

KOMGIS

AI Incident Response Plan

Detailed plan for AI incident severity, containment, investigation, communication, corrective action and reactivation

Core incident response commitment

KOMGIS will treat harmful, wrong, unauthorized or suspicious AI behavior as an operational incident when it creates actual or reasonably foreseeable impact. The response priority is to stop ongoing harm, preserve evidence, determine scope, communicate accurately, correct the failed controls and return the system to service only after verification and authorized approval. Serious AI incidents must never be silently fixed or dismissed as normal model error.

Policy owner: Security / AI Governance / Technical Leadership

Version: 1.0

Effective date: 8 July 2026

Classification: Internal Policy

Document Control

Field	Value
Document title	KOMGIS AI Incident Response Plan
Document owner	Security / AI Governance / Technical Leadership
Version	1.0
Effective date	8 July 2026
Review frequency	At least annually and after high or critical AI incidents or material response-process changes
Approved by	[Executive approver / role]
Classification	Internal Policy
Applies to	KOMGIS employees, contractors, AI systems, client implementations, infrastructure and approved third parties
Related policies	Responsible AI and AI Management Policy; AI Security Policy; Data Privacy Policy

Status and interpretation note

This plan is an internal operational framework for responding to AI incidents in KOMGIS systems and client implementations. It expands the incident requirements in KOMGIS's Responsible AI and AI Management Policy and coordinates with KOMGIS's AI Security and Data Privacy policies. It is informed by the management-system approach of ISO/IEC 42001:2023 and relevant AI, privacy and cybersecurity obligations. Adoption of this plan does not by itself determine whether a specific event is legally reportable, a personal data breach, or a serious incident under applicable law. The incident commander shall obtain qualified privacy, legal or compliance review when notification or regulatory interpretation may be required.

Policy Statement

KOMGIS responds to AI incidents as failures of a controlled business and software system. The model may be the point at which harmful behavior becomes visible, but investigation must also examine identity, authorization, tenant isolation, retrieval, prompt and workflow versions, model and provider configuration, tool permissions, human oversight, validation, monitoring and change control.

The incident lifecycle is: detect and report; triage and classify; contain and preserve evidence; investigate and determine scope; communicate and assess notifications; correct affected outputs or actions; implement corrective controls; retest; reactivate through an approved gate; and feed lessons into the AI management system.

KOMGIS will use a single incident owner, a documented timeline and evidence-based decisions. The responsible team must distinguish confirmed facts from working hypotheses, protect client isolation during response, and avoid public or client speculation.

The minimum internal rule is that a high or critical AI incident shall not be closed by changing a prompt and resuming production without understanding the affected scope, addressing the failed control, verifying the fix and recording authorized reactivation.

Contents

1. Purpose
2. Scope
3. Relationship to KOMGIS AI Governance and Policy Hierarchy
4. Incident Response Language and Definitions
5. Incident Response Principles
6. Roles, Command and Accountability
7. Reportable AI Events and Incident Triggers
8. Detection and Reporting Channels
9. Initial Triage and Incident Declaration
10. AI Incident Severity Classification
11. Escalation and Response Targets
12. Incident Command and Coordination
13. Immediate Containment Strategy
14. AI Output and Workflow Containment
15. RAG and Cross-Client Data Containment
16. Agent, Tool and Unauthorized Action Containment
17. Credentials, Secrets and Provider Containment
18. Evidence Preservation and Incident Record
19. Investigation Methodology
20. AI-Specific Technical Investigation
21. Impact, Scope and Affected-Party Analysis
22. Root Cause and Contributing Cause Analysis
23. Communication and Notification
24. Client Communication
25. Privacy, Legal, Regulatory and Provider Coordination
26. Correction, Recovery and Remediation of Affected Outputs or Actions
27. Corrective and Preventive Action
28. Retesting and Verification
29. Reactivation Decision and Controlled Return to Service
30. Post-Incident Review and Lessons Learned
31. Records, Metrics and Trend Analysis
32. Exercises, Preparedness and Training
33. Exceptions, Nonconformity and Enforcement
34. Plan Review and Continual Improvement
- Appendix A. AI Incident Severity Matrix
- Appendix B. First 60 Minutes Checklist
- Appendix C. Evidence Preservation Checklist
- Appendix D. AI Incident Record Template

Appendix E. Communication and Notification Decision Matrix

Appendix F. AI Incident Playbooks

Appendix G. Root Cause Taxonomy

Appendix H. Reactivation Checklist and Approval Record

Appendix I. Incident Response RACI and Escalation Matrix

Appendix J. Required Incident Records

Appendix K. Alignment and Reference Documents

1. Purpose

This plan defines how KOMGIS prepares for, detects, reports, classifies, contains, investigates, communicates, corrects and recovers from incidents involving artificial intelligence systems or AI-enabled business processes.

The purpose is to ensure that AI incidents are handled consistently and proportionately, with priority given to stopping ongoing impact, protecting clients and affected persons, preserving evidence and restoring only verified, controlled operation.

- provide a common definition of AI incidents and reportable events;
- classify incidents by severity using impact, scale, data, autonomy and reversibility;
- define rapid containment for AI outputs, RAG, agents, tools, secrets and model providers;
- preserve prompts, retrieved sources, logs, approvals, tool calls and release evidence needed for investigation;
- determine affected systems, clients, users, data, outputs, decisions and external actions;
- coordinate security, privacy, legal, client and provider communication without speculation;
- correct affected outputs or actions and implement root-cause controls;
- require relevant retesting and authorized reactivation before a disabled AI capability returns to service;
- use incidents and near misses to improve risk assessments, tests, monitoring, training and policy.

