ODA3 INSTITUTE

ADOPTION GUIDE

Classifying AI Incidents

A Practical Guide to UAIF™

Doc ID: ODA3-2026-08-WHP-COM-012
Classification: Public – GEL v1.0
UAIF™ publication status: Final Publication v1.0

Revision: FINAL R5

Executive Summary

AI incident classification is not simply the act of attaching a label after something goes wrong. An organisation may observe anomalous or harmful system behaviour before it knows the cause, the full consequence, whether the condition is an incident or a vulnerability, or whether a regulatory reporting obligation exists.

UAIF™ provides a machine-readable incident interchange, classification, severity-scoring, and conformance architecture for AI incidents. Its seven-layer architecture separates identity and workflow, causality, severity, acute and chronic harm, incident-versus-vulnerability status, generative and agentic characteristics, and regulatory or cross-sector metadata.

> **KEY PRINCIPLE**
>
> Record what the evidence supports. Preserve what remains uncertain. Revise the classification when the evidence changes.

This guide explains how practitioners can apply that discipline without treating a provisional classification, severity score, causal hypothesis, or technical routing flag as a legal conclusion.

> **METHODOLOGY NOTE**
>
> This is an adoption and practitioner-orientation guide derived from ODA3 Institute's UAIF™ v1.0 technical architecture. It does not replace the controlled UAIF specification or machine-readable schema. Examples are hypothetical and do not represent client incidents, proprietary telemetry, deployment outcomes, or incident-frequency statistics. Where UAIF uses controlled terminology, this guide preserves that terminology. Where the guide adds explanatory workflow language, that language is informative rather than a new UAIF lifecycle or taxonomy.

Source and Traceability

This adoption guide is an informative WHP publication. It does not create or modify UAIF normative requirements.

- Controlling technical source: ODA3-2026-06-TCR-STD-002 — UAIF™ v1.0 Technical Specification — Final Publication v1.0.
- Machine-readable companion: TCR-STD-003 — UAIF™ schema specification / schema package.
- Current public status: UAIF™ v1.0 is a Final Publication (TCR-STD-002, TCR-STD-003). Final status means the normative architecture, controlled vocabulary and conformance model are frozen under ODA3 Institute change control — it does not by itself mean every sector calibration, legal interpretation or empirical reliability claim has been validated in production.
- Operational reliance: before implementation, confirm the current schema version, publication state, checksum/release metadata, GEL terms, and any open validation gates in the ODA3 Publication Catalog and JSON Schema Center.

COM-012 reaching a final publication state as a practitioner guide is a separate control from UAIF™'s own publication status, and the two happen to align in this revision: both are now final. Document publication status and framework maturity remain separate controls that must not be conflated going forward — a future COM-012 revision could still be FINAL while UAIF™ moves to a later version, or vice versa. Framework maturity is stated in full above and in the Publication Boundary at the end of this guide; where it is referenced again in the body, it is noted briefly rather than restated in full.

> **HOW TO READ THIS GUIDE**
>
> All field names, values and examples shown in this guide are illustrative. The normative authority for required, conditional and prohibited fields is the TCR-STD-003 JSON Schema for your organisation's selected conformance profile, together with the current release state published in the ODA3 Publication Catalog and JSON Schema Center. This guide never overrides the schema; where the two differ, the schema governs.

1. What UAIF™ Is — and Is Not

UAIF™ is designed as an incident interchange and classification framework for AI systems. It provides structured severity scoring, conformance architecture, and machine-readable incident information.

It is not:

- a universal AI governance framework;
- an operational containment-and-recovery framework;
- a safety certification scheme;
- a legal risk-assessment methodology;
- a legal determination of regulatory applicability or reportability;
- evidence that an identified root cause, actor, or intent has been proven merely because a field can record it.

Within the GAISSF Ecosystem, the functional distinction matters:

Framework	Primary role during incident work
GAISSF™	Governance, safety/security controls, assurance context and lessons feeding back into control improvement
UAIF™	Incident information, classification, severity and interchange
AI-IRF™	Operational preparation, detection, classification support, containment, recovery, notification and learning
PAI-SF™	More detailed physical-AI security controls where physical/cyber-physical characteristics place them in scope

UAIF classification and AI-IRF response are complementary. Response need not wait for every classification question to be resolved.

2. The UAIF™ Seven-Layer Architecture

UAIF structures an incident record across seven layers.

Layer	Name	What it makes visible
L0	Record Identity & Workflow	Identity, classification (record_type), conformance-profile declaration, deduplication, audit trail and workflow state
L1	Causal Separation Chain	Separation of observed manifestation from causal analysis
L2	Severity Scoring	Analytical and presentation severity with a machine-auditable rationale
L3	Acute vs Chronic Harm	Immediate harm and longer-term or cumulative degradation
L4	Incident–Vulnerability Relationship	Links a classified Incident to an associated Vulnerability and captures exposure context; not the classification decision itself
L5	Generative, Agentic & Control Plane	AI-specific characteristics including RAG leakage, hallucination, agent escalation and control-plane issues
L6	Regulatory & Cross-Sector Metadata	Sector, jurisdiction and technical regulatory-routing information

A prior L7 control-plane extension was merged into L5 in UAIF™ v1.0. Practitioners should not reconstruct L7 as a separate conformance layer.

Why separation matters

A single incident can simultaneously involve:

- a confirmed observed output;
- an unconfirmed causal hypothesis;
- realized privacy harm;
- an unresolved question about wider blast radius;
- an AI-specific attack or failure mechanism;
- a technical routing flag requiring legal review.

Compressing these into one prose label destroys useful uncertainty. The layered model keeps different kinds of information separate enough to be revised independently.

3. Start With the Record Type

record_type is an L0 field — part of the record's identity and workflow declaration alongside profile and the other L0 fields — and it explicitly distinguishes Incident from Vulnerability: dynamic occurrences versus static exposures. In the controlling UAIF™ technical specification (TCR-STD-002), Incident is used in a normative conditional rule governing which further fields a record requires. L4 is a separate layer (Incident–Vulnerability Relationship) that links an already-classified Incident to an associated Vulnerability once both exist; it does not perform the classification itself. The specification does not establish Observation as a controlled record_type value, so this guide does not introduce it as one.

