ODA3 INSTITUTE

Practitioner Cheat Sheet

AI Security Incident Response Playbook

DocID: ODA3-2026-07-CHT-SEC-011
Classification: Public — Practitioner Reference

Version 1.2 | Public | 17 July 2026

Publisher: ODA3 Institute

Publication type: Practitioner Cheat Sheet

Review status: Publication edition

How to use this playbook: Prepare the authority, telemetry and recovery mechanisms before an incident; create and maintain a structured incident record when a trigger occurs; execute containment and recovery against the complete AI system boundary; preserve uncertainty and evidence; then return verified findings to governance and control improvement.

> **Key definition — AI security incident:** An event involving an AI system that jeopardizes or may jeopardize confidentiality, integrity, availability, safety, authorized behaviour, accountable decision-making or compliance within a defined operational boundary.

1. Executive Overview

AI security incident response extends conventional incident response rather than replacing it. The difference is the object being investigated: an AI-enabled service may combine a model, prompts, data pipelines, retrieval stores, agents, tools, identities, third-party APIs, human approvals and downstream actions. Restoring a server does not establish that the model, data, behaviour or decision pathway is trustworthy.

Response decisions can also change evidence. Disabling a model endpoint may stop harm but destroy volatile context; retraining can overwrite the affected state; modifying a system prompt can remove the exact condition needed for reproduction; and a supplier may rotate an opaque model without preserving the version used during the event. Preparation must therefore define evidence, authority and rollback mechanisms before production use.

This playbook uses the ODA3 framework suite as its primary operating architecture. **GAISSF™** establishes preparedness, control objectives, ownership, evidence expectations and assurance. **UAIF™** structures incident identity, classification, impact, causality and evidence. **AI-IRF™** structures preparation, identification, containment, eradication, recovery and lessons learned. Confirmed lessons return to GAISSF™ so that controls and assurance scope change rather than the incident closing as an isolated ticket.

NIST SP 800-61 Rev. 3 integrates incident response with cybersecurity risk management and NIST CSF 2.0. The EU AI Act contains role- and classification-dependent logging, post-market monitoring and serious-incident duties. These sources support external alignment; they do not establish equivalence with the ODA3 suite or determine legal applicability for a particular organization.

Primary actors are the incident commander, system owner, SOC/IR team, AI security architect, MLOps/data engineering, privacy and legal counsel, safety specialists, supplier management, communications and independent assurance. Each must know who can suspend an endpoint, revoke agent tools, roll back a model or retrieval index, preserve evidence and authorize recovery.

> **Notably Absent / Evidence Boundaries:** This playbook does not establish that every harmful or inaccurate output is a security incident, that every AI incident requires regulatory notification, or that containment proves root cause. It does not provide a legal opinion or a verified UAIF™ severity identifier. Exact ODA3 control, severity, routing, role and phase identifiers were not available in the authoritative source set for this edition and are not inferred. No claim is made that the ODA3 suite is regulator-recognized, accredited or the only operational approach.

2. Key Concepts & Definitions

Term	Operational definition
GAISSF™	ODA3 governance and assurance framework used here for preparedness, control ownership, evidence expectations and improvement.
UAIF™	ODA3 incident framework used here to structure incident identity, classification, causality, impact and evidence without inventing unavailable identifiers.
AI-IRF™	ODA3 response framework used here to structure preparation, identification, containment, eradication, recovery and lessons learned.
AI system boundary	Approved combination of model, prompts, data, retrieval, tools, identities, infrastructure, users and downstream actions.
AI security event	Observable occurrence that may be benign, a control failure or an incident; it requires triage before classification.
AI security incident	Event that has or may have material adverse security consequences within the AI system context.
Incident identity	Stable identifier joining reports, evidence, affected components, decisions and related incidents.
Incident commander	Person authorized to coordinate response, set objectives, assign actions and maintain decision cadence.
Causality status	Explicit state—unknown, suspected, supported or confirmed—describing evidentiary confidence in a causal explanation.
Prompt injection	Untrusted instructions influence an AI system to disregard or conflict with intended instructions.
Indirect prompt injection	Malicious instructions enter through data the system retrieves or observes, such as a document, message or webpage.
Agent hijacking	Manipulation that causes an AI agent to pursue an unauthorized objective or use tools in an unintended way.
Data poisoning	Manipulation of training, tuning, evaluation or retrieval data to alter system behaviour.
Model extraction	Unauthorized reconstruction or acquisition of model functionality, parameters or proprietary behaviour through access or queries.
Model inversion	Attempts to infer information about training data or sensitive attributes from model access or outputs.
Retrieval compromise	Unauthorized modification, insertion, deletion or prioritization of content used by a retrieval-augmented system.
Tool abuse	Unauthorized or unsafe invocation of APIs, code execution, messaging, payment or other capabilities exposed to an AI agent.
Decision trace	Record connecting input, model/configuration, retrieved context, output, human intervention and downstream action.
Model rollback	Restoration of a previously approved model and its compatible configuration, data and prompt policy—not model weights alone.
Kill switch	Tested mechanism that suspends an endpoint, agent, tool or action path through an independent control channel.
Nonhuman identity (NHI)	Service account, API credential, OAuth grant, workload identity or other machine credential used by an AI component; it remains a separate containment surface even after the agent process is stopped.
Version-pinned release bundle	Reconstructable approved combination of model, prompt policy, retrieval/index state, tool schema, guardrails and runtime configuration that is released, rolled back and tested as a compatible unit.
Blast radius	Users, data, decisions, systems, identities, suppliers and downstream actions that may have been affected—not merely the component where the anomaly appeared.
Evidence manifest	Index of collected artefacts with source, time, collector, hash, custody, access

	restriction and relevance.
Chain of custody	Documented history of evidence possession, transfer, access and transformation.
Containment boundary	Components, users, identities, data and actions isolated to prevent propagation while preserving essential evidence.
Recovery criterion	Testable condition that must be satisfied before restricted or full service restoration.
Substantial modification	Change that may affect intended purpose, compliance, threat exposure or control effectiveness and triggers reassessment.
Serious incident	Defined legal or framework-specific category; do not apply it without confirming the governing instrument and organizational role.
Evidence tier	T1 primary verified; T2 corroborated authoritative; T3 limited/single-source; T4 unverified or hypothesis.
Notably Absent	Material negative finding or evidence limit stating what the investigation did not establish.

3. Threat and Incident Landscape

3a. Actor taxonomy

Actor	Motivation	Capability	Likely vectors
External attacker	Access, fraud, exfiltration, disruption, influence	Medium–high	Prompt injection, credential theft, tool abuse, endpoint exploitation, supply chain
Malicious insider	Gain, sabotage, concealment	Medium–high	Data poisoning, model theft, evidence tampering, privileged tool use
Well-intentioned user	Complete task quickly	Low–medium	Sensitive prompt submission, unsafe reliance, shadow AI, unintended action
Content/data contributor	Manipulation, visibility, poisoning	Medium	Malicious document, webpage, message, dataset or retrieval content
Supplier compromise	Upstream intrusion or malicious change	High	Model/API change, dependency compromise, malicious update, log loss
Automated system	No motivation; failure condition	Variable	Drift, feedback loop, runaway agent, integration error, misrouting
Governance-process attacker	Avoid scrutiny or obtain approval	Medium	Under-classification, evidence omission, exception abuse, retrospective approval

3b. Incident paths

MITRE ATLAS references are functional families; confirm exact technique identifiers against the current release before formal reporting.