Mandatory response boundary

An AI model, prompt or model-provider statement is not the incident response authority. Severity, containment, data scope, tool shutdown, client isolation, notification review and reactivation must be determined by authorized KOMGIS roles using system evidence and deterministic controls.

2. Scope

This plan applies to suspected and confirmed AI incidents involving systems developed, configured, integrated, deployed or operated by KOMGIS for internal or client use. It applies in development, test, staging, pilot, production, support and incident environments when there is actual or reasonably foreseeable impact.

- AI assistants, chatbots, voice interfaces and AI-enabled avatars;
- retrieval-augmented generation (RAG), enterprise search and document intelligence;
- document ingestion, OCR, parsing, chunking, embedding, extraction, classification and summarization;
- AI agents, LangGraph-style workflows, tool-using automation and long-running tasks;
- model gateways, hosted and local models, model routers and provider integrations;
- embedding services, vector databases, graph retrieval, caches and search indexes;
- AI-generated reports, messages, structured data, code, SQL, scripts, recommendations and decisions;
- MCP servers, plugins, connectors, APIs, webhooks and external tool interfaces exposed to AI systems;
- client-specific AI systems and internal AI tools, including AI-assisted engineering when company or client assets may be affected.

The plan applies to KOMGIS employees, contractors, developers, system owners, technical owners, human reviewers, administrators, support staff, project managers, security and privacy personnel, and approved third parties participating in an incident response.

Conventional security, privacy, service outage or business continuity incidents that involve an AI component may be managed under multiple KOMGIS response processes. One incident record should identify the lead process and the linked supporting processes so that evidence, decisions and communications remain coordinated.

3. Relationship to KOMGIS AI Governance and Policy Hierarchy

This plan is a supporting document under KOMGIS's Responsible AI and AI Management Policy. That policy requires immediate containment, evidence preservation, severity classification, root-cause investigation, corrective action and

approval before reactivation. This plan defines the operational procedure and records used to satisfy those requirements.

Document or record	Incident-response role
AI Incident Response Plan	Incident declaration, severity, containment, investigation, communication, corrective action, recovery and reactivation.
Responsible AI and AI Management Policy	AI governance, risk classification, lifecycle controls, human oversight, monitoring, incidents and continual improvement.
AI Security Policy	Security incident handling for prompt injection, access control, RAG, agent tools, secrets, output security and secure development.
Data Privacy Policy	Personal and client data scope, breach coordination, providers, retention, deletion and data requests.
AI System Inventory / Risk Assessment	System ownership, intended use, models, data, tools, controls, residual risk and impact assumptions.
Client contract / DPA / system requirements	Client-specific notification, cooperation, retention, data residency, service and approval obligations.

Where an applicable law, contract, approved system requirement or another KOMGIS policy imposes a stricter response or notification requirement, the stricter requirement applies. Potential conflicts must be escalated to the incident commander and the relevant privacy, legal or compliance owner before the response team assumes that an obligation does not apply.

4. Incident Response Language and Definitions

The words shall and must indicate mandatory requirements. Should indicates a recommended control that may be omitted only when the incident owner documents why it is not appropriate. May indicates an optional response choice.

Term	Incident response meaning
AI event	Observed AI behavior, alert, output, action or control condition that may require review but has not yet been declared an incident.
Suspected AI incident	An event with credible indications of actual or reasonably foreseeable harmful, wrong, unsafe, unauthorized or policy-violating impact.
AI incident	An event involving an AI system or AI-enabled process that creates actual or potential harm, client impact, security or privacy exposure, policy failure, contractual risk or loss of trust and requires coordinated response.
Serious / material incident	An incident whose impact, scale, data exposure, autonomy, affected persons or regulatory significance requires high or critical handling and enhanced review.
Incident commander	The person with authority to coordinate the response, assign workstreams, approve response priorities and maintain the incident operating rhythm.
Incident owner	The accountable person assigned to ensure the incident record, actions, decisions, follow-up and closure are complete.
Containment	Action that stops, limits or isolates ongoing impact while preserving evidence and enabling investigation.
Correction	Immediate action that fixes a specific affected output, record or external action. A correction does not necessarily remove the root cause.
Corrective action	Action that removes or reduces a root or contributing cause and is verified for effectiveness.
Recovery	Controlled restoration of business processing, data or service after containment and remediation.

Term	Incident response meaning
Reactivation	Formal authorization to return a disabled or restricted AI feature, model, retrieval source, agent or tool to operational use.
Affected output / action	An AI-generated result, recommendation, decision-support item, record change, communication or tool action within the incident scope.
Evidence set	Preserved prompts, context, sources, model and workflow versions, logs, approvals, tool calls, configurations and related records needed to reconstruct the incident.
Working hypothesis	A provisional explanation used to direct investigation. It must not be communicated as confirmed fact.
Residual risk	Risk remaining after containment and corrective controls are implemented and tested.

5. Incident Response Principles

5.1 Stop ongoing harm first

When harmful output, data exposure or unauthorized action may continue, KOMGIS shall contain the affected capability before completing the full root-cause analysis.