That distinction is operationally important.

An attempted prompt injection that is blocked and produces no realized incident consequence should not automatically be described as a successful incident. A discovered unsafe configuration may instead be handled as a vulnerability or retained in the organisation's existing investigation workflow until the UAIF record type is supported by the available evidence.

> **OBSERVATION IS NOT A UAIF record_type VALUE**
>
> Practitioners will naturally speak about an "observation" while an event is still being investigated. In this guide, observation is an investigative concept, not an additional UAIF record_type. If the evidence does not yet support classifying the record as Incident or Vulnerability, keep the matter in the organisation's existing intake, triage, ticket or investigation workflow; preserve the observed facts, evidence and uncertainty; and set record_type only when the available evidence supports Incident or Vulnerability specifically. Do not invent an Observation enum value to make the record look complete sooner.

Classification discipline

Before forcing a record into an Incident state, ask:

1. What actually occurred?
2. What evidence demonstrates that occurrence?
3. Is there a realized dynamic occurrence, a static exposure, or insufficient evidence yet to assign the UAIF record type?
4. What remains unknown?
5. Has any harm actually been realized?
6. What response is justified even if incident status remains unresolved?

> **KEY PRINCIPLE**
>
> Uncertainty is recordable information. It is not a defect that must be hidden by choosing a stronger label.

> **RELATIONSHIP TO FIRST-RESPONSE CHECKLISTS**
>
> WHP-COM-011's 15-minute first-response checklist instructs responders to “Start the UAIF™-structured record” at the 10–15 minute mark. In this guide's terms, that means beginning to collect and stage UAIF-mappable information — L0 identity fields other than record_type itself, and the directly observed L1 facts — not yet setting record_type. Early triage remains a local SOC/case-management observation. record_type is set to Incident or Vulnerability only once the available evidence supports that specific determination, exactly

> as described earlier in this section. This reconciliation reflects WHP-COM-011 FINAL R2; if COM-011 is revised after this guide's publication, verify the two documents still agree before relying on this callout.

Practitioner classification sequence

The following is an informative decision sequence for using this guide. It is not a new UAIF lifecycle and does not replace the controlled specification.

Practitioner question	UAIF layer(s)	Recording discipline
What record are we maintaining, and how is its history preserved?	L0 — Record Identity & Workflow	Preserve identity, classification, deduplication, workflow state and audit trail.
Does the evidence support a dynamic incident or a static vulnerability?	L0 — record_type field	Do not force record_type = Incident merely to keep investigation moving.
What was directly observed, and what is only a causal explanation?	L1 — Causal Separation Chain	Separate manifestation, hypothesis, and unresolved cause.
What harm is evidenced, and is it immediate or cumulative?	L3 — Acute vs Chronic Harm	Record realized harm separately from potential or still-unverified consequence.
What severity does the UAIF method produce from the supported inputs?	L2 — Severity Scoring	Preserve the rationale and the provisional sector-validation limitation.
Are generative, agentic, retrieval or control-plane characteristics material?	L5 — Generative, Agentic & Control Plane	Use controlled AI-specific terms only where the evidence supports them.
Does the record need sector or jurisdiction routing?	L6 — Regulatory & Cross-Sector Metadata	Treat routing as technical assistance, not a legal reporting conclusion.

These questions are not required to be answered once, in this order, or before response begins. New evidence can change several layers independently.

A starting point for the record_type determination

The following is a simplified aid for triage, not a substitute for the classification discipline questions above or the controlled specification.

If the evidence shows…	Consider recording as…
A dynamic occurrence with a realized consequence	record_type = Incident, subject to the classification discipline questions above
A static exposure with no realized occurrence	record_type = Vulnerability, subject to the same discipline
Insufficient evidence yet to distinguish the two	Retained in local intake/triage workflow — not yet assigned a UAIF record_type

A dynamic event that was blocked or produced no realized harm is not automatically a Vulnerability merely because the immediate consequence was zero — the occurrence itself may still be the relevant fact. Where the evidence supports neither Incident nor Vulnerability under the classification discipline questions above, the correct action is to retain the matter in local triage, not to select whichever row of this table looks closest. Apply the classification discipline questions rather than this table alone; where the specification's own criteria differ from this simplification, the specification governs.

A practical starting record

The following is a representative starting point, not a replacement for the mandatory fields of the selected UAIF conformance profile.

L0 representative fields verified in the UAIF technical specification include:

- incident_uuid
- record_type
- profile
- extensions

- reporting_timestamp
- reporter_confidence_level
- remediation_status
- deduplication_hash

> **PROFILE AND EXTENSIONS**
>
> Every UAIF record declares a base profile — Core, Enterprise, or Regulatory — which determines which further fields the record requires; Enterprise extends Core with additional harm-characterization fields, and Regulatory extends Enterprise with additional regulatory metadata. A record may separately declare optional extensions (for example, fields that map to external security-tooling references or to generative/agentic AI characteristics); extensions are additive and never reduce what the base profile already requires. This guide does not restate the full conformance rules — treat profile and extensions as structural context for why a given record may require more or fewer fields than another, and confirm the applicable rules in the controlling specification.

L1 representative causal-analysis fields include:

- root_cause_primary_domain
- root_cause_specific_type
- root_cause_confidence
- exploit_path_description

A responder should populate only what the evidence and applicable schema support. A field's existence does not justify inventing a value to make the record look complete. The controlled profile/schema remains the authority for required, conditional and prohibited fields.

4. Separate Manifestation From Cause

UAIF L1 is the Causal Separation Chain.

This is one of the most important disciplines in AI incident classification because temporal sequence can easily be mistaken for causality.

For example:

- a model was updated on Monday;
- anomalous outputs appeared on Tuesday;
- the timing is confirmed;
- the update caused the outputs is still a hypothesis until supported.

UAIF provides fields for root-cause domain, specific type, root-cause confidence and exploit-path description. Those fields create a place to record causal analysis; they do not remove the evidentiary burden required to support it.

