symptom-only closure
Phase 5 - Recovery	Recover	Use existing recovery/change processes with AI-specific validation, model/configuration provenance, rollback and sector return-to-service gates
Phase 6 - Post-Incident & Regulatory	Recover plus Improvement across CSF functions	Feed lessons into governance, architecture, provider management, controls, playbooks and legally authorized notification processes

Adoption principle: COM-013 should normally operate as an AI-specific overlay to an existing SOC/CSIRT model, not as a replacement incident management system.

PART XI — ENTERPRISE INTEGRATION

39. SOC / SIEM / SOAR Integration Pattern

A practical implementation can retain existing enterprise systems:

Alert / report
→ existing ticket/case
→ AI-specific intake
→ UAIF-compatible classification fields
→ AI-IRF response tasks/playbook
→ evidence links
→ classification re-evaluation
→ recovery / closure
→ governance feedback

Possible implementation:

keep the organization’s existing incident ID as the operational case key;

store a UAIF-compatible structured record alongside or within the case;

map AI-IRF response actions to existing playbook tasks;

preserve evidence references rather than duplicating large artifacts;

trigger human review when severity/classification changes;

avoid automatic legal reporting based solely on technical routing fields.

---

40. Automation Boundaries

Machine-readable UAIF and AIIS artifacts can support:

validation;

controlled vocabularies;

structured playbooks;

GRC mapping;

case enrichment;

deterministic calculations defined by the schema;

routing to human review.

Automation cannot by itself establish:

truth of the evidence;

root cause;

legal reportability;

negligence;

adequacy of safety;

effectiveness of containment;

conformance or certification.

AIIS v1.0.0 explicitly reserves the optional conformance_test structure for a future release. It is not populated in the current control-record release, and records validate without it. This guide therefore makes no claim that AIIS v1.0.0 provides embedded automated conformance testing.

40A. Role-Specific Response Cards

Incident Commander

Ask: - What decision is needed now? - What evidence supports it? - What remains unknown? - Who has authority? - What is the next checkpoint? - What business/safety function must be present?

Do not: - allow hypothesis to become fact in leadership communications; - let technical teams make legal or safety-authorization decisions by default; - let the response proceed without an explicit owner.

SOC / CSIRT

Focus on: - preservation; - timeline; - scope; - identities; - telemetry; - containment; - case history.

Escalate when: - autonomous action has external consequence; - sensitive data is involved; - shared/provider dependency is suspected; - physical consequence is possible; - the event crosses a jurisdiction/sector threshold requiring specialist review.

AI / Model / Platform Owner

Provide: - model/provider/version; - configuration; - evaluation/change history; - system prompt / policy configuration where authorized; - tool and retrieval architecture; - rollback/fallback options; - provider contacts.

Do not treat: - evaluation success as incident closure evidence; - a provider model card as proof of the incident cause.

Data / RAG Owner

Provide: - source repositories; - permissions; - ingestion/change history; - index/version; - provenance; - chunk/retrieval records; - quarantine/rebuild options.

IAM / Cloud / Infrastructure

Provide: - authentication; - service principal; - token; - authorization; - network/API; - deployment; - cloud audit evidence.

Distinguish: - AI-specific mechanism from conventional account/infrastructure compromise.

Legal / Privacy / Compliance

Determine: - privilege/investigation handling; - legal characterization; - reporting/notification obligations; - data-protection impacts; - contractual duties.

Technical routing informs this work; it does not replace it.

Safety / OT / Clinical / Engineering Authority

Own or participate in: - safe state; - emergency procedure; - operating-envelope restriction; - specialist validation; - return-to-service authorization.

Cyber responders should not override the authorized safety process merely to preserve evidence.

Vendor / Procurement Owner

Coordinate: - provider escalation; - evidence requests; - contractual notice; - service/change records; - residual evidence gaps; - remediation commitments.

Feed systemic provider limitations into vendor-risk and GAISSF supply-chain control improvement.