5.2 Preserve before modifying

Relevant evidence shall be captured before logs are rotated, indexes rebuilt, prompts overwritten, queues cleared or deployments replaced, unless immediate safety or confidentiality requires faster action. The reason for any unavoidable evidence loss shall be documented.

5.3 Security and permissions remain deterministic

The response team shall disable tools, block data access and enforce tenant boundaries in application or infrastructure controls. Asking the model to ignore a source, stop calling a tool or behave safely is not containment.

5.4 Severity follows impact, not model novelty

A simple extraction model may create a critical incident if it causes large-scale cross-client exposure. An unusual model response may remain low severity if it is caught before use and has no material impact.

5.5 One source of incident truth

The incident record shall contain the authoritative severity, current facts, affected scope, decisions, action owners, communications and reactivation status.

5.6 Facts are separated from hypotheses

Internal and external communication shall clearly distinguish confirmed facts, suspected scope and unknowns. The team shall avoid blaming a user, developer or provider without evidence.

5.7 Client isolation continues during response

Incident access, exports, screenshots, test datasets and evidence copies must preserve tenant and confidentiality controls. A response team shall not combine multiple clients' data merely for convenience.

5.8 Reactivation is a control gate

Business urgency does not replace evidence of containment, corrective control, testing, notification review and authorized risk acceptance.

5.9 Reporting is protected

Personnel must promptly report suspicious AI behavior and shall not be penalized for a good-faith report through approved channels.

5.10 Incidents drive improvement

High and critical incidents, and recurring medium incidents, shall update relevant risk assessments, tests, monitoring, architecture, provider review, training or policy.

6. Roles, Command and Accountability

Every declared high or critical incident shall have a named incident commander and incident owner. Medium incidents shall have an incident owner and may use a dedicated commander when multiple teams or clients are involved. Low incidents may be handled by the system owner or technical owner through the normal issue process if escalation criteria are not met.

Role	Minimum incident responsibilities
Executive leadership	Provide authority and resources; review critical impact; approve critical reactivation and material risk acceptance; support client or public communication decisions.
Incident commander	Set priorities; maintain severity; assign workstreams; approve containment direction; run coordination cadence; ensure decision and communication logs are current.
Incident owner	Maintain the formal incident record; track actions, evidence, affected scope, CAPA and closure; ensure post-incident review is completed.
AI system owner	Explain intended purpose and business process; identify affected workflows; coordinate manual fallback; approve residual risk within authority.
Technical owner	Lead architecture and technical investigation; disable features; preserve technical evidence; implement fixes; coordinate testing and rollback.
Security owner	Investigate attack or control failure; advise containment; coordinate credential rotation, forensics, vulnerability handling and security revalidation.
Privacy / data protection owner	Assess personal and client data scope, breach implications, data subject impact, evidence handling and privacy notification requirements.
Legal / compliance	Assess contractual and regulatory obligations, serious-incident implications, legal preservation and external communication review where required.
Project / client owner	Maintain client context, contractual contacts and approved client updates; coordinate client evidence or action requests.
Human reviewer / business SME	Assess output correctness and real-world impact; identify affected cases; support correction and verification.
Operations / support	Execute approved operational controls, collect reports, preserve support evidence and coordinate service restoration.
Vendor / provider coordinator	Open provider cases, preserve provider response IDs and notices, obtain relevant technical facts and track provider actions.
All personnel	Stop using clearly harmful output, preserve what they observed, report promptly and avoid independent client or public statements.

Command rule

The incident commander directs the response process but does not replace system, security, privacy or legal accountability. Required subject-matter approvals remain with the authorized role. During a critical incident, unresolved ownership ambiguity shall be escalated to executive leadership immediately.

7. Reportable AI Events and Incident Triggers

Personnel shall report AI events when there is a reasonable possibility that a model output, retrieval result, tool request, automated action or control failure has created or may create material impact. Reporting does not require proof of root cause.

- wrong or misleading output that was used, sent externally or incorporated into an important business process;
- dangerous, harmful or materially inappropriate recommendation;
- unsupported or fabricated claim presented as document-grounded or source-cited;
- wrong-client retrieval, cross-tenant data reference or unauthorized document access;
- confidential, personal, secret or restricted information exposed in an AI output, prompt, log or provider request;
- successful or plausibly successful direct or indirect prompt injection that changes data access, tool behavior or trusted workflow state;
- unauthorized agent action, tool call, record update, email, file share, payment-related step, deletion or external submission;
- human approval gate bypass, missing approval or action execution before the approval state was valid;
- repeated model or extraction failure above the accepted error threshold, especially when reviewers miss the errors;
- bias, unfair treatment or uneven performance with material impact on people or important decisions;
- unsafe fallback behavior after model, provider, retrieval or validation failure;
- unexpected model or provider behavior after a model alias, version, region, configuration or provider service change;
- monitoring, audit logging or traceability failure that prevents reconstruction of a material AI action;
- abnormal autonomous loops, repeated tool retries, excessive data retrieval, bulk export or denial-of-wallet/resource consumption;
- client complaint, support case, audit finding or provider notice indicating a possible AI control failure.

A minor formatting issue or non-sensitive internal error caught before use may be handled as a low-level event.

Recurrence, pattern, reviewer failure or evidence of broader scope shall trigger incident review even when each individual error appears small.