Controlled root_cause_specific_type values include terms such as Prompt_Injection, Model_Poisoning, Data_Poisoning, RAG_Leakage, Hallucination, Agent_Escalation, Supply_Chain_Compromise, System_Failure, and Human_Error.

> **CONTROLLED VOCABULARIES ARE VERSIONED**
>
> Always verify the values above against the current TCR-STD-002 enum before populating a record. UAIF™ v1.0 is a Final Publication, meaning this controlled vocabulary is frozen under ODA3 Institute change control for v1.0; a future version could still extend or revise it, so confirm you are working from the current release before implementation.

Use a controlled term only when the evidence supports it. An investigator should not convert "possible prompt injection" into Prompt_Injection simply because that is the most plausible early explanation.

DOCUMENTING COMPETING HYPOTHESES

When multiple root causes are plausible, record each working hypothesis with its own confidence level and the specific evidence that would confirm or exclude it, rather than collapsing them into a single early guess. This keeps the causal record honest as evidence accumulates and makes it possible for a later reviewer to see what was actually ruled out and why.

5. Severity: Two Scores, One Explicit Limitation

UAIF L2 separates:

- severity_analytical_score — 0.0–10.0;
- severity_presentation_score — 0.0–10.0;
- severity_level — Level 1 through Level 5;
- severity_rationale — the machine-auditable explanation of the calculation.

WHY TWO SCORES

TCR-STD-002 describes severity_analytical_score as the primary analytical metric, calculated before impact-scale and presentation treatment are applied. severity_presentation_score is derived from that analytical result by applying UAIF's impact-scale factor (the larger of the population and blast-radius multipliers) and a temporal-confidence adjustment; severity_level is then mapped from the presentation score, not the analytical score. regulatory_trigger is likewise evaluated against the presentation score. The two scores can therefore differ, and where they do, the presentation score is the operative figure for banding, stakeholder communication and regulatory-routing comparison.

The normative presentation-score bands are:

Level	Label	Presentation score
1	Info	0.0 ≤ score < 2.0
2	Low	2.0 ≤ score < 4.0
3	Medium	4.0 ≤ score < 6.5
4	High	6.5 ≤ score < 8.5
5	Critical	8.5 ≤ score ≤ 10.0*

** Critical is the only closed-interval band (score = 10.0 is included); all other bands are half-open at the upper bound. This is intentional, not a transcription inconsistency — 10.0 is a valid, reachable score and must fall inside a band.*

UAIF uses an impact-dominant model with context and scale adjustments. Controlled realized-harm categories include physical, environmental, security-integrity, privacy, financial, psychological, property and reputational harm.

The limitation that must travel with the score

UAIF v1.0's default harm weights are provisional pending sector validation.

Therefore a severity score should not be represented as:

- an empirically universal measure of harm;
- a regulator's severity determination;
- a legal conclusion;
- proof of materiality;
- a substitute for sector-specific safety, clinical, financial, operational or regulatory assessment.

A numerical output does not eliminate uncertainty about the underlying evidence.

How a severity output would be communicated

ILLUSTRATION (NOT A CALCULATED EXAMPLE)

Consider a single unauthorized RAG disclosure that has been contained, with no confirmed financial impact. Under UAIF's normative calculation, the resulting score depends on the dominant harm category and its weight, any secondary harms and their contribution, the applicable context modifiers (for example, whether RAG_Leakage or another context applies), the impact scale (the larger of the population and blast-radius multipliers), and a temporal-confidence adjustment — not on the scenario description alone. A UAIF severity record should capture the supported realized harm, scope, confidence and relevant context, then let the authoritative scoring method in TCR-STD-002 §4 produce the analytical score, presentation score and severity level. This guide does not state a level, band or score for this scenario, because doing so without running the actual calculation would carry the same false precision as stating a number directly.

A template for reporting a severity score

Practitioners reporting a UAIF severity score to stakeholders can use language such as:

TEMPLATE

“UAIF Severity Level [X] ([Label]), Presentation Score [N.N]. Score calculated using UAIF v1.0 default harm weights (provisional pending sector validation). This score reflects analytical severity based on currently available evidence; it is not a regulatory determination or legal conclusion. Rationale: [brief summary].”

6. Acute and Chronic Harm

L3 separates Acute vs Chronic Harm.

This prevents classification from focusing only on immediately visible damage.

An acute event might include a single unauthorized disclosure or unsafe autonomous action. Chronic harm may emerge through repeated biased outcomes, accumulated degradation, persistent exposure, or a pattern that is individually subtle but collectively consequential.

For practitioners, the important question is not simply:

“What happened at the moment of detection?”

It is also:

“Is there evidence that the condition existed before detection, persisted after it, or produced cumulative effects?”

Do not infer chronic harm merely because a system operated for a long period. Establish the relevant evidence.

WHAT COUNTS AS CHRONIC-HARM EVIDENCE

Evidence supporting a chronic-harm finding may include: a sustained or analytically supported outcome disparity across time periods, repeated user complaints referencing similar outputs, audit findings showing persistent configuration drift, or longitudinal monitoring data demonstrating cumulative degradation. Statistical significance testing is one possible way to support such a finding but is not itself the criterion UAIF requires; the relevant question is whether the available evidence supports a persistent or cumulative pattern, by whatever analytically appropriate method fits the data. Absent evidence of this kind, record acute harm only and note the chronic-harm assessment as unresolved rather than assumed.

7. Generative and Agentic Incidents

L5 makes AI-specific incident characteristics explicit.

It covers generative, agentic and control-plane concerns including RAG leakage, hallucination, agent escalation, MCP, OAuth and policy-enforcement points.

This matters because a conventional security ticket may record “unauthorized transaction” while omitting whether:

- an agent selected the action autonomously;
- a tool invocation crossed an authorization boundary;
- a retrieval source introduced malicious instructions;
- an OAuth pivot expanded the agent's effective privileges;
- the model hallucinated an instruction;
- a policy-enforcement point failed or was bypassed.

UAIF does not require investigators to claim an AI-specific mechanism before it is established. It gives them a structured place to record the mechanism when evidence supports it.

8. Regulatory Routing Is Not a Legal Conclusion

L6 captures regulatory and cross-sector metadata.