Path	Description	Prerequisites	Impact	ATLAS / security relationship
Direct prompt injection	User input changes system behaviour beyond authorized purpose	Model accepts untrusted text; weak instruction separation	Disclosure, policy bypass, unsafe output	Prompt injection family

Indirect injection / agent hijacking	Retrieved content instructs an agent to invoke tools or alter objective	RAG/agent reads attacker-controlled content	Credential theft, unauthorized messages/actions	Prompt injection; abuse of ML-enabled product
Retrieval poisoning	Malicious content enters or is prioritized in knowledge base	Weak ingestion, provenance or access control	Persistent output manipulation	Data poisoning / supply-chain relationship
Training or tuning poisoning	Data changes create targeted or broad behaviour	Access to pipeline or source data	Backdoor, degradation, biased or unsafe output	Poisoning family
Model/API credential compromise	Attacker uses privileged inference, management or deployment access	Credential theft or excessive privilege	Model theft, misuse, configuration change	Conventional IAM TTPs plus ML impact
Tool/agent privilege abuse	Agent invokes action outside intended scope	Over-broad tools; insufficient authorization	Fraud, deletion, lateral movement, unsafe act	Abuse impact; conventional ATT&CK applies
Supplier/model drift	Upstream model changes without validated notice	Mutable service; no pinning or regression gate	Broken controls, changed behaviour	ML supply chain
Evidence impairment	Logs, prompts, configuration or approvals are deleted/altered	Mutable storage; conflicted access	Unreconstructable event, false assurance	Impair-defenses relationship
Resource exhaustion	High-cost or high-volume inference/tool loop	Weak budgets, quotas or termination logic	Cost, outage, downstream saturation	Denial-of-service relationship

3c. Impact analysis

Impact	Scenario	Triage priority	UAIF™ record requirement	Stakeholders
Confidentiality	Agent sends protected data to unauthorized destination	Critical	Data classes, recipients, access path, exposure confidence	Privacy, legal, customers
Integrity	Poisoned retrieval changes recommendations	High–critical	Changed content, provenance, affected decisions, causality	Product, data, business owner
Availability	Agent loop exhausts quota and blocks service	High	Start/end, resources, dependencies, recovery	Operations, customers
Safety	AI action influences physical or clinical process outside boundary	Critical	Harm/exposure, human intervention, uncertainty	Safety, regulators, affected people
Financial	Unauthorized action creates payment or contractual commitment	Critical	Transaction trace, authorization, reversibility	Finance, legal, customers
Regulatory	Applicable high-risk system lacks logs or serious-incident process	High	Role/classification, duties considered, counsel decision	Compliance, governing body
Reputational	Unsupported disclosure or inaccurate public response worsens event	High	Communications decisions and confirmed facts	Board, communications

Priorities above are playbook triage labels, not verified UAIF™ severity-tier identifiers or legal classifications.

3d. Documented findings and cases

Finding / case	Evidence	What occurred	Response lesson	Notably absent
NIST CAISI agent-hijacking experiments	**T1:** NIST technical publication	Indirect prompt-injection evaluations showed that malicious	Preserve retrieved content, full tool trace and authorization	Controlled evaluation, not evidence of a production

(2025)		instructions embedded in data can redirect agents toward harmful tasks in controlled environments.	decisions; isolate tools independently of the model.	compromise or population-wide incident rate.
NIST CAISI evaluation of DeepSeek models (2025)	**T1:** NIST evaluation report	In simulated agent scenarios, hijacked agents performed actions including phishing, malware execution and credential exfiltration; comparative results were model- and test-specific.	Response testing must cover repeated adaptive attacks and consequential tool actions, not single-turn chatbot refusal.	No claim that these exact simulated actions occurred in a customer production environment.
Moffatt v. Air Canada (2024)	**T1:** tribunal decision	A customer-facing chatbot provided inaccurate policy information and the organization was held responsible in the dispute.	Treat harmful AI output as an organizational incident candidate; preserve content, source policy, model/configuration and human resolution path.	Not a confirmed malicious cyberattack; the decision does not establish universal liability for chatbot errors.
Samsung source-code exposure via public LLM (2023)	**T2:** multiple contemporaneous news reports; corroborated by subsequent corporate policy change	Employees pasted proprietary source code into a public generative-AI tool with no enterprise gateway, inventory or egress control in place.	Shadow-AI incidents require network-level containment (block the destination, not just the user); response readiness must include a pre-tested egress-blocking capability, not only an acceptable-use policy.	No verified evidence of subsequent third-party exploitation of the exposed code; this is a governance/control gap finding, not a confirmed adversarial compromise.

4. Technical and Structural Deep Dive

4.1 Response architecture

```
GAISSF™ preparedness
  owners • controls • evidence • authority • exercises • recovery criteria
                              ↓ trigger
UAIF™ incident record
  identity • classification • impact • causality • evidence • related events
                              ↓ routing
AI-IRF™ response
  prepare → identify → contain → eradicate → recover → learn
                              ↓ verified findings
GAISSF™ improvement
  control change • assurance test • residual-risk decision • portfolio search
```

Plane	Components	Response requirement
Control	Governance registry, approval, exception and evidence store	Know approved boundary and decision authority
AI runtime	Model gateway, system prompt, policy, inference service	Capture version, trace and policy results
Data	Training/tuning data, retrieval index, evaluation sets	Preserve provenance and change history
Agent/	Tool broker, service identities, transaction controls	Revoke or restrict action independently

action		
Infrastructure	Cloud, container, endpoint, network, secrets	Apply conventional IR and forensic methods
Human decision	User, reviewer, approver, override and downstream consumer	Reconstruct reliance and intervention
Supplier	Model/API provider, data source, platform, subprocessors	Trigger joint response and obtain evidence

4.2 Incident mechanics

1. **Trigger:** monitoring, user report, supplier notice, red-team result or external disclosure identifies a possible condition.
2. **Stabilize:** protect people and critical assets; avoid uncontrolled changes to the affected state.
3. **Create incident identity:** assign one record before parallel teams create disconnected narratives.
4. **Bound the system:** identify exact model, prompt policy, data, retrieval, tools, identities, users and downstream actions.
5. **Preserve evidence:** collect volatile traces and authoritative configuration before rollback where safe.
6. **Classify provisionally:** record impact, confidence, causality status and unknowns; do not wait for perfect certainty.
7. **Contain by capability:** disable the smallest unsafe action path that materially reduces risk.
8. **Investigate:** reproduce under controlled conditions; compare affected and approved baselines.
9. **Eradicate:** remove compromised input, permission, dependency, configuration or process weakness.
10. **Recover under conditions:** stage restoration; test security, business and safety outcomes.
11. **Close the loop:** update GAISSF™ controls and assurance; finalize UAIF™ record and AI-IRF™ lessons.