PART XII — PRACTITIONER TOOLS

41. Incident Intake Worksheet

Case ID:
Detected/reported at:
Reporter/source:
System:
Model/provider/version:

Business process:
Initial observation:
Immediate safety concern:
Potential data/privacy concern:
Potential physical consequence:
Affected tools/integrations:
Third-party dependencies:
Evidence preserved:
Working hypotheses:
Unknowns:
Containment taken:
Decision authority:
Next review time:

---

42. Evidence Preservation Checklist

☐ Initial prompt/input

☐ System prompt/configuration where authorized

☐ Retrieved context

☐ Output

☐ Tool calls

☐ Agent intermediate state where available

☐ Model/provider/version

☐ Deployment/configuration

☐ Identity/authentication

☐ Authorization

☐ API/network

☐ Data/RAG ingestion

☐ Model release/change

☐ Human approval/override

☐ Provider notices

☐ Sensor telemetry

☐ Actuator commands

☐ Safety-controller logs

☐ Case/ticket timeline

☐ Evidence integrity/provenance record

---

43. Containment Decision Record

Observed condition:
Risk if no action is taken:
Proposed containment:
Potential business/safety impact of containment:
Evidence that may be lost:
Alternative actions considered:
Decision:
Authorized by:
Time:
Review/rollback condition:

---

44. Classification Change Record

Previous state:
New state:
New evidence:
What changed in interpretation:
What remains unknown:
Response change required:
Reviewer/authority:
Timestamp:

---

45. Recovery Gate

Before return to service:

☐ affected scope is bounded;

☐ remediation is implemented;

☐ evidence supports remediation;

☐ relevant tests completed;

☐ residual uncertainty recorded;

☐ monitoring updated;

☐ rollback available where applicable;

☐ provider dependency reviewed;

☐ safety authority approved where applicable;

☐ Legal/Privacy/Compliance conditions addressed where applicable;

☐ business owner accepts residual operational risk under local governance;

☐ recovery decision and authority recorded.

---

46. Post-Incident Review Record

Incident summary:
Confirmed cause(s):
Unresolved hypotheses:
Realized harm:
Potential harm not realized:
Detection gap:
Evidence gap:
Containment lesson:
Recovery lesson:
Provider/supply-chain lesson:
Human-oversight lesson:
GAISSF™ control feedback:
PAI-SF™ feedback where applicable:
UAIF implementation feedback:
AI-IRF implementation feedback:
Actions / owners / due dates:

PART XIII — NOTABLY ABSENT

47. What This Field Guide Does Not Establish

This guide does not, by itself:

establish that an observed AI event is a legally defined incident;

determine legal reportability or notification deadlines;

establish negligence, liability, malicious intent, or actor attribution;

establish root cause from temporal correlation;

establish that a provider caused an incident;

establish that a model is safe or unsafe;

establish that a control is effective merely because an artifact exists;

establish GAISSF™, UAIF™, AI-IRF™ or PAI-SF™ conformance;

authorize a certification, compliance, approved, endorsed, assessor, or certification-body claim;

replace existing SOC/CSIRT processes;

replace sector-specific incident-response requirements;

replace privacy-breach analysis;

replace medical-device, automotive, aviation, industrial, defence, critical-infrastructure, or other safety regimes;

establish that default UAIF severity weights are validated for a particular sector;

convert a technical regulatory-routing field into a legal determination;

guarantee interoperability with a specific SIEM, SOAR, GRC, ticketing, cloud, model-provider, or case-management product;

establish that automated playbook execution is appropriate without local validation;

establish that return to service is safe merely because the incident is closed.

---

48. Source, Traceability, and Conflict Rule

This guide is an implementation/adoption artifact. It does not modify the controlled UAIF™, AI-IRF™, AIIS, GAISSF™, or PAI-SF™ source instruments.

Controlled/source identifiers used by this edition

UAIF™ technical source: ODA3-2026-06-TCR-STD-002.

UAIF machine-readable companion: TCR-STD-003 / current UAIF schema package - verify current version, checksum, release state and profile before implementation.

AI-IRF™ technical source: ODA3-2026-06-TCR-HAI-001 (corrected technical source; 79 controls / 19 domains).

AIIS machine-readable companion: aiis-v1.0.schema.json, schema version 1.0.0; optional conformance_test structure reserved for a future release and not populated in v1.0.0.