UAIF includes regulatory_trigger, trigger_confidence, and human_review_recommended. The trigger is technical routing assistance.

It is not a legal determination that an incident is reportable.

UAIF v1.0's technical routing logic includes a default threshold of 6.5 against the presentation score for its configured routing conditions. That threshold must not be converted into a universal rule such as:

“Every UAIF score of 6.5 or above must be reported.”

> **NOTE ON THE SHARED 6.5 VALUE**
>
> The regulatory routing threshold and the Medium/High severity-band boundary in Section 5 currently default to the same value (6.5), but they are independently configurable mechanisms. A presentation score crossing into the High band does not itself mean the regulatory trigger fired, and a routing trigger firing does not itself mean severity is High. Treat them as two separate signals that happen to share a default number, not as one derived from the other. This threshold must not be hard-coded into an automatic reporting workflow.

Legal counsel and the applicable regulatory framework determine actual obligations, definitions and timelines.

The same discipline applies to sector metadata: structured routing can help direct a record to the right review process; it does not decide the law.

9. Classification Can Change

An incident record should evolve when material evidence changes.

Examples include:

- a suspected prompt injection later traced to a non-adversarial system failure;
- an apparent single-system event later found across multiple systems sharing a provider;
- a privacy concern later confirmed as a realized disclosure;
- an apparent physical-AI malfunction later traced to a non-AI mechanical failure;
- a condition under investigation later shown to be a vulnerability without realized exploitation.

The objective is not to preserve the first classification for consistency. The objective is to preserve the audit trail explaining why the classification changed.

For each material change, retain the prior state, the new evidence that prompted reconsideration, the resulting classification change, the decision time, and the responsible reviewer or authority in the organisation's case history and the UAIF fields permitted by the applicable profile/schema. This guide does not invent a separate audit_trail field where the controlled schema does not define one.

> **WHERE TO RECORD THE CHANGE**
>
> Where the applicable UAIF profile has no dedicated field for change history, record the prior state, new evidence, decision time and responsible reviewer in the organisation's own case-management or incident-tracking system alongside the UAIF payload — consistent with the local-case-record approach used

> elsewhere in the ODA3 adoption series. Do not leave the change undocumented merely because no UAIF field exists to hold it. At minimum, record: (1) the prior classification or state; (2) the new evidence that prompted reconsideration; (3) the resulting classification or state; (4) the decision timestamp; (5) the responsible reviewer or authority.

That distinction is central to defensible incident records.

10. Third-Party and Foundation-Model Incidents

Provider-dependent AI systems create an evidence boundary.

Separate:

Category	What it means
Enterprise-observed facts	What your systems, users, logs and controls actually demonstrate.
Provider representations	What the provider says occurred.
Provider-supplied evidence	Reports, notices, technical records or other artifacts supplied by the provider.
Independently reproduced observations	What your organisation can validate through its own testing or telemetry.
Provider-side unknowns	Training, model-update, infrastructure, or causal information that remains unavailable.

Do not fill provider-side gaps with certainty.

A provider notice may be relevant evidence. It is not automatically equivalent to independently established root cause.

A provider-visibility limitation is itself recordable uncertainty. It does not automatically require an incident to remain open indefinitely. Where organisational closure criteria permit closure despite unresolved provider-side facts, preserve the missing-evidence boundary, residual uncertainty and decision basis, and feed material provider limitations into procurement/vendor-risk review and the relevant GAISSF™ control improvement process.

A worked recording example

> **EXAMPLE RECORDING**
>
> Working hypothesis: a provider model update contributed to the behaviour change (confidence: Low). Provider visibility limitation: training-data composition and other provider-side implementation detail not disclosed. Enterprise-observed fact: the behaviour change correlates with the provider-confirmed update timestamp. Independent reproduction: not yet attempted. This is written as explanatory practitioner notation, not as UAIF schema field names or controlled values — it separates the provider representation, the enterprise-observed fact and the unresolved causal hypothesis without implying that these labels themselves appear in the schema. Where a record is actually being populated against the schema, use the field names and controlled values the applicable conformance profile defines, not the labels shown here.

11. Multi-System and Cascading Incidents

A shared model, retrieval service, agent platform, identity service or third-party provider can affect more than one AI system.

Classification should therefore examine:

- common dependencies;
- shared versions or update windows;
- affected systems;

- whether manifestations are identical or merely correlated;
- blast radius;
- whether one incident record or linked records better preserves the facts;
- whether the causal conclusion is common across systems or still unresolved.

A broad blast radius can affect severity. It does not prove a common root cause.

12. Physical-AI Events

Physical-AI incidents require particularly careful separation of safety action from causal classification.

If an autonomous or robotic system presents a credible safety concern, organisations may need to stop, isolate, constrain or otherwise make the system safe before root cause is established.

Classification should preserve:

- sensor/perception evidence;
- decision or model outputs where available;
- command and actuation records;
- safety-interlock state;
- human intervention;
- software/model/configuration versions;
- physical consequence;
- environmental conditions;
- unresolved non-AI mechanical or electrical explanations.

GAISSF™ remains relevant according to determined scope, including D9 where applicable. PAI-SF™ provides the more detailed physical-AI control set where its scope is relevant.

UAIF classification alone does not establish that AI caused a physical consequence, that a device is safe to return to service, or that sector-specific safety obligations have been satisfied.

> **PAI-SF™ HANDOFF**
>
> Where PAI-SF™ applies to the system, preserve sensor, actuator and physical-state evidence per PAI-SF™'s own evidence-preservation requirements before or alongside attempting UAIF L1 causal analysis — physical evidence can be time-limited in a way digital evidence is not. UAIF classification alone does not satisfy PAI-SF™ return-to-service validation; that determination follows PAI-SF™'s own process.

13. Worked Example 1 — Blocked Prompt Injection

Hypothetical scenario: An internal RAG assistant retrieves a document containing instructions intended to manipulate model behaviour. The application-layer policy blocks the attempted external action. No sensitive information is observed leaving the approved environment.

Observed event: Malicious or manipulative instructions were present in retrieved content and reached the model context.