8. Detection and Reporting Channels

KOMGIS shall maintain practical reporting paths appropriate to its operating environment. A report may originate from automated monitoring, human review, a client, a provider, testing or any employee or contractor.

Detection source	Examples
Automated monitoring	Cross-tenant reference alerts, denied tool calls, approval bypass attempts, sensitive-data detection, abnormal retrieval, model error spikes, validation failures, cost anomalies and agent loops.
Human oversight	Reviewer rejection, source mismatch, unsupported citation, dangerous recommendation, unexplained model confidence or unexpected workflow action.
Client / user report	Wrong answer, privacy concern, unexpected record change, unapproved communication, cross-client content or loss of trust.
Security / privacy review	Prompt injection evidence, credential exposure, unauthorized access, breach indicator, abnormal logs or control nonconformity.
Testing / audit	Failed tenant-isolation, retrieval permission, tool scope, output validation, deletion, monitoring or reactivation test.
Provider / vendor notice	Model service issue, security event, data-processing change, model behavior change, region issue or availability incident that may affect KOMGIS systems.

Reports should include the system name, client or workspace if known, what was observed, when it occurred, the output or action involved, and whether the behavior is still ongoing. A reporter shall not delay escalation to gather perfect evidence.

- do not forward confidential incident evidence to broad email or chat groups;
- do not delete the prompt, conversation, document, output or screenshot simply because it contains harmful material;
- do not independently reproduce a suspected cross-client or security issue against production data;
- use the approved incident channel and restrict evidence access to personnel with a response need.

9. Initial Triage and Incident Declaration

The first qualified responder shall perform a rapid triage sufficient to decide whether an incident record is required, whether immediate containment is necessary and who must be escalated. Triage is not the full investigation.

- 1. Confirm the observation.** Identify the AI system, affected workflow and the specific output, retrieval, tool request or action that raised concern.
- 2. Ask whether impact is ongoing.** Determine whether the model, workflow, agent, batch job, queue or external action can continue producing additional impact.
- 3. Identify the most consequential plausible impact.** Consider confidentiality, client isolation, unsafe action, people, financial or legal effect and external communication.
- 4. Preserve the first evidence.** Capture relevant IDs, timestamps, visible output, source references and system versions before changes are made.
- 5. Choose provisional severity.** Use the severity matrix based on the highest credible impact and available evidence. Uncertainty shall not be used to downgrade a potentially serious incident.
- 6. Assign ownership and containment.** Create or update the incident record, assign an incident owner and initiate containment appropriate to the provisional severity.
- 7. Escalate notification review.** Notify security, privacy, legal or client owners when data exposure, unauthorized action, affected persons or contractual obligations may be involved.

Triage default

When the team cannot determine whether a suspected cross-client exposure, sensitive data disclosure or unauthorized AI action is contained, the event shall be treated at least as High until evidence supports a lower classification.

10. AI Incident Severity Classification

Severity is an internal response classification. It determines response urgency, leadership involvement, evidence and review rigor. It does not replace separate legal, privacy, cybersecurity or regulatory classification.

The incident shall be classified using the most consequential reasonably foreseeable impact supported by current facts. Severity may be increased or decreased as the investigation develops, but every change must be recorded with the reason and approver.

Severity	Typical impact	Required response posture
1 - Low	Minor, non-sensitive error caught before material use; no client or external impact; easily corrected; no evidence of broader control failure.	Correct or quarantine the output; record when recurring; owner review; normal engineering follow-up.
2 - Medium	Wrong internal summary or extraction caught in review; repeated quality failure; limited reversible client workflow impact; isolated monitoring or traceability gap.	Contain affected workflow; notify system and technical owners; review affected outputs; document cause and correction; monitor recurrence.
3 - High	Client-facing wrong information; confidential exposure; wrong-client retrieval; unauthorized tool or external action; approval failure; sensitive workflow or affected-person impact.	Immediate containment; incident commander/owner; preserve evidence; determine affected scope; security/privacy/legal/client review as relevant; formal corrective action; approval before reactivation.

Severity	Typical impact	Required response posture
4 - Critical	Large-scale or systemic cross-client leak; severe personal-data exposure; dangerous high-impact automated action; serious harm; major regulatory or contractual risk; uncontrolled privileged agent behavior.	Stop affected system or capability; executive, security, privacy/legal response; urgent scope and harm reduction; required notifications; formal RCA and CAPA; management approval before staged reactivation.

Severity factors include:

- actual or potential harm to people, safety or fundamental interests;
- client business, financial, legal, contractual or reputation impact;
- confidentiality, privacy and data sensitivity;
- unauthorized access, cross-client scope or integrity of official records;
- number of clients, users, records, outputs or actions potentially affected;
- autonomy, write capability, external communication and reversibility;
- whether the incident was caught before use or passed required human review;
- ability to reconstruct the incident and identify the affected version and time window;
- regulatory, contractual or serious-incident notification implications.

11. Escalation and Response Targets

The following are internal operational objectives intended to reduce response delay. They do not replace shorter legal, contractual or client-specific deadlines. The incident commander may require faster action based on impact.