Current public documentation portal: docs.oda3.org - use the portal inventory to confirm publication state and current packages before implementation.

External SOC integration reference: NIST SP 800-61 Rev. 3, April 2025.

48.1 Source hierarchy used for COM-013

When sources differ, apply the following order for this publication:

Current controlled normative / technical source for framework architecture, fields, control inventories, lifecycle phases, schemas, and conformance structures.

Current ODA3 public publication-state and GEL / claims-boundary material for market-facing maturity, licensing, and representation language.

Current sector-calibration packages for bounded sector interpretation.

WHP practitioner/adoption guides for explanatory workflow and cross-document consistency.

Historical drafts or superseded language only as research context; they do not override the controls above.

If this guide conflicts with a controlling source, the controlling source governs.

48.2 Core source register

Source	Role in COM-013	Status / boundary carried into this guide
ODA3-2026-06-TCR-STD-002 — UAIF™ v1.0 Technical Specification	UAIF architecture, L0 (Record Identity & Workflow, including record_type/profile/extensions) through L4 (Incident–Vulnerability Relationship), severity/routing boundary	Final Publication v1.0; not a legal reporting determination
TCR-STD-003 / UAIF machine-readable schema package	UAIF payload/profile implementation companion	Check current schema version, release state and package metadata before implementation
ODA3-2026-06-TCR-HAI-001 — corrected AI-IRF™ technical source	19-domain / 79-control inventory and canonical AI-adapted lifecycle	Current public comparison lists AI-IRF™ v1.0 as Final Publication; older certification language is not automatically carried forward
AIIS v1.0.0 machine-readable schema	AI-IRF control/playbook encoding and automation boundary	79 controls / 19 domains; optional conformance_test reserved for a future release and not populated in v1.0.0
ODA3-2026-08-WHP-COM-011 — When AI Breaks	Short response-orientation baseline	Informative practitioner guide
ODA3-2026-08-WHP-COM-012 — Classifying AI Incidents	UAIF classification, uncertainty, provider-evidence and pre-classification boundary	Final practitioner guide; underlying UAIF™ v1.0 is now Final Publication
PAI-SF™ Sector Calibration Packages 012-016	Physical-AI / regulated-sector boundary calibration	Sector guidance only; does not replace safety, approval, clinical, airworthiness, roadworthiness, machinery, OT, emergency-management or regulatory regimes

48.3 PAI-SF sector sources used

ODA3-2026-07-SCP-PAISF-012 — Robotics / Industrial Automation

ODA3-2026-07-SCP-PAISF-013 — Autonomous Vehicles / Mobility Systems

ODA3-2026-07-SCP-PAISF-014 — Drones / Uncrewed Systems

ODA3-2026-07-SCP-PAISF-015 — Medical Robotics

ODA3-2026-07-SCP-PAISF-016 — Smart Infrastructure / Physical AI

These packages are used to preserve sector boundaries and evidence interfaces. They are not treated as regulatory crosswalks or approvals.

48.4 Publication metadata rule

A FINAL status for **COM-013** would mean that this field guide completed its own ODA3 publication gates. It would not:

change UAIF™ framework maturity;

create an AI-IRF™ certification scheme;

populate the future AIIS conformance_test structure;

establish sector validation of severity weights;

create regulator recognition;

authorize external certified/compliant/approved/endorsed claims.

Before public release, re-check the current ODA3 publication catalog, GEL terms, and any framework-specific release-state notices.

---

49. Next Reading

Classifying AI Incidents: A Practical Guide to UAIF™ — ODA3-2026-08-WHP-COM-012

When AI Breaks: The ODA3 Guide to AI Incident Response — ODA3-2026-08-WHP-COM-011

The UAIF™ Incident Classification Handbook — planned practitioner handbook

Securing Physical AI: A Practitioner's Guide to PAI-SF™ — forthcoming in the WHP adoption/practitioner series

Building an AI Assurance Evidence Pack — forthcoming

APPENDIX A - SYSTEM-SPECIFIC AI INCIDENT READINESS CARD

Use one card per material AI system or service. This is an implementation worksheet, not a conformance form.