Confirmed facts: The policy layer blocked the attempted action; no confirmed external action or disclosure has been identified.

Working hypothesis: The content represents an indirect prompt-injection attempt.

Unknowns: Whether the document was intentionally poisoned; whether similar documents exist; whether earlier sessions were affected.

Classification consideration: The blocking outcome does not by itself resolve the record_type determination either way; apply the classification discipline questions in Section 3 to the actual facts rather than assuming the absence of realized harm settles the record type. Where the evidence is not yet sufficient to make that determination, preserve the uncertainty in the investigation workflow rather than inventing an additional UAIF record_type.

Response already justified: Preserve the document, retrieval record, model/context trace where lawfully available, policy decision and affected-source metadata; investigate repository exposure and similar content.

What could change classification: Evidence of successful unauthorized action, disclosure, persistent compromise or realized harm.

Notably Absent: No conclusion here that the event is legally reportable, that a threat actor has been identified, or that prompt injection is the established root cause of any separate anomalous behaviour.

14. Worked Example 2 — RAG Data Disclosure

Hypothetical scenario: A user receives content from a restricted document that the user was not authorized to access.

Observed event: Restricted information appeared in an output.

Confirmed facts: The output and authorization state are preserved; the user lacked access to the source document.

Working hypothesis: Retrieval-filter failure, index-permission mismatch, cached context, application authorization defect, or another mechanism — the specific mechanism is not yet established, and the event should not be labelled RAG_Leakage as root cause until it is.

Unknowns: Whether other users received restricted content; how long the condition existed; whether the source was retrieved directly.

Classification consideration: Privacy or security-integrity harm can be assessed from the confirmed facts independently of the unresolved causal mechanism — the two are separate dimensions of the record. Do not label the mechanism RAG_Leakage as root cause until the retrieval pathway is specifically established; that causal question can remain open while the realized harm is characterized from what is already known.

Response already justified: Preserve the output, authorization state, retrieval logs and access-control evaluation; begin reproduction testing where safe to do so.

What could change classification: Retrieval logs, access-control evaluation, reproduction testing and evidence of wider exposure.

Notably Absent: Presence of restricted information in an output does not by itself establish the complete causal chain, total affected population, legal breach status or notification obligation.

15. Worked Example 3 — Unauthorized Agent Transaction

Hypothetical scenario: An enterprise agent invokes a payment tool and submits a transaction outside the user's intended authorization.

Observed event: A payment tool was invoked and a transaction was submitted outside the user's intended authorization.

Confirmed facts: Tool invocation and transaction records show the action occurred.

Working hypothesis: Agent escalation, policy-enforcement failure, prompt manipulation, authorization misconfiguration, or human configuration error.

Unknowns: What authority was actually delegated to the agent; what instruction the agent received; which policy-enforcement point evaluated the call; whether the transaction can be reversed.

Classification consideration: Relevant UAIF dimensions include L1 causal analysis, L2 severity, L5 agent/control-plane metadata, and L6 sector/jurisdiction metadata; an autonomous action does not by itself establish malicious compromise.

Response already justified: Stop further unauthorized action, preserve the invocation chain, protect funds/accounts, and begin AI-IRF response activity while classification continues.

What could change classification: Confirmation of which policy-enforcement point failed; evidence of prompt manipulation or a specific misconfiguration; reversal or containment outcome.

Notably Absent: An autonomous action does not establish malicious compromise. A financial consequence does not establish which technical mechanism caused it.

16. Worked Example 4 — Provider Model Change

Hypothetical scenario: Several enterprise applications begin producing materially different outputs after a third-party model provider deploys an update.

Observed event: Multiple dependent applications began producing materially different outputs within the same time window.

Confirmed facts: Behaviour changed within a known time window; multiple dependent applications are affected; the provider confirms an update occurred.

Working hypothesis: The provider update caused the behaviour change; temporal alignment alone is not sufficient for causal attribution and must be treated as a hypothesis, not a finding.

Unknowns: Whether the update actually caused the observed behaviour; what internal provider changes were material; whether all affected applications share the same mechanism.

Classification consideration: Preserve provider evidence separately from enterprise-observed evidence; determine whether linked records or a common incident relationship best preserves the multi-system impact.

Response already justified: Preserve before/after versions and the change timeline for each affected application; request provider evidence through established channels.

What could change classification: Provider-supplied technical detail on the update; independent reproduction isolating the change; evidence that only some applications share the mechanism.

Notably Absent: No conclusion that the provider caused the incident unless the causal evidence supports that conclusion.

17. Worked Example 5 — Clinical Decision-Support Error

Hypothetical scenario: An AI decision-support system produces an incorrect recommendation. A clinician reviews the recommendation before treatment.

Observed event: An AI decision-support system produced an incorrect recommendation.

Confirmed facts: The recommendation was generated and reviewed by a clinician before treatment.

Working hypothesis: The incorrect output reflects model behaviour, input-data error, integration failure, or another cause — the mechanism is not yet established.

Unknowns: Whether the recommendation affected the clinical decision; which of the candidate causes is responsible.

Classification consideration: Separate the incorrect output from realized patient harm; do not equate incorrectness with the severity or reportability of the event.

Response already justified: Preserve the relevant input/output, system/version information, clinical workflow record and authorized review evidence while applying healthcare-specific safety and reporting procedures.

What could change classification: Evidence that the recommendation was relied upon rather than overridden; confirmation of the technical mechanism; evidence of patient impact.

Notably Absent: This guide does not establish clinical negligence, medical-device reportability, patient harm, causality, or adequacy of human oversight.

18. Worked Example 6 — Physical-AI Malfunction

Hypothetical scenario: An autonomous inspection robot deviates from its expected path and triggers an emergency stop. No injury is observed.

Observed event: An autonomous inspection robot deviated from its expected path and triggered an emergency stop.

Confirmed facts: Path deviation and emergency-stop activation occurred; no injury is observed.

Working hypothesis: Perception error, localization failure, model/configuration change, sensor fault, environmental condition, mechanical defect or control-system issue — do not infer AI causality solely because the system contains AI.

Unknowns: Which of the candidate mechanisms is responsible; whether the condition is reproducible without physical risk.