Severity	Ownership target	Containment target	Escalation target	Status cadence
1 - Low	Owner assigned by next business day	Within 2 business days or normal issue priority	System owner / technical owner	As needed
2 - Medium	Owner assigned within 4 hours during business operations	Same business day; target within 8 hours	System owner, technical owner; security/privacy as relevant	Daily until stable
3 - High	Incident owner and commander within 30 minutes of confirmation	Immediate action; target meaningful containment within 4 hours	AI governance, security and relevant privacy/legal/client owner within 1 hour	At least every 4 hours while active, or agreed cadence
4 - Critical	Incident commander within 15 minutes	Immediate stop or strongest safe containment	Executive leadership, security, privacy/legal and AI governance within 30 minutes	Continuous coordination; formal updates every 1-2 hours during active critical phase

A response target may be missed because the incident is discovered outside normal staffing hours or because the required action is technically complex. The incident record shall state the actual timeline and the reason. Targets shall not be manipulated by delaying incident declaration or keeping severity artificially low.

12. Incident Command and Coordination

High and critical incidents shall use a structured coordination process. The incident commander shall establish a restricted incident channel or war room and identify the official incident record.

- **Incident header.** Incident ID, title, current severity, system, client scope, commander, owner and start time.
- **Current facts.** What is confirmed, what is suspected and what remains unknown.
- **Containment status.** What is disabled, restricted, quarantined or operating in safe/manual mode.
- **Action tracker.** Action, owner, deadline, status and evidence link.
- **Timeline.** Detection, reports, deployments, model calls, retrieval, tool actions, approvals, communications and major decisions.
- **Decision log.** Severity changes, scope decisions, containment tradeoffs, notification reviews and reactivation approvals.
- **Evidence index.** Where preserved logs, prompts, source IDs, screenshots, exports and provider records are stored.

- **Communication log.** Internal, client, provider, regulatory and public communications, including approver and exact send time.

One person shall facilitate status meetings and one person shall update the authoritative incident record. Technical discussion may occur in workstream channels, but material facts and decisions must be copied into the incident record.

Shift handover for an extended incident shall include current severity, known impact, active containment, unsafe actions to avoid, highest-priority unknowns, open communications and the next decision time.

13. Immediate Containment Strategy

Containment shall reduce ongoing impact as quickly as possible while preserving sufficient evidence for investigation. The response team should choose the least disruptive control that reliably stops the harmful path; however, service continuity shall not justify leaving a high-impact action or data exposure active.

A preferred containment hierarchy is:

1. Stop or block the specific external or sensitive action type.
 2. Disable the affected tool, write capability, workflow step or queue consumer.
 3. Disable or isolate the affected retrieval source, vector namespace, tenant path or document collection.
 4. Restrict the feature to safe read-only or draft mode with mandatory human review.
 5. Pin or roll back the model, prompt, workflow, configuration or release associated with the incident.
 6. Block provider egress or disable the model gateway route for the affected service.
 7. Disable the affected AI feature or system completely when a narrower control cannot be trusted.
- record who authorized the containment, the exact time and the affected scope;
 - record whether containment is reversible and the rollback path;
 - preserve relevant configuration before overwriting or rebuilding it;
 - avoid deleting suspicious source documents, prompts or malicious inputs before evidence preservation;
 - assess whether the same component or control is used by other clients or systems and requires wider containment.

Containment rule

Do not use a prompt such as “ignore this document,” “stop sending emails,” or “do not reveal client data” as the primary containment for a security, privacy or unauthorized-action incident. Disable or constrain the underlying retrieval, tool, permission or workflow control.

14. AI Output and Workflow Containment

When the incident involves wrong, misleading, unsupported or harmful outputs, KOMGIS shall stop downstream reliance and identify the population of outputs produced under the affected conditions.

- mark the specific output and related case or workflow as suspect or quarantined;
- stop automated delivery, report generation, case closure, structured export or downstream API transmission;
- pause batch jobs and queued tasks that use the same model, prompt, workflow, data source or configuration;
- identify the affected model, prompt/workflow version, deployment and approximate time window;
- query audit records for outputs generated under the same conditions and create an affected-output list;
- require additional human review before use of potentially affected outputs;
- correct or withdraw externally sent information when business, legal or safety impact requires it;
- preserve the original output and its disposition even after a corrected version is issued.

Where an AI output created an official record, the response team shall distinguish the original generated content, the human edits, the final approved record and any downstream changes. Correction must preserve traceability rather than silently replacing the historical record when audit or client requirements require evidence of the change.

15. RAG and Cross-Client Data Containment

RAG incidents require containment at the retrieval and authorization layers. A model shall not remain connected to a data path that may retrieve unauthorized source material while the team investigates whether the model will expose it.

- **Isolate the data path.** Disable the affected retriever, source connector, vector namespace, graph expansion path, cache or collection.
- **Preserve authorization context.** Capture the authenticated identity, tenant/workspace, user role, object permissions and retrieval filters applied at the time.
- **Preserve source traceability.** Record retrieved document IDs, chunk IDs, metadata, scores and source locations where available.
- **Block cross-tenant scope.** Apply server-side deny rules or temporarily restrict retrieval to a verified allowlist. Do not rely on the model to avoid unauthorized context.
- **Stop ingestion when poisoning is suspected.** Pause document ingestion, OCR, chunking and indexing for the affected source until the content path is reviewed.
- **Invalidate unsafe caches.** After evidence capture, invalidate caches that may retain unauthorized retrieval results or stale permission decisions.
- **Test adjacent systems.** Identify other clients, indexes or systems using the same retriever code, metadata mapping or authorization component.