A.1 System identity

System / service:
Business owner:
Technical owner:
AI/model owner:
Provider:
Model / service / version:
Environment:
Criticality:
Primary users:
Primary business process:

A.2 AI capability profile

Check or describe:

☐ Generative text/content

☐ RAG / enterprise knowledge retrieval

☐ Agent / tool use

☐ Autonomous transaction/action

☐ Decision support

☐ Automated decision

☐ Computer vision / perception

☐ Physical actuation

☐ Third-party embedded AI

☐ Foundation-model dependency

☐ Safety/clinical consequence

☐ Critical-infrastructure consequence

Autonomy boundary:
Human approval / override:
Physical consequence potential:
Sensitive-data exposure:

A.3 Evidence locations

Evidence class	Source / location	Owner	Retention / availability limitation
Inputs / prompts			
System/configuration prompts			
Retrieved context			
Outputs			
Tool/API calls			
Agent state / plan			
Model/provider/version			
Identity / authorization			
Cloud / network			
Data / RAG ingestion			
Deployment / release			
Human approval / override			
Provider notices			
Safety / sensor / actuator			

A.4 Emergency action

Can AI functionality be disabled independently?
Can tool permissions be narrowed?
Can retrieval sources be quarantined?
Can the model/app version be rolled back?
Is manual fallback available?
Is a safe/degraded mode defined?
Who can authorize each action?
Which actions require two-person or safety approval?

A.5 Provider escalation

Security support route:
Incident escalation route:
Account / contract owner:
Change-notification mechanism:

Available independent assurance artifacts:
Known provider evidence limitations:

A.6 Recovery authority

Technical restoration authority:
Business restoration authority:
Safety/clinical/OT authority where applicable:
Legal/Privacy conditions where applicable:
Minimum recovery tests:
Rollback condition:

APPENDIX B - FIRST-HOUR OPERATIONS WORKSHEET

B.1 Minute 0-15

Initial observation:
Reporter / source:
Time detected:
Affected system:
Immediate safety concern:
Immediate business concern:
Sensitive-data concern:
Physical consequence concern:

Evidence captured before containment:
1.
2.
3.

Protective action taken:
Authority:
Evidence altered/lost by action:

B.2 Minute 15-30

Known facts:
1.
2.
3.

Working hypotheses:
1.
2.
3.

Provider representations:
1.
2.

Unknowns:
1.
2.
3.

AI involvement:
- [ ] affected asset - [ ] attack surface - [ ] causal mechanism - [ ] enabling mechanism - [ ] evidence source only - [ ] incidental presence - [ ] unresolved

Potential shared dependencies:

B.3 Minute 30-60

Initial blast radius:
Containment state:
UAIF classification status:
AI-IRF response actions active:
Provider escalation:
Legal/Privacy/Compliance routing:
Safety/OT/Clinical routing:
Next checkpoint:
Decision needed:
Decision owner:

B.4 First-hour leadership brief

What happened - confirmed only:

What is affected:

What we have done:

What may have happened - hypothesis:

What we still do not know:

Next material decision and time:

APPENDIX C - END-TO-END WORKED RECORD

C.1 Hypothetical incident

An enterprise procurement agent is authorized to gather quotations and draft a purchase recommendation. It can access a vendor catalog and an internal purchasing API, but production purchases require a human approval token.

A user reports that the agent created a purchase order without the expected approval step.

This example demonstrates record evolution. It is not a statement that every similar event should receive the same classification or response.

C.2 Initial signal

Observed: a purchase order exists in the purchasing system. The user states they did not approve it.

Immediately known: - order identifier; - timestamp; - agent identity; - purchasing API endpoint; - affected vendor/account; - user statement.

Not yet known: - whether the agent bypassed approval; - whether a valid approval token existed; - whether the purchasing platform misapplied policy; - whether prompt manipulation occurred; - whether another user or service account authorized the order.

C.3 First containment

The incident commander authorizes: - suspension of the affected agent's purchasing permission; - preservation of agent/tool/API logs; - verification of whether the order can be cancelled without financial loss.

The team does **not** delete the agent session or rebuild the application before preservation because there is no immediate safety reason requiring destructive action.