Classification consideration: Maintain a safe state before attempting reproduction where reproduction could create physical risk; causal classification should not precede the safety action.

Response already justified: Preserve sensor data where available, actuator commands, safety-controller records, model/software versions, environment records, maintenance state and human observations.

What could change classification: Sensor and actuator log analysis; safety-controller record review; evidence ruling in or out a non-AI mechanical or electrical cause.

Notably Absent: No conclusion that the robot is safe to return to service, that AI caused the deviation, or that sector-specific engineering/safety requirements have been satisfied.

19. Worked Example 7 — Multi-System Cascading Incident

Hypothetical scenario: A shared retrieval index used by three internal AI assistants begins returning outdated pricing information after a connector misconfiguration during a scheduled content sync.

Observed event: Three internal AI assistants that share a retrieval index began surfacing outdated pricing information.

Confirmed facts: All three assistants began surfacing the outdated pricing content within the same maintenance window; the connector configuration change is confirmed via change-management records.

Working hypothesis: The connector misconfiguration is the common cause; a secondary hypothesis is that one assistant's caching layer independently masked an unrelated, pre-existing data-freshness defect.

Unknowns: Whether all three assistants are affected identically or to different degrees; whether any customer-facing decision relied on the outdated pricing before detection; whether the connector defect also affects index content beyond pricing.

Classification consideration: Do not assume identical impact or a single root cause across all three systems merely because they share the dependency; maintain a parent incident view with linked system-specific records so each assistant's evidence and blast radius can be verified independently.

Response already justified: Suspend or roll back the affected connector sync, preserve pre- and post-sync index snapshots for each assistant, and begin AI-IRF containment for the shared dependency while system-specific verification continues.

What could change classification: Verification that one assistant's issue has a distinct, unrelated cause; evidence of a customer-facing decision made on outdated information; discovery that the misconfiguration affects additional content categories.

Notably Absent: A shared dependency and a shared symptom window do not by themselves establish that every affected system shares the same severity, root cause, or notification obligation.

20. Classification Review Checklist

Before accepting a material classification decision, ask:

1. Is the record type supported by what actually occurred?
2. Are observed facts separated from causal hypotheses?
3. Are unknowns visible rather than silently resolved?
4. Is realized harm distinguished from potential harm?
5. Does severity use the UAIF normative method rather than an improvised score?
6. Is the provisional status of default harm weights understood?
7. Are provider representations distinguished from independently observed evidence?
8. Has blast radius been established rather than assumed?
9. Are generative/agentic mechanisms recorded only where supported?
10. Is a regulatory trigger being treated as technical routing rather than legal determination?
11. Can a reviewer understand why the classification changed over time?
12. Has operational response proceeded where justified without waiting for unnecessary classification certainty?

21. Relationship to AI-IRF™

UAIF and AI-IRF solve different but overlapping operational problems.

UAIF gives responders a structured way to record and classify the incident. AI-IRF structures the operational response.

In practice:

classification informs response → response produces new evidence → new evidence may change classification → changed classification may alter response or escalation

This is a feedback loop, not a one-way handoff.

Illustrative integration pattern

A practical enterprise implementation may look like:

existing alert / case → AI-specific intake and evidence preservation → UAIF field mapping and schema validation → local triage / escalation → AI-IRF™ response activities → new evidence → UAIF re-evaluation → governance / GAISSF™ control feedback

> **MAKING THE FEEDBACK LOOP SPECIFIC**
>
> The final step is more useful when tied to a specific control domain rather than left as general governance feedback. For example, a confirmed RAG_Leakage root cause (L1) would typically prompt reassessment of the GAISSF™ control domains covering supply-chain and runtime data-handling, not only the general governance domain. Record the specific GAISSF™ domain or control implicated in the post-incident review so the feedback loop produces a traceable control change rather than a general note.

For example, a SOC analyst can retain the organisation's existing ticket or case identifier while a UAIF-compatible record carries the AI-specific classification data alongside it. UAIF severity may inform local escalation policy, but provisional weights and technical routing flags should not be hard-coded into automatic legal-reporting decisions. As response activities produce new evidence, the UAIF record can be re-evaluated and the change history preserved.

Organisations can integrate UAIF information into existing SOC/CSIRT tickets, case-management systems, SIEM/SOAR workflows and regulatory review processes. This guide does not require replacement of those systems or proprietary ODA3 tooling.

22. Machine-Readable Interchange

UAIF is designed for machine-readable incident interchange.

The technical specification defines profile-based conformance and points to the separately maintained normative schema in TCR-STD-003. The ODA3 JSON Schema Center identifies the UAIF core JSON Schema as the normative machine-readable companion and instructs implementers to confirm version, publication state, checksums and release metadata before reliance. This guide uses only the controlled terms needed for practitioner orientation and does not reproduce the complete schema.

Machine readability can support:

- consistent field validation;
- controlled vocabularies;
- deterministic severity calculation;
- conditional field requirements;
- incident exchange between operational and governance systems;
- routing into sector or jurisdiction review processes.

It does not make the underlying evidence correct.

A schema can validate that a required field exists and contains an allowed value. It cannot, by itself, prove that the factual assertion recorded in that field is true.

> **SCHEMA VALIDATION ≠ EVIDENCE SUFFICIENCY**
>
> A schema-valid UAIF record confirms structural conformance: the required fields are present and contain allowed values. It does not confirm that the evidence supporting each field meets an organisation's own assurance threshold, that provider representations recorded in the payload have been independently verified, or that the record satisfies any particular regulatory or contractual evidence standard. Tool vendors and integrators in particular should not treat schema validation as equivalent to evidentiary or assurance sufficiency.