4.3 Enabling conditions

Condition	Diagnostic test
Model/version opacity	Can operations identify the exact model used for a disputed trace?
Missing decision trace	Can input, context, output, reviewer and downstream action be joined?
Unbounded tools	Can the agent initiate irreversible or privileged actions without separate authorization?
Shared credentials	Can a tool action be attributed to the agent, user and approval?
Mutable retrieval	Can investigators prove which documents and ranks were used at event time?
Supplier evidence gap	Does contract require incident notice, logs, preservation and change history?
No independent kill path	Can the affected agent disable or bypass its own containment control?
No independent ingestion stop	Can operations suspend external-content ingestion or refresh without depending on the chat interface?
Untested rollback	Does the previous model remain compatible with prompt, index and tool schema?
Content over-collection	Are raw prompts retained without minimization, access controls or legal basis?

4.4 Stage indicators

Stage	Observable artefacts
Entry	Anomalous prompt patterns, malicious retrieved instructions, unusual OAuth/tool grant, supplier alert
Objective change	System/developer instruction conflict, goal deviation, unexplained plan change
Capability use	New tool sequence, sensitive query, privileged API, unusual destination or transaction
Persistence	Poisoned document/index entry, modified prompt policy, changed model/config, scheduled action

Impact	Unauthorized disclosure/action, harm report, degraded output, cost spike, outage
Concealment	Missing trace spans, disabled logging, altered evidence, approval created after event

4.5 Hardened versus vulnerable response

Vulnerable baseline	Hardened response pattern
Generic ticket “AI gave wrong answer”	Stable incident record with system boundary, impact and causality status
Disable whole service immediately	Preserve critical evidence, then contain the smallest unsafe capability consistent with safety
Roll back model weights only	Roll back compatible model, prompt policy, index, tools and configuration
Screenshot as primary evidence	Export authoritative logs, manifests and hashes with screenshot as context
Assume prompt injection caused event	Maintain competing hypotheses until evidence supports causality
Supplier says “resolved”	Obtain affected versions, timeline, evidence and validation conditions
Restore when outputs look normal	Restore only after defined security, business and safety tests pass

4.6 Minimum incident record

```
{
  "incident_id": "aii-2026-0042",
  "status": "contained",
  "system_id": "ai-000123",
  "detected_at": "2026-07-17T08:22:14Z",
  "reported_by": "monitor:agent-tool-anomaly",
  "affected_boundary": {
    "model": "provider/model@immutable-version",
    "prompt_policy_hash": "sha256:...",
    "retrieval_index": "kb-prod@2026-07-16",
    "tools": ["mail.read", "mail.send"],
    "identities": ["svc:agent-support"]
  },
  "impact": {"confirmed": [], "potential": ["credential disclosure"]},
  "causality": {"status": "suspected", "hypotheses": ["indirect_prompt_injection"]},
  "containment": ["mail.send revoked", "index version isolated"],
  "evidence_manifest": "sha256:...",
  "unknowns": ["whether downstream recipient accessed message"],
  "notably_absent": ["no confirmed production-data modification"]
}
```

4.7 Safe triage query

```
SELECT trace_id, system_id, model_version, principal_id,
       tool_name, authorization_result, destination_class,
       policy_result, event_time
FROM agent_tool_events
WHERE event_time BETWEEN :start AND :end
  AND system_id = :system_id
  AND (authorization_result = 'denied'
       OR destination_class = 'external'
       OR tool_name IN (:sensitive_tools));
```

5. Practitioner Decision Framework

5a. Triage scoring aid

This matrix prioritizes response; it is not a verified UAIF™ severity method, CVSS score or regulatory classification. Score 1–4, multiply by weight and divide by 100. Any hard stop overrides the numeric result.

Factor	Weight	1 — Low	2 — Moderate	3 — High	4 — Critical
Confirmed consequence	20	No adverse effect	Reversible internal effect	Material external/financial effect	Safety, rights or major loss
Capability exercised	15	Output only	Read access	Bounded write/action	Privileged or irreversible action
Data exposure	15	Public	Internal	Confidential/personal	Secrets, regulated or special-category
Propagation	10	Single trace	One workflow	Multiple users/systems	Broad, autonomous or supply-chain
Persistence	10	None	Session	Stored content/config	Model, pipeline or supplier persistence
Detectability	10	Full trace/replay	Strong logs	Partial trace	Material actions unobservable
Reversibility	10	Immediate	Bounded effort	Partial/uncertain	Irreversible or safety-critical
Regulatory/safety trigger	10	None identified	Contractual	Potential notification	Serious/prohibited-use concern

Suggested bands: 1.00–1.49 Low; 1.50–2.24 Moderate; 2.25–2.99 High; 3.00–4.00 Critical. Validate thresholds against organizational incidents and risk appetite.

Hard stops: active safety threat; continuing unauthorized action; suspected secret/credential exposure; compromised response identity; evidence destruction; prohibited-use concern; unknown system owner for a consequential system.

5b. Decision tree

```
Is harm active or is an agent still able to perform unauthorized actions?
  YES → Revoke unsafe tools/identities or suspend endpoint through independent control.
        Preserve volatile evidence where this does not delay protection.
  NO  → Preserve state and start bounded investigation.

Is the exact affected boundary known?
  NO  → Restrict change and external action; discover model, data, tools and dependencies.
  YES → Compare runtime with approved boundary.

Is there confirmed or plausible sensitive-data, safety, rights or material financial impact?
  YES → Escalate to executive owner, legal/privacy/safety and communications as applicable.
        Assess notification duties by role, system classification and jurisdiction.
  NO  → Continue technical triage with documented review time.

Is causality confirmed?
  NO  → Maintain competing hypotheses; do not publicly attribute.
  YES → Eradicate root condition and search portfolio for the same exposure.

Have recovery criteria passed on the exact restored configuration?
  NO  → Keep restricted; remediate and retest.
  YES → Stage restoration with enhanced monitoring and authorized residual-risk decision.
```

5c. Prioritization

Scenario	Priorit	Immediate action	Owner	Target response

	y			
Agent actively sending unauthorized messages/actions	Critical	Revoke tool and service identity; isolate agent; preserve traces	Incident commander	Immediate
Suspected credential disclosure in model context	Critical	Rotate credential; block use; identify access and destinations	Security lead	Immediate
Poisoned retrieval document with unknown reach	High	Remove from active index without destroying evidence; identify affected traces	Data/RAG owner	Same operational period
Supplier changes model during investigation	High	Pin or restrict service; demand preservation/version evidence	Supplier manager	Immediate escalation
Harmful output with no downstream action	Moderate–high	Preserve trace; restrict use case; assess recurrence and affected users	System owner	Risk-based
Red-team finding with no production evidence	Moderate	Reproduce, scope exposed systems, prioritize remediation	Security testing lead	Before next material release