A confirmed wrong-client retrieval or cross-client data exposure is normally High. Large-scale, systemic or repeated exposure, or exposure of highly sensitive personal or regulated data, may be Critical. The response team shall not downgrade severity merely because the model output did not reproduce the full source document.

16. Agent, Tool and Unauthorized Action Containment

An agent incident shall be contained at the tool execution, credential, queue or workflow policy layer. The response team shall assume that the model may continue requesting the action until deterministic controls prevent it.

- disable the affected write tool, action type or endpoint while keeping unrelated read tools available where safe;
- revoke or narrow the service credential or token used by the tool;
- stop workers, schedulers or queue consumers executing pending AI-requested actions;
- freeze or quarantine pending tasks and preserve their state before cancellation;
- block the action server-side by tenant, user, tool, target, parameter or workflow state;
- inspect approval state and verify whether the action executed before, after or without a valid approval;
- identify retries, alternate tool selection, action splitting or loops that may have bypassed intended controls;
- reconcile external side effects such as sent email, shared files, CRM changes, deleted records, submitted forms or financial instructions;
- preserve tool requests, validated parameters, authorization decisions, approval records and external response IDs.

Agent kill-switch requirement

Every production AI agent with write capabilities shall have a documented way to disable the agent, revoke or disable a tool, and stop queued or scheduled actions without requiring the model to cooperate.

17. Credentials, Secrets and Provider Containment

If a prompt, output, log, document, tool result or provider request may have exposed a secret or credential, the credential shall be treated as compromised until assessed. Containment must focus on revocation or rotation rather than asking recipients or the model not to use it.

1. Identify the secret type, owner, privilege, environment and systems that trust it.
2. Revoke, rotate or disable the exposed credential using an approved secure process.

3. Invalidate active sessions or downstream tokens derived from the credential where relevant.
4. Review access and usage logs for unauthorized use before and after the suspected exposure.
5. Determine whether the secret was sent to a model provider, stored in logs, embedded in documents or included in a retrievable index.
6. Contact the approved provider when provider-side exposure, retention or deletion support must be assessed.
7. Remove or redact active copies after the required evidence is preserved and verify that prompts, logs, caches and indexes no longer expose the secret.
8. Update secret scanning, prompt construction, logging or development controls that permitted the exposure.

Provider containment may include disabling a model route, pinning a known model version, changing a model alias, restricting data sent to the provider, switching to a pre-approved fallback provider or temporarily operating in manual mode. Provider changes remain controlled changes and must not create an unreviewed data residency, retention or security risk during the incident.

18. Evidence Preservation and Incident Record

The evidence set shall be sufficient to reconstruct what the system received, retrieved, generated, requested, approved and executed. Evidence collection must be proportionate and privacy-aware, but insufficient logging shall not be hidden by reconstructing a preferred narrative after the fact.

- incident ID, exact timestamps and timezone;
- AI system, environment, release, model provider, model or model alias and model configuration;
- prompt, workflow, agent graph or orchestration version;
- authenticated user or service identity and tenant/workspace identifier;
- request, trace, conversation, workflow, job, queue and provider response IDs;
- relevant AI inputs and system-controlled instructions, with sensitive handling restrictions;
- retrieved source IDs, document IDs, chunk IDs, metadata filters, scores and citations;
- model outputs, structured outputs, parser or validation results and final user-visible content;
- tool calls, requested and validated parameters, authorization decisions and external action results;
- human approval, rejection, edit and override records;
- application, security, infrastructure, model gateway, provider and integration logs;
- deployment commits, release records, configuration, feature flags and permission changes;
- alerts, screenshots, support reports, client reports and provider notices;
- containment actions and the system state immediately before and after each material action.

Evidence should be copied to a restricted incident location with access limited to the response need. Where integrity is important, KOMGIS should preserve original exports or immutable copies and record hashes, export commands or collection methods. Any legal hold, privileged legal analysis or regulator-specific evidence handling requirement shall be coordinated with Legal / Compliance.

The team shall preserve evidence before rebuilding an index, rotating logs, clearing a queue, overwriting a prompt, replacing an agent state store or deleting a malicious document unless immediate safety or confidentiality requires otherwise. The decision and expected evidence loss must be recorded.

19. Investigation Methodology

The investigation shall explain the incident path and the control failure, not merely identify the model output that became visible. The response team should use parallel technical, business-impact, privacy/security and communication workstreams when severity warrants.

1. **Define expected behavior.** Document what the system should have done under the intended purpose, permission model, workflow and approval requirements.

- 2. Build the factual timeline.** Correlate deployment, request, retrieval, model, tool, approval, alert and communication timestamps.
- 3. Identify the entry point.** Determine the user input, document, provider change, workflow state, permission, data condition or internal change that triggered the incident path.
- 4. Identify the control that should have stopped the impact.** Authorization, tenant filter, tool scope, approval gate, output validation, human review, monitoring or another deterministic control.
- 5. Determine why the control failed or was absent.** Inspect design assumptions, implementation, configuration, change review, test coverage and operational use.
- 6. Determine why detection did or did not work.** Review logs, alerts, reviewer evidence, support path and escalation behavior.
- 7. Establish scope.** Use evidence to identify affected systems, tenants, outputs, actions, data and time window.
- 8. Reproduce safely.** Create the smallest controlled test that confirms the failure without exposing additional client data or repeating harmful external actions.
- 9. Compare unaffected cases.** Identify what differs between affected and unaffected tenants, documents, users, releases or workflow states.
- 10. Maintain hypotheses.** Record competing explanations and actively seek evidence that disproves them before declaring root cause.