C.4 Evidence pack

Collected: - user request; - agent plan/state available from the platform; - tool-call request; - API response; - identity/token logs; - approval service logs; - purchasing-system order history; - agent/app/model version; - recent deployment changes.

Provider evidence requested: - service-side execution trace; - any known model/tool-routing change; - relevant provider incident reference.

C.5 Initial hypotheses

H1 - Agent privilege abuse / approval bypass.
Plausible because the action exceeded intended authority.

H2 - Approval-token validation defect.
Plausible because the API may have accepted a malformed/stale token.

H3 - Prompt injection.
Possible but not established.

H4 - Human/account action outside the agent.
Possible until identity records are reconciled.

C.6 Classification evolution

T0: signal under triage.
T1: managed incident because an unauthorized production purchase order is confirmed under local criteria.
T2: identity evidence shows the request used the agent service principal.
T3: approval-service logs show no valid human approval token.
T4: application logs reveal a deployment change that disabled an approval check for one execution path. Prompt injection is not observed in the preserved context.

The working causal view therefore changes from "possible agent privilege abuse or prompt manipulation" to "application authorization / approval-path defect."

C.7 Response evolution

Initial: - suspend purchasing permission; - cancel order where possible; - preserve evidence.

After T3: - review all orders created through the same path; - review affected service-principal permissions; - inspect deployment/change records.

After T4: - correct approval-path defect; - test every purchasing execution branch; - inspect whether other agents use the same component; - update release test/gate; - restore only after evidence-supported validation.

C.8 Recovery gate

Required: - approval check restored; - negative tests prove no purchase occurs without valid approval; - affected historical orders reviewed; - service-principal scope reviewed; - monitoring added

for purchase attempts without approval; - rollback available; - business owner authorizes limited restoration.

C.9 Closure

Confirmed cause: application authorization-path defect, if evidence and local review support that conclusion.

Not confirmed: prompt injection.

Realized consequence: one unauthorized purchase order; financial impact depends on cancellation/settlement facts.

Residual actions: release-control improvement, shared-component review, provider follow-up if applicable.

GAISSF feedback: authorization, agent autonomy, change control, evidence and monitoring controls should be reassessed.

Notably Absent: this worked example does not establish legal liability, fraud, provider fault, financial materiality or framework conformance.

APPENDIX D - PLAYBOOK READINESS REVIEW

For each production playbook, review:

D.1 Trigger quality

☐ Trigger condition is defined.

☐ Trigger source is known.

☐ False-positive handling exists.

☐ Trigger does not imply root cause automatically.

☐ Shared-dependency impact is considered.

D.2 Evidence preservation

☐ Volatile evidence is identified.

☐ Preservation precedes destructive action where feasible.

☐ Evidence source/owner is known.

☐ Provider evidence route is known.

☐ Physical/safety telemetry is included where relevant.

D.3 Authority

☐ Incident commander authority defined.

☐ Credential revocation authority defined.

☐ Production rollback authority defined.

☐ Legal/Privacy routing defined.

☐ Safety/clinical/OT authority defined where applicable.

☐ Dual-control conditions defined where required.

D.4 Branch logic

☐ Low/medium/high consequence branches are distinguishable.

☐ Human review branch exists.

☐ Stop/rollback condition exists.

☐ Automation cannot run indefinitely without review.

☐ Classification change can alter the playbook branch.

D.5 Recovery

☐ Remediation evidence is defined.

☐ Validation/test evidence is defined.

☐ Monitoring after recovery is defined.

☐ Rollback is defined.

☐ Recovery authority is defined.

☐ Physical-sector approval remains outside the playbook where applicable.

D.6 Learning

☐ Incident findings feed control improvement.

☐ Playbook changes are version-controlled.

☐ Exercise findings create owned actions.

☐ Repeated evidence gaps are tracked.

☐ Provider limitations feed vendor-risk review.

APPENDIX E - PROVIDER ESCALATION EVIDENCE REQUEST

Use as a request framework, subject to contract and local process.

Customer incident reference:
Provider support / incident reference:
Affected service/model:
Observed time window:
Customer-observed symptoms:
Affected applications/processes:

Request, where available:

Confirm whether the provider identifies a service/model/change event during the stated period.