5d. Containment selection

Option	Use when	Evidence risk	Business trade-off
Tool revocation	Model output can remain but action capability is unsafe	Low if logs preserved	Retains advisory function
Identity suspension	Service principal may be compromised or over-privileged	Session/token evidence may expire	Stops dependent workflows
Endpoint restriction	One external surface is affected	Live state may change	Bounded outage
Model rollback	Behaviour changed with model/config version	Current state may be lost	Compatibility and regression risk
Retrieval quarantine	Specific content/index suspected	Ranking/served-context evidence may change	Reduced knowledge coverage
Ingestion-path suspension	External content continues entering or refreshing an affected retrieval pipeline	Queued items and transformation state may expire or change	Knowledge freshness and dependent workflows degrade
Human-only mode	Automation unsafe but service essential	Manual decisions still need logging	Slower throughput
Full shutdown	Safety, broad compromise or unknown propagation	Highest operational disruption	Maximum immediate risk reduction

5e. Incident-archetype routing

Use this table to select the first specialized response path after creating the common UAIF™-structured incident record. An archetype routes work; it does not prove cause or replace the triage score.

Incident archetype	First specialist path	Immediate containment	Evidence that is easy to lose
Prompt or agent hijacking	Agent/action investigation	Revoke high-risk tools; isolate affected context source	Full retrieved content, hidden instructions, plan and tool trace
Compromised skill, plugin or connector	Software/supply-chain response	Disable distribution and execution; revoke grants and tokens	Package version, manifest, publisher identity, install population, network destinations
Retrieval poisoning	Data/RAG investigation	Quarantine affected content or index version	Served document IDs, rank, ingest actor, index snapshot and access history
Training/tuning	ML pipeline	Freeze promotion and affected	Dataset snapshot, lineage,

poisoning	forensics	pipeline; isolate model versions	commits, training configuration and evaluation results
Sensitive-data disclosure	Privacy and data-breach response	Stop destination/action path; rotate exposed secrets	Exact recipients, content class, access/download evidence and deletion status
Unauthorized consequential action	Transaction/ business-process response	Suspend write capability; preserve and reverse transactions where safe	Authorization result, human approval, downstream receipt and reversal state
Supplier model or API compromise	Joint supplier response	Pin, restrict or replace affected service	Provider version, change history, tenant logs, notice and preservation request
Harmful output without confirmed attack	Product safety/governance response	Restrict use case or require human review	Decision trace, affected population, policy source and reliance evidence
Evidence impairment	Insider/forensic integrity response	Restrict evidence administration; preserve alternate telemetry	Audit trail, deletion events, access history and backup copies

5f. Notification decision record

Do not copy a deadline from a generic table into an incident ticket. Legal, contractual and regulatory duties depend on the facts, organizational role, system classification, jurisdiction and governing instrument. Record the decision even when notification is not required.

Required field	Question to answer
Governing instrument	Which law, regulation, contract, policy or sector rule was assessed?
Organizational role	Provider, deployer, importer, distributor, processor, controller, operator or another defined role?
Trigger facts	What confirmed facts may satisfy the reporting threshold?
Classification	What system, data, harm or incident classification applies, and who approved it?
Time origin	Which event starts the clock: awareness, confirmation, materiality decision or another defined point?
Deadline and channel	What deadline and authority/party apply according to current controlled guidance?
Evidence status	Which facts are confirmed, inferred, unknown or notably absent?
Decision	Notify, defer pending facts, voluntarily disclose or no notification—and why?
Owner and approval	Who owns submission, legal review and executive authorization?
Follow-up	What supplemental, correction, affected-person or preservation action remains?

6. Security Controls Matrix

Mappings are functional relationships, not equivalence or conformity determinations. Exact ODA3 identifiers require authoritative source verification; ISO audit use requires licensed standards text.

Control	Requirement	GAISSF™	UAIF™	AI-IRF™	Evidence	Type; eff./compl ex.	External alignment
IR-01 System and owner registry	Every production AI system has stable ID, boundary, owner and critical dependencies.	Preparedness/ accountability	Incident identity context	Mobilization context	Registry reconciliation	P; H/M	NIST CSF ID.AM; AI RMF GOVERN/MAP
IR-02 Response authority	Pre-authorize incident command, stop, rollback, evidence hold and communication	Decision rights	Routing stakeholders	All phases	Authority matrix, exercise	P; H/L	NIST CSF GV.RR/RS; ISO 27001 incident controls

	decisions.						
IR-03 AI telemetry baseline	Record model/config, trace, retrieved content references, tool authorization and downstream action.	Evidence expectation	Classification/evidence	Identification	Log schema and test event	P/D; H/H	NIST CSF DE.CM/DE.AE; EU AI Act logging where applicable
IR-04 Evidence manifest and custody	Hash, restrict and track collected artefacts and transformations.	Assurance evidence	Evidence integrity	Investigation	Manifest, custody log	D; H/M	NIST SP 800-61r3; SP 800-86
IR-05 Independent containment channel	Disable agent/tool/endpoint outside affected orchestration path.	Resilience control	Containment record	Containment	Kill-switch test	P/C; H/H	NIST CSF PR/RS; EU AI Act oversight context
IR-06 Scoped agent identities	Separate identities, allow-list tools and impose transaction/time/data limits.	Access control	Affected identity/action	Containment	IAM state, action tests	P/D; H/H	OWASP LLM06 (2025); ISO 27001 access controls
IR-07 Component version binding	Bind model, prompt, index, tools and policy to each release and trace.	Approved boundary	Incident component identity	Rollback/ recovery	Signed manifest	P/D; H/H	NIST AI RMF MAP/MEASURE; CSF PR.PS
IR-08 Retrieval provenance	Record source, ingest/change, authorization, index version and served references.	Data control	Causality/ evidence	Eradication	Lineage/export	P/D; H/H	NIST AI 100-2e2025; ISO data governance context
IR-09 Supplier incident clauses	Require notice, preservation, affected versions, logs, cooperation and remediation evidence.	Supplier governance	External-source evidence	Joint response	Contract and exercise	P; H/M	NIST CSF GV.SC; ISO 27001 supplier controls
IR-10 Recovery test pack	Predefine security, functional, safety and regression criteria for staged restoration.	Assurance criteria	Closure evidence	Recovery	Signed test results	C; H/M	NIST CSF RC; AI RMF MANAGE
IR-11 Notification matrix	Map potential triggers by role, jurisdiction, facts and decision authority; keep legally controlled.	Compliance governance	Impact/ routing facts	Escalation/ recovery	Counsel-approved matrix	P; H/M	EU AI Act Art. 73 where applicable; privacy/sectoral law
IR-12 Exercise program	Test prompt injection, poisoning, agent abuse, supplier failure, evidence loss and rollback.	Assurance/ improvement	Taxonomy coverage	Preparation	Exercise reports	P/C; H/M	NIST SP 800-61r3; CSF GV/RS/RC
IR-13 Portfolio hunt	Search related systems for the same model, supplier, data source, tool or control failure.	Portfolio risk	Related-incident links	Eradication	Query and results	D/C; H/M	NIST CSF ID/DE/RS
IR-14 Lessons-to-control closure	Verify corrective action changes controls and operating evidence before incident closure.	Continual improvement	Final classification	Lessons learned	Finding, retest, acceptance	C; H/M	NIST CSF ID.IM; ISO management-system improvement
IR-15 Independent ingestion suspension	Stop new or refreshed untrusted content entering an affected retrieval pipeline without relying on the user-facing application.	Data/ resilience control	Affected pathway and containment	Containment	Suspension test, queue state, index diff	P/C; H/H	NIST AI RMF MANAGE; OWASP LLM01/LLM08 context