Production reproduction using real client data shall be avoided unless necessary, specifically approved and tightly controlled. Write tools, external communications and irreversible actions must remain disabled during reproduction unless the test environment provides a safe simulation.

20. AI-Specific Technical Investigation

The following investigation areas shall be selected according to the architecture and incident type.

Investigation area	Required questions and evidence
Model output / hallucination	Compare output to source evidence; inspect retrieved context; sample similar outputs; identify model and provider version; review prompt/workflow change; inspect sampling and fallback behavior; determine whether citations or confidence UI misled reviewers.
RAG / retrieval	Reconstruct identity and tenant context; inspect query filters, metadata, namespace and object permissions; review cache keys; verify source/chunk lineage; inspect graph expansion or hybrid retrieval; test deletion/permission propagation.
Prompt injection	Identify direct or indirect input; inspect hidden text, PDF layers, comments, OCR artifacts, webpages, emails or tool output; trace how untrusted content entered model context; inspect requested tools and data export behavior.
Agent / tools	Inspect tool inventory, allowlists, schemas, parameter validation, per-call authorization, approval state, retries, alternate tools, loop limits, memory and queued actions; reconcile external side effects.
Structured output / automation	Inspect schema validation, parser fallback, type conversion, allowed fields, default values, error handling and downstream mapping; determine whether malformed output was accepted as valid.
Provider / model change	Review provider notice, model alias resolution, version changes, region, configuration, rate limits, context limits, safety or filtering behavior and response identifiers; compare with a pinned or previous approved version.
Human oversight	Review what evidence the reviewer saw, system limitations communicated, workload, interface design, approval authority, edit/rejection behavior and whether automation bias or unclear source citations contributed.
Monitoring / traceability	Identify missing fields, disabled alerts, retention gaps, excessive redaction or inaccessible logs that delayed detection or scope determination.

AI investigation rule

“The model hallucinated” is a symptom description, not a root-cause conclusion. The investigation must explain why that output was accepted, delivered or acted upon and which control should prevent the same impact from recurring.

21. Impact, Scope and Affected-Party Analysis

The response team shall create an evidence-backed scope statement. Scope shall be updated as new evidence is found. Unknown scope is not the same as no impact.

- **Systems and versions.** Which application, feature, deployment, model, prompt/workflow, retriever, agent and tool versions were affected?
- **Time window.** What is the earliest plausible start, confirmed start, containment time and latest affected action?
- **Client scope.** Which tenants, workspaces, projects or clients are confirmed, suspected or specifically tested as unaffected?
- **User and person scope.** Which employees, client users, end users or affected persons received or were influenced by the output or action?
- **Data scope.** What confidential, personal, special-category, regulated, secret or public data was retrieved, sent, generated or exposed?
- **Output scope.** How many answers, summaries, classifications, extractions, reports or recommendations were produced under the affected conditions?
- **Action scope.** What emails, record updates, shares, submissions, deletions, API calls or other actions executed?
- **External recipients.** Who outside KOMGIS or the relevant client received affected content, files or actions?
- **Business impact.** What rework, financial, legal, operational, safety or trust impact occurred or may occur?
- **Reversibility.** Which outputs or actions can be corrected, recalled, restored or reversed, and which cannot?

The scope statement should use the labels Confirmed affected, Suspected affected, Tested not affected and Unknown. The basis for “tested not affected” must be stated. For cross-client incidents, the team shall test adjacent tenants or shared components using a controlled method rather than assuming the incident is isolated to the first reported client.

22. Root Cause and Contributing Cause Analysis

High and critical incidents require a formal root-cause analysis. Medium incidents require root-cause analysis when they recur, reveal a significant control weakness or affect a sensitive workflow. The analysis shall avoid reducing the explanation to individual error when system design or process conditions enabled the impact.

The record shall distinguish:

- **Trigger.** The immediate input, event or change that started the incident path.
- **Root cause.** The fundamental design, implementation, configuration or process condition that allowed the impact to occur.
- **Contributing cause.** A condition that increased likelihood, scale or severity but was not sufficient alone.
- **Detection gap.** Why monitoring, validation, human review or escalation did not detect the problem earlier.
- **Governance gap.** Why risk assessment, change control, testing, provider review, training or ownership did not prevent the weakness.

Root-cause methods may include five-whys analysis, fault trees, control mapping, change comparison or causal timeline analysis. The method must identify evidence for the conclusion and explain why competing hypotheses were rejected.

A root-cause conclusion such as “AI is unpredictable,” “user prompted badly,” “reviewer missed it,” or “provider issue” is insufficient unless the analysis also identifies the KOMGIS control, process or accepted risk that allowed the event to become a material incident.

23. Communication and Notification

Incident communication shall be timely, factual, audience-specific and coordinated. The incident commander shall identify the communication owner for internal, client, provider and external updates.