23. Notably Absent

This guide does not, by itself:

Legal & Regulatory

- establish legal liability, negligence, fault or malicious intent;
- determine whether an event is legally reportable;
- replace regulatory or sector-specific incident definitions;
- establish a regulator's severity or materiality determination;
- convert regulatory_trigger into a legal reporting determination;

Technical & Causal

- establish root cause because a plausible mechanism has been identified;
- establish actor attribution;
- establish causal linkage between an AI event and downstream physical, financial, clinical, environmental or societal harm without supporting evidence;
- establish that a third-party provider caused an incident;
- establish that UAIF's provisional default harm weights are empirically validated for a particular sector;

Framework & Certification

- prove that a GAISSF™, PAI-SF™ or other control failed;
- establish GAISSF™, UAIF™, AI-IRF™ or PAI-SF™ conformance or certification;
- treat a schema-valid record as equivalent to a UAIF™-conformant record, or a conformant record as equivalent to a certified or endorsed one — schema validity, conformance and certification are separate claims and none is established merely by the others;
- replace AI-IRF™ response activities.

Operational & Interoperability

- guarantee interoperability with a particular SIEM, SOAR, GRC or case-management platform;
- define a universal observability/logging tier, retention period, or evidence-quality ladder;
- establish that provider-side missing evidence prevents or permits incident closure in every organisation.

24. Next Reading

- WHEN AI BREAKS: The ODA3 Guide to AI Incident Response — operational incident-response orientation.
- The First 30 Days with GAISSF™ — implementation pathway for the wider assurance baseline.

License and Commercial Use Boundary

This guide, and integration of UAIF™ information into an organisation's own SOC/CSIRT, case-management, SIEM/SOAR or regulatory-review workflows for that organisation's internal incident handling, is within the permitted use described by GEL v1.0. Commercial embedding of UAIF™ into a third-party product, paid implementation services built on UAIF™, or other commercial exploitation of the framework are separate uses and are governed by the specific GEL v1.0 provisions and any applicable ODA3 partner terms that apply to that activity, not by this guide. Where commercial use is contemplated, confirm the applicable terms directly rather than inferring permission from this practitioner guide.

Publication Boundary

TCR-STD-002 (UAIF™ v1.0 Technical Specification) and TCR-STD-003 (UAIF™ v1.0 Schema Specification) are both published as Final Publication v1.0. Final Publication means the v1.0 normative architecture, controlled vocabulary, machine-readable conformance model and reference calculation semantics are frozen under ODA3 Institute change control. It does not mean that every sector calibration, legal interpretation, integration pattern or

empirical reliability claim has been validated in production — sector-specific calibration, independent inter-rater studies and broad SIEM/GRC interoperability testing remain outstanding validation work per TCR-STD-002.

Accordingly, a final publication of COM-012 means that this practitioner guide has completed its own publication gates, consistent with UAIF™ v1.0 itself having reached Final Publication status. It does not establish that this guide, or UAIF™, has completed the sector calibration and production-validation work TCR-STD-002 identifies as still outstanding.

Revision History

This guide has been revised through R1 (initial publication), R2 (editorial and rendering corrections), R3 (technical corrections and practitioner-usability improvements, developed across three internal review passes), R4 (direct verification against the controlling TCR-STD-002 and TCR-STD-003 specifications and the normative JSON schema), and R5 (architecture reconciliation — correcting stale L0/L4 layer attributions that survived R4's source verification). A summary appears below; the full itemized record is preserved afterward for traceability.

Summary

Revision	Summary
R1	Initial publication.
R2	Editorial and table-rendering corrections; one factual error introduced (a severity-boundary mislabel, corrected in R3).
R3	Corrected the R2 severity-boundary error; resolved the COM-011 checklist wording tension directly rather than deferring it; added practitioner-usability material (worked severity illustration, change-history template, competing-hypotheses guidance, chronic-harm evidence criteria, PAI-SF™ handoff, schema-validity/evidence-sufficiency distinction); standardized all seven worked examples to one structure; grouped Notably Absent by theme; corrected two items introduced during R3's own review process (an inferred severity level with no supporting calculation, and an example provider value presented in schema-like notation without confirmation).
R4	First revision with direct access to TCR-STD-002, TCR-STD-003 and the normative uaif-core-v1.0.json schema. Updated UAIF™ publication status from Pilot-Ready Specification Candidate to Final Publication v1.0 throughout (cover, Source and Traceability, controlled-vocabulary callout, Publication Boundary), matching both specifications' cover pages. Rewrote the Section 5 “Why Two Scores” and severity-illustration callouts to state the confirmed calculation mechanism directly. Confirmed the guide's exclusion of Observation as a record_type value was correct and required no change, but did not catch that the surrounding text still attributed record_type to L4 rather than its actual layer, L0 (corrected in R5). Confirmed Section 10's cautious treatment of the provider example was justified, since the example value is not in fact a valid schema value.
R5	Architecture reconciliation. A subsequent check against TCR-STD-002's own layer table found that record_type is an L0 field (Record Identity & Workflow), not L4, and that L4 is a distinct layer (Incident–Vulnerability Relationship, covering linked_vulnerability_id and exposure_duration_hours) that does not perform the Incident/Vulnerability classification itself — R4's source check had verified the Observation exclusion but missed this separate, pre-existing layer mislabelling throughout Sections 2–3 and Worked Example 1. Corrected the seven-layer architecture table, Section 3's opening paragraph and Observation callout (retitled to name record_type specifically), the COM-011 reconciliation callout, the practitioner classification sequence table, the record_type triage aid and its table cells, Worked Example 1, and the L0 representative-fields list (added the previously missing record_type and profile fields, with a new callout explaining profile/extensions). Also removed two residual stale references (a leftover “Publication Candidate” citation and a leftover “Pilot-Ready” sentence) that R4 had not fully swept from Source and Traceability.