7. Detection & Monitoring

7a. Logging requirements

Source	Required fields	Retention basis	Priority
Model gateway	trace, system/model/config, user/service ID, policy result, input/output protected reference, time	Risk, privacy, legal and investigation schedule	P1
Agent/tool broker	tool, arguments classification, authorization, approver, destination, result, rollback	Reconstruct consequential actions	P1
Retrieval	query, index/version, served document IDs/ranks, ingest/change actor, provenance	Reproduce context and poisoning path	P1
MLOps/data pipeline	model/data/prompt hashes, deployment, tests, lineage, release actor	Lifecycle plus response/recovery need	P1
Identity/secrets	principal, scopes, grants, token events, rotation/revocation, source	Security schedule	P1
Infrastructure/ network	endpoint, container, DNS, egress, process, cloud control-plane actions	Conventional IR schedule	P1
Human oversight	reviewer, information presented, decision, override, time, downstream action	Consequence and audit need	P1
Supplier	version/change, outage, incident notice, affected scope, evidence supplied	Contract and regulatory need	P2
User reports	reporter, trace/system, alleged effect, supporting material, disposition	Incident and privacy schedule	P2

Retention is not indefinite content capture. Minimize raw sensitive content, separate security metadata from content, restrict access, document legal holds and validate that deletion does not break required evidence.

7b. Detection logic

```
Rule: Unauthorized Agent Action
Source: Tool broker + registry
Logic: tool_call NOT IN approved_tools
       OR destination NOT IN approved_destinations
       OR transaction_value > approved_limit
Threshold: Any consequential match
False-positive risk: Low–Medium
Tuning: Bind policy to system version; treat emergency actions separately.
```

```
Rule: Indirect Instruction Influence
Source: Retrieval + agent plan + tool trace
Logic: retrieved_content contains instruction-like pattern
       AND plan/objective changes after retrieval
       AND sensitive_tool requested
Threshold: Correlated trace; block high-impact action pending review
False-positive risk: High
Tuning: Use semantic and structural signals; never alert on keywords alone.
```

```
Rule: Retrieval Integrity Drift
Source: Index lineage + approved corpus
Logic: served_document.version NOT IN approved_corpus
       OR provenance_status != verified
       OR privileged_change outside approved pipeline
Threshold: Any production match
False-positive risk: Low
```

```
Tuning: Account for approved emergency ingestion and delayed metadata.
```

```
Rule: Model Boundary Mismatch
Source: Runtime manifest + governance registry
Logic: runtime(model,prompt,index,tools,region) != approved_boundary
Threshold: One material mismatch
False-positive risk: Low–Medium
Tuning: Normalize supplier aliases and immutable versions.
```

```
Rule: Decision Trace Gap
Source: Business transaction + AI trace store
Logic: consequential_action = true AND trace_id is null
Threshold: Any match
False-positive risk: Low
Tuning: Exclude documented non-AI/manual paths with separate evidence.
```

7c. Indicators

Indicator	Confidence	Interpretation
Sensitive tool invoked immediately after untrusted content retrieval	High	Strong agent-hijacking signal; confirm authorization and intent
Same hidden instruction pattern across unrelated documents	High	Possible coordinated injection/poisoning
Runtime model/config differs from approved manifest	High	Change-control or supplier-drift incident
Output includes system prompt, secret pattern or unrelated private context	Medium–high	Potential disclosure; validate actual secret and source
Repeated denied tool calls followed by alternate tool sequence	Medium	Adaptive bypass attempt or faulty planning
Index document changed by unexpected identity	High	Retrieval compromise candidate
High-cost loop with no task progress	Medium–high	Resource exhaustion or runaway automation
Missing trace spans only for disputed actions	High	Evidence impairment or telemetry failure

7d. Telemetry architecture

Normalize events around `incident_id`, `trace_id`, `system_id`, `component_id`, `version`, `principal_id`, `tool`, `authorization`, `data_class`, `destination`, `event_time` and `evidence_hash`. Preserve time synchronization and immutable storage for high-value evidence. Build correlation across AI and conventional security telemetry; an AI incident may begin as identity compromise and end as model/tool abuse.

7e. Response health metrics

Metric	Definition	Decision supported
Trace completeness	Consequential actions with complete model/context/tool trace / consequential actions sampled	Can incidents be reconstructed?
Time to capability containment	Validated trigger to revocation/restriction of unsafe capability	Does authority work under pressure?
Boundary identification time	Trigger to confirmed affected system/component boundary	Can teams scope efficiently?

Unknown-impact backlog	Open incidents with material unresolved impact questions	Is management accepting uncertainty consciously?
Recovery test pass rate	Recovery candidates passing all required tests / candidates tested	Are restorations evidence-based?
Loop-closure rate	Closed incidents with verified GAISSF™ control improvement / incidents closed	Does response improve governance?
Recurrence	Incidents sharing previously identified causal condition	Did corrective action address root cause?

8. AI Incident Response Runbook

Preparation

Step	Action	Owner	Evidence	Decision point
1	Inventory systems, owners, dependencies and approved boundaries	AI governance lead	Reconciliation record	Is any consequential system unowned?
2	Pre-authorize command, containment, evidence hold and recovery decisions	Executive/ CISO	Authority matrix	Who can act outside business hours?
3	Define telemetry and evidence manifest requirements	IR/ architecture	Schema and test trace	Can one action be reconstructed?
4	Test tool revocation, endpoint isolation, model/index rollback and human-only mode	System owner	Exercise results	Which mechanism failed?
5	Maintain counsel-controlled notification and communications matrix	Legal/ privacy	Approved matrix	Which facts trigger legal review?

Identification

Step	Action	Owner	Evidence	Decision point
1	Create incident ID and appoint incident commander	SOC/IR lead	Case record	Formal incident or monitored event?
2	Record confirmed facts, potential impact, unknowns and notably absent evidence	Incident commander	Initial situation report	Executive escalation?
3	Freeze change where safe; capture exact runtime boundary	MLOps/ forensics	Manifest and hashes	Will containment destroy evidence?
4	Preserve prompts/protected references, retrieved context, tool actions, identities and approvals	Forensics	Evidence manifest	Legal hold or restricted access?
5	Classify provisionally using UAIF™ function and assign causality status	IR + system owner	Rationale	Priority and routing?