Identify affected service/model/version or routing layer.

Provide provider-confirmed incident/change timeline.

Provide known impact and customer populations if available.

State whether root cause is established, still under investigation, or not disclosed.

Provide mitigations already applied.

Identify customer-side action required.

Identify rollback, pinning, regional routing or alternative-version options.

Confirm preservation of relevant provider evidence.

Provide future update/notification reference where appropriate.

Provider evidence received:
Provider statements not independently verified:
Customer evidence that corroborates provider statement:
Material unresolved provider facts:
Effect on recovery/closure:
Residual uncertainty owner:

APPENDIX F - TABLETOP EXERCISE OBSERVER SHEET

F.1 Scenario

System:
AI capability:
Incident pattern:
Sector:
Physical/safety consequence:
Third-party dependency:

F.2 Observer questions

Did the team:

identify the system owner quickly?

preserve the initial observation?

distinguish facts from hypotheses?

perform the AI Involvement Test?

identify volatile evidence?

recognize provider dependency?

choose a proportionate containment action?

document authority?

recognize evidence destruction risk?

engage Legal/Privacy/Safety at the right point?

maintain classification while response evolved?

establish recovery evidence?

define who could authorize return to service?

preserve unresolved questions?

create corrective actions?

F.3 Friction log

Friction	Operational consequence	Owner	Due date
Missing log			
Unknown owner			
Missing provider route			
Unclear containment authority			
Destructive action risk			
Missing recovery test			

F.4 Exercise outcome

What worked:
What failed:
Evidence gap:
Authority gap:
Architecture gap:
Provider gap:
Sector/safety gap:
Playbook changes:
GAISSF / PAI-SF feedback:

APPENDIX G - RETURN-TO-SERVICE PROMPTS BY CONTEXT

G.1 SaaS / API AI

Is the provider issue confirmed resolved?

Has local behavior been re-evaluated?

Are fallback/version options known?

Are affected integrations reviewed?

Is residual provider uncertainty accepted?

G.2 RAG / enterprise knowledge

Are poisoned/untrusted sources removed or quarantined?

Is provenance re-established?

Was the index rebuilt only where justified?

Were retrieval and authorization boundaries tested?

Are previously affected outputs/users reviewed?

G.3 Agentic systems

Are tool permissions correct?

Are approval gates demonstrably enforced?

Are credentials rotated/re-scoped where needed?

Are downstream actions reconciled?

Is autonomous scope bounded and monitored?

G.4 Medical / clinical systems

Has the authorized clinical/device safety process approved restoration?

Has technical remediation been validated?

Are patient/clinical consequences reviewed under the applicable process?

Is operator/clinician override functional?

Are product/regulatory conditions satisfied where applicable?

G.5 Automotive / mobile physical AI

Is the ODD / operating envelope still valid?

Are sensor/perception and actuation paths validated?

Is fallback/takeover behavior validated?

Are safety/vehicle engineering authorities satisfied?

Is fleet/update propagation controlled?

G.6 Aviation / drones

Is the approved contingency/recovery process complete?

Are navigation, communications and command paths validated?

Are payload/autonomy boundaries restored?

Is airworthiness/operational authority satisfied where applicable?

Is mission authorization separate from the cyber incident closure?

G.7 Manufacturing / OT / robotics

Are interlocks and safe states validated?

Is manual/local control available?

Are PLC/controller and AI boundaries tested?

Are machinery/functional-safety authorities satisfied?

Is staged production restoration appropriate?

G.8 Critical infrastructure / smart infrastructure

Is service continuity protected?

Are public-safety/emergency-management conditions satisfied?

Are distributed controllers/gateways validated?

Is local/manual control available?

Are physical command and degraded-state records reviewed?

APPENDIX H - PRACTITIONER CONCEPT TO UAIF IMPLEMENTATION REFERENCE

This appendix is intentionally bounded. It helps implementers connect common COM-013 practitioner language to UAIF concepts without inventing new controlled values. The current UAIF technical specification and machine-readable schema remain authoritative.