Verification status

SOURCE VERIFICATION STATUS

This guide has been checked directly against TCR-STD-002 (UAIF™ v1.0 Technical Specification, Final Publication v1.0), TCR-STD-003 (UAIF™ v1.0 Schema Specification, Final Publication v1.0) and the normative uaif-core-v1.0.json schema, across two verification passes (R4 and R5). Items previously flagged as requiring verification are now resolved: (1) the Section 5 “Why Two Scores” explanation matches the confirmed calculation order in TCR-STD-002 §4.2 — dominant harm, context modifiers, impact scale, temporal adjustment, then presentation score and severity_level; (2) record_type's two normative values (Incident, Vulnerability) and the exclusion of Observation are confirmed identically across TCR-STD-002 §2.1, TCR-STD-003 §9 and the JSON schema enum, and this guide's layer attribution now correctly places record_type in L0 (Record Identity & Workflow) rather than L4, with L4 correctly described as the separate Incident–Vulnerability Relationship layer — R4 verified the Observation exclusion but missed this layer mislabelling, which R5 found and corrected against TCR-STD-002's own layer table; (3) UAIF™ v1.0's publication status is confirmed as Final Publication v1.0 on both TCR-STD-002 and TCR-STD-003, and this guide is updated throughout to match, while retaining the specification's own distinction that Final Publication

freezes the normative architecture without certifying sector calibration or production validation. One caution stands on its own merits rather than as an open verification item: Section 10's provider example deliberately avoids schema-value-looking notation, since Provider_Model_Update is confirmed not to be a valid root_cause_specific_type value in the actual schema — the guide's explanatory-notation approach there was the right call.

Full itemized record

R3 — first pass

Area	Correction
Section 8 (correction, not R2 addition)	R2's callout mislabelled 6.5 as the “High/Critical” severity-band boundary. Per the Section 5 table, 6.5 is the Medium/High boundary (High begins at 6.5 and runs to below 8.5; Critical begins at 8.5).
Section 3	Added a “Relationship to First-Response Checklists” callout resolving the COM-011 wording tension directly, rather than deferring it to COM-011's next revision cycle as R2 had done.
Section 3	Replaced “recovered TCR-STD-002” with “controlling UAIF™ technical specification” throughout.
Section 3	Added a simplified L4 (Incident vs Vulnerability) triage table, with a footnote against over-reading it as a substitute for the classification discipline questions.
Source and Traceability	Added a “How to Read This Guide” callout stating that all field names and examples are illustrative and that the TCR-STD-003 schema governs where it differs from this guide.
Section 4	Promoted the controlled-vocabulary version warning to a callout box, and added guidance on documenting competing causal hypotheses with individual confidence levels.
Section 5	Added a worked severity-scoring illustration (revised twice in subsequent passes — see below) and a ready-to-use severity-reporting statement template.
Section 6	Added a callout on what does (and does not) constitute evidence of chronic harm (later softened — see below).
Section 10	Added a worked example showing how to populate provider-related fields when evidence is incomplete (later corrected — see below).
Section 12	Added an explicit PAI-SF™ evidence-preservation handoff sentence.
Worked Example 1	Softened the classification-consideration wording so a blocked attempt with no realized harm is not read as implying the record defaults to Vulnerability.
Worked Examples 2–7	Standardized every example to the same seven-field structure used in Example 1.
Section 21	Added a worked illustration tying a specific root-cause finding to a specific GAISSF™ control domain.
Section 23 (Notably Absent)	Grouped the 16 items under four subheadings for scannability, and added an explicit schema-valid ≠ conformant ≠ certified item.
Next Reading	Removed the “forthcoming” field-guide reference; added a License and Commercial Use Boundary.
Throughout	Normalized spelling to British convention; reduced repetition of the “Pilot-Ready” caveat to one clear statement per relevant section.
Structural	Subheadings converted to Heading 2 style; table header rows marked as semantic headers.

R3 — second pass

Area	Correction
Section 5 (correction)	The first-pass worked severity example presented specific numeric scores (3.2) without showing how they were derived; replaced with a non-numeric illustration.
Section 5	Added a “Why Two Scores” callout distinguishing severity_analytical_score from severity_presentation_score.
Worked Example 2 (correction)	First-pass wording implied realized harm could not be assessed until the causal mechanism was established, inconsistent with the guide's own separation of manifestation, harm and cause. Corrected to state harm can be assessed independently of the unresolved mechanism.
Section 3	Strengthened the L4 triage-table footnote to explicitly address the case where the evidence supports neither Incident nor Vulnerability.
Section 6	Softened “statistically significant” in the chronic-harm evidence callout.
Section 3 (COM-011 callout)	Added a version-pinning note tying the reconciliation to COM-011 FINAL R2.
Section 9	Added a five-field minimum template for change-history recording.
Section 22	Added a “Schema Validation ≠ Evidence Sufficiency” callout.

R3 — third pass

Area	Correction
Section 5 (correction)	The second-pass severity illustration removed the numeric scores but still inferred a specific severity_level (Level 2 / Low) from the scenario — a smaller version of the same false-precision problem, since the published band is itself a calculated consequence of the presentation score. Corrected to state the discipline (record evidence, apply the method, take the result from the calculation) without inferring any score, band or level.
Section 5 (“Why Two Scores”)	The explanation asserted a specific mechanism for score divergence (“asymmetric context or scale adjustments”) that had not been checked against any source. Revised to describe the relationship more generally and to state explicitly that it is a practitioner-level summary requiring confirmation against the controlling specification.
Section 10 (provider worked example)	The example showed ‘root_cause_primary_domain: Provider_Model_Update’, placing an unconfirmed label directly after a real controlled-field name. Rewritten as explanatory practitioner notation with an explicit statement that the labels are not schema field names or controlled values.

R2 (carried forward, one item later corrected in R3)

Area	Correction
Tables (Sections 1, 2, 3, 5, 10)	Rebuilt as properly structured tables.
Section 3	Moved the "observation is an investigative concept, not a UAIF record_type" clarification ahead of the classification discipline question list.
Section 5	Added a footnote clarifying the Critical band's closed interval is intentional.
Section 8	Added a callout disambiguating the shared 6.5 default — mislabelled at the time as the “High/Critical” boundary; corrected to Medium/High in R3.
Section 9	Added guidance on where to record classification change history absent a dedicated field.
Section 19	Added Worked Example 7 (multi-system cascading incident).
Executive Summary	Promoted the thesis statement to a Key Principle callout box.

Not actioned: a physical split of the document into a core guide and a separate examples appendix, page reordering, checklist relocation, and new visual diagrams were proposed during review but are layout and publication-architecture decisions reserved for the document owner. The COM-011 wording tension is resolved on this document's side (Section 3); a corresponding small wording update to COM-011's own checklist is recommended but falls within COM-011's document, not this one.