Containment

Step	Action	Owner	Evidence	Decision point
1	Protect people and stop continuing unauthorized actions	Incident commander	Decision/action times	Full shutdown or scoped restriction?
2	Revoke affected tools, tokens and	Security	Before/after state	Credential rotation required?

	destinations through independent controls	/IAM		
3	Quarantine poisoned data/index content without deleting evidence	Data/ RAG owner	Snapshot, provenance, hashes	What traces consumed it?
4	For indirect-injection or retrieval incidents, suspend the affected ingestion/refresh path and preserve queued/transformed content	Data/ RAG owner	Suspension record, queue snapshot, pipeline version	Can new poisoned content still enter through another source?
5	Restrict model endpoint or enable human-only workflow	Service owner	Change record	Essential service continuity?
6	Block propagation to downstream agents, queues and transactions	Architec t	Dependency/actions map	Which actions require reversal?

Eradication

Ste p	Action	Owner	Evidence	Decision point
1	Test competing hypotheses and reproduce under isolated conditions	IR/red team	Experiment record	Causality supported or still unknown?
2	Remove compromised content, credential, permission, dependency or configuration	Engineering	Remediation diff	Root condition eliminated?
3	Patch instruction/data separation and authorization, not only attack string	Security architect	Design/test evidence	Does adaptive testing still succeed?
4	Search portfolio for same supplier, component, tool or control weakness	Threat hunting	Hunt query/results	Enterprise-wide containment?
5	Obtain supplier remediation and affected-version evidence	Supplier manager	Supplier report	Accept, compensate or exit?

Recovery

Ste p	Action	Owner	Evidence	Decision point
1	Restore a version-bound package: model, prompt, index, tools and configuration	MLOps	Signed manifest	Exact approved package?
2	Run security, functional, safety and regression tests	Independen t tester	Raw results and failures	All mandatory criteria passed?
3	Reconcile affected historical decisions/actions and reverse where possible	Business owner	Review sample, reversals	Notify or remediate affected parties?
4	Stage service restoration with enhanced monitoring and limits	Service owner	Release approval	Expand, hold or roll back?
5	Obtain authorized residual-risk decision with expiry/conditions	Risk owner	Signed decision	Continued operation justified?

Lessons learned

Ste p	Action	Owner	Evidence	Decision point
1	Complete technical and governance causal analysis	Incident commander	Timeline and causal graph	What failed, was absent or was bypassed?
2	Finalize UAIF™-structured record, evidence	IR lead	Final incident	Classification supported?

	disposition and related incidents		package	
3	Convert findings into GAISSF™ control and assurance changes	AI governance lead	Corrective-action plan	Portfolio or policy change?
4	Retest corrective action and track recurrence	Independent assurance	Retest report	Close, extend or reopen?
5	Update AI-IRF™ playbook, exercises and competence requirements	IR lead	Updated runbook/training	What must change before next release?

Do not close solely because service is restored. Closure requires evidence disposition, authorized residual risk, corrective-action ownership and a verified plan to test operating effectiveness.

Communication templates

Internal situation report

```
Incident ID / system / accountable owner:
Time detected and current status:
Confirmed facts:
Potential impacts:
Containment completed:
Evidence preserved:
Unknowns and competing hypotheses:
Notably absent:
Decision required / authority / deadline:
Next update:
```

External holding statement

```
We identified an issue affecting [bounded service or function] on [date/time].
We [suspended/restricted] the affected capability and are investigating.
Confirmed impact: [facts only]. We have not confirmed: [material unknowns].
We are assessing affected parties and applicable notification obligations.
Updates will be provided through [authorized channel].
```

Scenario response card — compromised AI skill, plugin or connector

Apply this card when a packaged capability, extension, tool integration or connector is suspected of malicious behaviour or supply-chain compromise.

Window	Action	Lead	Completion evidence
First 15 minutes	Identify package, version, publisher, distribution channel and affected environment; open incident record; preserve the original artefact.	SOC / platform owner	Incident ID, artefact hash, acquisition source and time
First operational hour	Disable new installation and execution; revoke OAuth grants, tokens and service identities where exposure is plausible; block confirmed malicious destinations.	Incident commander / IAM	Before/after control state, revocation and block records
Same operational period	Determine installed population, permissions, accessed data, invoked tools, network destinations and persistence; perform controlled static and dynamic analysis.	IR / threat analysis	Scope statement, analysis report, evidence manifest and confidence
After	Remove persistence, rotate affected	Platform / endpoint /	Remediation record, clean validation

containment	secrets, clean or rebuild installations, validate dependent workflows and monitor for recurrence.	cloud teams	and hunt results
Before closure	Coordinate supplier/platform notice, assess affected-party and regulatory duties, publish indicators through approved channels, and convert findings into GAISSF™ controls and AI-IRF™ exercises.	Legal / supplier / governance	Notification decision, supplier response, control owner and retest date

Do not delete the package from the registry or affected host before preserving the distributed bytes, metadata and access history. If the package is hosted by a third party, issue an evidence-preservation request at the same time as the takedown request.

9. Audit & Assurance Checklist

9a. Documentation

Artefact	Purpose	Owner	Review
AI incident response policy	Scope, principles and authority	CISO	Annual/material change
System/owner registry	Mobilization and boundary context	AI governance	Continuous reconcile
Response authority matrix	Command, containment and recovery decisions	Executive/ CISO	Semiannual/exercise
AI incident taxonomy	Consistent event and incident records	IR lead	Annual/threat change
Evidence collection standard	Artefacts, custody, privacy and access	Forensics/ legal	Annual
Notification matrix	Role/jurisdiction/fact-specific escalation	Legal/privacy	Controlled update
Supplier response procedure	Joint investigation and evidence	Procurement	Contract/change
Recovery criteria	Safe staged restoration	System owner	Each material release
Exercise plan	Scenario and capability coverage	IR lead	Annual cycle
Lessons/corrective register	Governance loop closure	AI governance	Monthly

9b. Technical evidence

Area	Evidence	Method	Assessor expectation
Traceability	Input/context/output/tool/decision trace	Reproduce sample	Exact version and action joined
Containment	Kill-switch, token revocation, index quarantine	Controlled exercise	Works without affected component
Evidence integrity	Manifest, hashes, custody, access logs	Sample incident	Transformations transparent
Classification	Impact and causality rationale	Reperform sample	Facts separated from inference
Recovery	Package manifest and test results	Observe staged recovery	Criteria defined before approval
Supplier	Notice and evidence exchange	Tabletop and contract sample	Timely cooperation is enforceable
Loop closure	Finding to control to retest	Trace one incident	Governance changed and was verified

9c. Interview questions

Incident commander

12. Show how you establish one incident identity across security, MLOps, privacy and supplier workstreams.
13. What evidence would be destroyed by your default containment action?
14. Who can override you, and where is that authority documented?