Practitioner concept in COM-013	UAIF implementation treatment	Boundary
Signal / event under triage	Keep in existing intake/case workflow until evidence supports creation/population of the applicable UAIF record	“Signal” is not introduced as a new UAIF enum
Suspected AI incident	Local workflow state; begin preserving UAIF-relevant evidence and identifiers	Does not itself establish record_type=Incident
Managed incident	Local operational state once organization declares/manages the event under its IR process	Local state, not a UAIF controlled value
Observed manifestation / confirmed observation	Record in the applicable manifestation/description and evidence structures defined by the current UAIF schema/profile	Observation should not be rewritten as causality
Working causal hypothesis	Use the applicable causal-hypothesis structure/field supported by the current UAIF schema/profile	Preserve hypothesis status until supported
Incident vs vulnerability determination	Use the controlled record_type field in L0 — Record Identity & Workflow when evidence supports it	Do not force an early record_type merely to satisfy workflow tooling
Linkage between a live incident and a related vulnerability/exposure	Use L4 — Incident–Vulnerability Relationship (e.g. linked_vulnerability_id, exposure_duration_hours)	L4 records relationship/exposure context; it does not itself set record_type
Base profile and additive extensions	Confirm the record’s base profile (Core, Enterprise, or Regulatory) and any adopted extensions (e.g. SOC_SIEM, AI_Security) before treating a field as required	Do not assume Core-profile field requirements apply under Enterprise/Regulatory or extension-scoped records
Provider representation	Preserve as externally supplied evidence with source/provenance differentiation	Provider statement is not automatically independently verified fact

Practitioner concept in COM-013	UAIF implementation treatment	Boundary
Known unknown / unresolved question	Preserve in the record/case using schema-supported notes/evidence plus local workflow metadata as appropriate	Do not invent a controlled UAIF value if the schema has none
Classification change	Update the applicable UAIF fields and preserve the audit/change history	Record new evidence, author/authority and time
Technical severity	Use the applicable UAIF severity profile/version	Current weights remain subject to the framework's stated maturity/calibration limits
Regulatory routing indicator	Treat as technical routing information for authorized legal/regulatory review	Not a legal-reportability determination

Minimum Viable UAIF-Linked Case

At the earliest defensible point, an enterprise case should be able to preserve:

enterprise incident/case identifier;

detection/observation timestamp;

affected system/service identity;

current local workflow state;

observed manifestation;

evidence references;

current causal hypothesis, if any;

incident/vulnerability determination when supported;

current severity/profile information when used;

unresolved questions;

classification-change history.

Before implementation, verify the exact field names, required/optional status, schema version, checksum/release state and applicable profile in the current UAIF machine-readable package.

USE BY / GOVERNANCE OF USE

Who should read this guide

Primary operators - SOC/CSIRT leaders and responders; - AI security teams; - security architects; - AI/model/platform engineering; - incident commanders.

Decision participants - AI governance; - privacy; - legal/compliance; - risk; - vendor/procurement; - business/system owners; - safety/OT/clinical engineering where applicable.

Who should act on it

Operational actions should be executed only by roles authorized under the organization's incident, change, safety, clinical, OT, legal and business processes.

The presence of an action in this guide does not grant authority to execute it.

Who should govern its application

Organizations should assign ownership for: - local adoption/profile decisions; - evidence/retention design; - playbook approval; - severity calibration; - automation; - provider integration; - sector/safety boundaries; - review after framework/schema updates.

This guide is best governed as an overlay to the organization’s existing incident-response and AI-governance system.

Publication Boundary

FINAL publication.

The final solution-architecture review has been incorporated and the resulting 55-page production artifact has passed render and visual QA.

FINAL status freezes this practitioner guide only. It does **not**: - upgrade UAIF™ or AI-IRF™ framework maturity; - create certification or accreditation mechanics; - populate AIIS automated conformance_test capability; - establish sector validation of provisional severity weights; - establish legal reportability; - replace functional-safety, clinical, product-safety, OT, defence or other controlling sector authority; - create regulator recognition or endorsement.

Implementers should verify the current ODA3 publication catalog, GEL terms, machine-readable schema version/checksum/release state, and applicable sector requirements at the time of implementation.