Security architect / MLOps

15. Demonstrate how you identify and restore the exact model–prompt–index–tool package.
16. Can you revoke agent actions independently of the agent and model?
17. Show a trace connecting retrieved content to a consequential tool call.

AI governance / legal

18. How does provisional incident classification reach regulatory and executive decision-makers?
19. Which notification conclusions require facts that current telemetry cannot provide?
20. Show how a past incident changed a GAISSF™ control or assurance test.

CISO / independent assurance

21. Which incident scenario has not yet been exercised, and why?
22. How do you know systems outside the registry are not bypassing response readiness?
23. What evidence shows corrective actions operate rather than merely being closed in a tracker?

9d. Maturity

Level	Observable characteristics	Evidence
1 — Reactive	Generic IT response; AI context reconstructed manually	Incomplete traces, ad hoc decisions
2 — Defined	AI runbook, owners and baseline evidence requirements	Approved documents; uneven execution
3 — Operational	Tool isolation, version-bound rollback and UAIF™ records work	Exercises and incident samples
4 — Measured	Scope, containment, recovery and recurrence are quantified	Trend data and management decisions
5 — Adaptive	Incidents and red-team findings continuously change controls and assurance	Portfolio hunts, control retests, reduced recurrence

10. Research Insight ★

Research Insight

Finding: NIST CAISI, *Insights into AI Agent Security from a Large-Scale Red-Teaming Competition*, March 2026. **[T1 — primary government research summary]**

Summary: More than 400 participants conducted over 250,000 attempts against 13 frontier models in agentic scenarios. At least one successful hijacking attack was found against every target model, while success varied materially across models and scenarios.

Technical significance: Passing a static prompt-injection test does not establish agent security. Attack adaptation, model differences, tool context and scenario-specific authorization materially affect outcomes.

Validation of guidance: The finding supports IR-05 independent containment, IR-06 scoped agent identities, IR-10 recovery testing and IR-12 exercises. It also validates the playbook requirement to preserve tool/action traces and test adaptive attacks before recovery.

Operational lesson: Add one indirect-injection exercise in which an agent reads hostile external content and attempts a consequential tool action. Test whether authorization, detection, containment and evidence collection all work—not only whether the model refuses once.

Notably Absent: The competition does not establish a production incident rate, universal ranking of models or that every deployment is exploitable. Results are bounded to tested models, scenarios, defenses and attack effort.

11. Common Mistakes

Mistake	Why it matters / correction
Treating every hallucination as a security incident	Triage consequence, control deviation and malicious influence; preserve non-security quality events separately.
Treating AI IR as model rollback	Investigate data, prompts, retrieval, tools, identities, humans and downstream actions.
Containing before preserving volatile evidence	Protect people first, but predefine rapid capture so safety and evidence are not false alternatives.
Saving screenshots without authoritative exports	Screenshots lack full fields and provenance; retain logs, manifests and hashes.
Declaring prompt injection as root cause from one malicious string	Maintain competing hypotheses and reproduce under controlled conditions.
Letting the agent enforce its own kill switch	Use an independent control plane and identity.
Rotating only one exposed secret	Hunt for derived tokens, logs, prompt/context copies and downstream access.
Removing poisoned content without finding affected traces	Preserve the item and identify every system/decision that consumed it.
Rolling back model but not retrieval and prompt policy	Restore a compatible version-bound package.
Trusting supplier status without evidence	Require affected versions, timeline, preservation and validation results.
Over-collecting raw prompts indefinitely	Balance investigation with minimization, access control and lawful retention.
Waiting for confirmed causality before escalation	Escalate on plausible consequence while clearly labelling uncertainty.
Treating service restoration as closure	Complete evidence, residual-risk, corrective action and control retest.
Reporting what happened but not what did not	Record notably absent harm, propagation, attribution and evidence.

12. Quick Reference

12.1 Incident-to-first-action

Incident candidate	First containment	Critical evidence
Indirect prompt injection	Revoke sensitive tools; isolate retrieved	Full retrieval, plan and tool trace

	item	
Retrieval poisoning	Quarantine index/content version	Source, ingest actor, served traces
Training/tuning poisoning	Freeze pipeline and affected model promotion	Data lineage, commits, training config
Credential disclosure	Rotate/revoke and block destinations	Context, secret source, access logs
Agent runaway/resource loop	Stop tool loop and impose budget	Plan steps, tool sequence, resource metrics
Supplier model drift	Pin/restrict version; invoke supplier response	Before/after version and regression results
Harmful consequential output	Restrict workflow; review affected decisions	Decision trace and downstream actions
Evidence impairment	Restrict evidence-admin access	Audit logs and alternate telemetry
Compromised skill/plugin/connector	Disable package and revoke grants	Package bytes/hash, manifest, install population, permissions and destinations
Active indirect-injection feed	Suspend ingestion/refresh and isolate affected index	Source content, queue/transformation state, ingest identity, index/version and served traces

12.2 Risk-to-control

Risk	Controls
Unreconstructable event	IR-03 telemetry; IR-04 evidence manifest; IR-07 version binding
Unauthorized agent action	IR-05 independent containment; IR-06 scoped identities
Persistent poisoning	IR-08 provenance; IR-13 portfolio hunt
Unsafe recovery	IR-07 version binding; IR-10 recovery test pack
Supplier evidence gap	IR-09 supplier clauses
Missed notification analysis	IR-11 notification matrix
Recurring control failure	IR-12 exercises; IR-14 loop closure

12.3 Implementation checklist

#	Check	Yes	No	Partial
1	Every production AI system has owner and boundary.	☐	☐	☐
2	Incident command and containment authority are pre-assigned.	☐	☐	☐
3	Consequential actions have complete decision traces.	☐	☐	☐
4	Model, prompt, retrieval, tools and policy are version-bound.	☐	☐	☐
5	Tool/agent identities are scoped and independently revocable.	☐	☐	☐
6	Retrieval provenance and served-document references are logged.	☐	☐	☐
7	External-content ingestion/refresh can be suspended independently.	☐	☐	☐
8	Evidence manifests include hashes, custody and access.	☐	☐	☐
9	Supplier contracts support incident evidence and preservation.	☐	☐	☐
10	Notification matrix is controlled and current.	☐	☐	☐
11	The complete version-pinned release bundle has been rolled back and tested.	☐	☐	☐
12	Human-only or restricted mode is available.	☐	☐	☐
13	Adaptive indirect-injection scenario—including ingestion suspension—has been exercised.	☐	☐	☐
14	Portfolio hunt queries are prepared.	☐	☐	☐
15	Recovery criteria include security and safety tests, and lessons receive independent retest.	☐	☐	☐

12.4 Governance checklist

#	Check	Status
1	AI incidents are included in enterprise incident governance.	☐
2	UAIF™-structured classification and evidence function is defined.	☐
3	AI-IRF™ response and recovery responsibilities are assigned.	☐
4	GAISSF™ control improvement closes every material incident.	☐
5	Safety, privacy, legal and supplier escalation are integrated.	☐
6	Communications distinguish facts, inference and unknowns.	☐
7	Executives receive material incident and recurrence metrics.	☐
8	Exceptions and residual risk have authorized owners.	☐
9	Response competence is tested, not self-attested.	☐
10	Notably Absent findings appear in material reports.	☐

12.5 Audit readiness checklist

#	Evidence	Status
1	Current response policy and authority matrix	☐
2	Registry reconciliation and dependency map	☐
3	Sample complete AI decision trace	☐
4	Kill-switch and rollback exercise results	☐
5	Incident classification and causality rationale	☐
6	Evidence manifest and custody sample	☐
7	Supplier incident exercise or response evidence	☐
8	Notification decision with applicable facts	☐
9	Recovery approval and raw test results	☐
10	Finding-to-control-to-retest trace	☐

13. Assessment and Certification Readiness

This section addresses readiness only. It does not state that ODA3 certification, accreditation, licensed assessment delivery or regulatory recognition is operational. Exact ODA3 scheme and assurance-tier identifiers require authoritative current publications.

Framework / standard	Assessor focus	Common nonconformity	Evidence	Maturity indicator
ODA3 suite: GAISSF™ / UAIF™ / AI-IRF™	Governance preparedness connects to structured incident record, response and verified improvement	Framework names appear but evidence and loop closure are disconnected	Authority, UAIF™ record, AI-IRF™ actions, GAISSF™ retest	One traceable loop from trigger through improvement
NIST SP 800-61 Rev. 3 / CSF 2.0	Response integrated into governance, identify, protect, detect, respond and recover	Standalone runbook with no enterprise-risk integration	Profile, plans, incidents, exercises, improvements	Risk decisions change from response evidence
NIST AI RMF 1.0	AI risks mapped, measured and managed with governance throughout	Generic cyber response ignores socio-technical context	AI boundary, impact, measurement and treatment	Model/data/tool context is operationalized
ISO/IEC 42001:2023	AIMS operation, evaluation,	Incident process exists	Records, corrective action, audit and	Incidents drive management-system

	nonconformity and improvement support incident learning	outside AIMS	review	improvement
ISO/IEC 27001:2022	Incident planning, assessment, response, learning and evidence work for AI assets	AI systems excluded from ISMS asset/incident scope	ISMS scope, controls, incident evidence	Unified conventional and AI response
EU AI Act	Correct role/classification and applicable logging, post-market and serious-incident duties	Notification table copied without applicability analysis	Classification, logs, monitoring and counsel decision	Duties embedded in operational routing
MITRE ATLAS	Threat-informed hypotheses and testing for AI-specific attack paths	Technique label treated as proof of attribution	Observations, mappings, validation	Hypotheses improve detection and exercises

Examiner traps

24. **Screenshot evidence:** dashboards look complete but authoritative data and version context cannot be exported.
25. **Kill-switch theatre:** shutdown is documented but untested, depends on the affected system or requires unavailable approval.
26. **Recovery theatre:** service is restored after a patch without testing model, retrieval, tools, safety and affected decisions.

Tier 1 versus Tier 2 assessment

Dimension	Tier 1 — documentation	Tier 2 — technical/operating test
Authority	Inspect delegation and contacts	Trigger exercise and observe decisions
Telemetry	Review required schema	Reconstruct sampled action end to end
Containment	Review procedures	Revoke tool/identity and measure outcome
Evidence	Inspect manifest template	Reperform collection and verify hashes
Recovery	Review criteria	Restore staged package and replay tests
Improvement	Inspect corrective register	Trace finding to control and retest

Preparation timeline

Period	Objective
Weeks 1–2	Confirm scope, owners, authority, critical systems and evidence gaps.
Weeks 3–6	Implement telemetry, incident schema, containment and supplier procedures.
Weeks 7–10	Build recovery packs, notification routing and portfolio hunts.
Weeks 11–14	Exercise agent hijacking, poisoning, supplier and rollback scenarios.
Weeks 15–18	Remediate findings, retest controls and complete readiness review.

The timeline is illustrative and depends on portfolio size, evidence maturity, supplier constraints and existing response capability.

Primary References

- [NIST SP 800-61 Rev. 3: Incident Response Recommendations and Considerations for Cybersecurity Risk Management](https://csrc.nist.gov/pubs/sp/800/61/r3/final)

- [NIST Cybersecurity Framework 2.0](https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf)
- [NIST AI Risk Management Framework 1.0](https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf)
- [NIST AI 100-2e2025: Adversarial Machine Learning Taxonomy and Terminology](https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf)
- [NIST CAISI: Strengthening AI Agent Hijacking Evaluations](https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations)
- [NIST CAISI: Insights into AI Agent Security from a Large-Scale Red-Teaming Competition](https://www.nist.gov/blogs/caisi-research-blog/insights-ai-agent-security-large-scale-red-teaming-competition)
- [NIST CAISI Evaluation of DeepSeek AI Models](https://www.nist.gov/system/files/documents/2025/09/30/CAISI_Evaluation_of_DeepSeek_AI_Models.pdf)
- [NIST SP 800-86: Guide to Integrating Forensic Techniques into Incident Response](https://csrc.nist.gov/pubs/sp/800/86/final)
- [MITRE ATLAS](https://atlas.mitre.org/)
- [OWASP Top 10 for LLM Applications 2025](https://genai.owasp.org/llm-top-10/)
- [Regulation (EU) 2024/1689 — Artificial Intelligence Act](https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng)
- [Moffatt v. Air Canada, 2024 BCCRT 149](https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html)

```
Doc ID: ODA3-2026-07-CHT-SEC-011 (pending confirmation against file listing)
Topic: AI Security Incident Response Playbook
Version: 1.2
Classification: Public
Intended Audience: CISOs, Incident Responders, SOC Leads, Security Architects,
MLOps Engineers, AI Governance Leads, Privacy and Legal Teams, Auditors
Evidence Confidence: T1 for primary NIST, regulatory and tribunal sources;
T2–T3 for operational synthesis and deployment-specific guidance
Review Cycle: Quarterly recommended
Framework Cross-References: GAISSF™ v1.0; UAIF™ v1.0; AI-IRF™ v1.0;
NIST SP 800-61 Rev. 3; NIST CSF 2.0; NIST AI RMF 1.0;
NIST AI 100-2e2025; ISO/IEC 42001:2023; ISO/IEC 27001:2022;
MITRE ATLAS; OWASP Top 10 for LLM Applications 2025; EU AI Act
Source Cut-off Date: 17 July 2026
Known Evidence Gaps: Exact ODA3 control, severity, routing, role and phase identifiers
were not verified from authoritative framework documents for this edition. External
crosswalks are functional and require licensed standards text before audit-grade use.
Samsung case entry is T2 (corroborated public reporting), not a primary regulatory or
tribunal document.
Related Publications: CHT-SEC-009 (AI Security Evidence Collection Guide);
CHT-SEC-010 (AI Security Governance Model)
```