- state what is known and separate it from what remains under investigation;
- identify the affected system or workflow and the current containment status;
- describe confirmed or suspected impact without minimizing or exaggerating it;
- state what KOMGIS is doing and the next decision or update time;
- state any action the recipient must take, such as disregarding an output, resetting credentials or pausing a workflow;
- avoid technical speculation, blame or statements that a model/provider is responsible without evidence;
- avoid sharing exploit details, secrets, cross-client evidence or unnecessary personal data;
- record the exact communication, approver, recipients and send time.

High incidents should receive formal internal status updates at least every four hours while active, unless the incident commander defines a more appropriate cadence. Critical incidents should use continuous coordination and formal status updates every one to two hours during the active phase. These are internal coordination objectives, not legal notification deadlines.

The team shall not wait for complete root cause before communicating a material confirmed impact when the recipient must act or when notification obligations may apply. Early communication may state that investigation is ongoing and provide a scheduled next update.

24. Client Communication

Client communication shall be coordinated through the project/client owner and approved according to severity and contractual requirements. KOMGIS shall communicate directly and plainly enough for the client to understand what is affected and whether the client must take action.

An initial client notice, when required or appropriate, should contain:

- the incident reference or clear incident title;
- the affected service, feature or workflow;
- when KOMGIS detected the issue and whether it is contained;
- the confirmed impact and any suspected scope still under investigation;
- whether client data, specific records, AI outputs or external actions may be affected;
- what KOMGIS has disabled, restricted or corrected;
- what the client should do immediately, if anything;
- the next update time and client contact path.

When KOMGIS later determines that a prior statement was incomplete or incorrect, the update shall explicitly correct the earlier statement. Client notices shall not imply that AI error is normal or unavoidable as a reason to dismiss material impact.

The incident record shall preserve client questions, evidence supplied by the client, decisions on client-specific containment and every approved outbound update. Where multiple clients are affected, communication content may share a common factual core but must not expose another client's identity, data or system details.

25. Privacy, Legal, Regulatory and Provider Coordination

The incident commander shall ensure that specialist review occurs early enough to meet applicable obligations. The technical team shall provide facts and evidence; it shall not independently decide that an incident is not legally reportable when privacy, contractual, serious-incident or other regulatory requirements may apply.

Coordination owner	Incident assessment focus
--------------------	---------------------------

Coordination owner	Incident assessment focus
Privacy / data protection	Determine whether personal data is involved; categories and affected persons; likely consequences; controller/processor role; client instructions; breach coordination; data subject or authority notification analysis; evidence and retention needs.
Legal / compliance	Assess contracts, confidentiality, regulated-domain obligations, applicable AI serious-incident requirements, litigation or preservation concerns and approval of public or regulator-facing statements.
Security	Assess unauthorized access, prompt injection, credential compromise, attack path, vulnerability disclosure, infrastructure impact and security notification dependencies.
Client / project owner	Check contractual reporting windows, designated contacts, cooperation requirements, service commitments and client-specific containment.
Model / cloud provider	Open a case when provider behavior, data processing, retention, region, security or model version may be relevant; preserve case IDs and responses; request deletion or evidence support where appropriate.
Insurance / external counsel / specialist	Engage only when authorized and relevant to the incident, contract, insurance condition or regulatory exposure. Preserve confidentiality and approved communication channels.

Notification deadlines shall be tracked in the incident action log using the actual applicable law, contract and KOMGIS role. The plan intentionally does not assume that one notification deadline or regulatory regime applies to every AI incident. When multiple regimes may apply, the response team shall track each assessment and decision separately.

26. Correction, Recovery and Remediation of Affected Outputs or Actions

Correction addresses the real-world consequences of the incident. The response team shall identify affected outputs and actions and determine whether each item should be withdrawn, corrected, reprocessed, reversed, restored or accepted with documented rationale.

- issue corrected answers, reports, classifications or extracted fields and preserve the link to the original affected item;
- recall or correct external communications where practicable and identify recipients who may have relied on the wrong information;
- reverse unauthorized record changes using audited business processes;
- restore deleted or corrupted records from verified backups or source systems;
- cancel queued tasks, scheduled actions or automation chains associated with the incident;
- re-run document ingestion, retrieval indexing, extraction or batch processing only after the relevant control has been corrected and verified;
- remove malicious or poisoned content from active retrieval after preserving evidence and verifying the source and deletion lineage;
- support client-side correction, notification or workflow review when KOMGIS outputs have already been incorporated into client processes;
- operate in manual or non-AI fallback mode until controlled recovery is approved.

Correction is not a substitute for corrective action. For example, correcting a hallucinated report does not address missing source verification; reversing an unauthorized agent action does not address the absent approval gate; and deleting an exposed vector record does not address a cross-tenant retrieval filter failure.

27. Corrective and Preventive Action

Corrective and preventive action, referred to in this plan as CAPA, shall convert investigation findings into owned, testable improvements. High and critical incidents require a formal CAPA record. Medium incidents require CAPA when the cause is systemic, sensitive or recurring.

CAPA element	Required purpose
Immediate correction	Fix a specific affected output, record or action.
Containment control	Stop additional impact while the system remains under investigation.
Root-cause corrective action	Remove or materially reduce the root cause of the incident.
Systemic preventive action	Apply the lesson to similar systems, tenants, workflows, retrievers, tools or provider integrations.
Detection improvement	Add or improve logs, alerts, review evidence, anomaly detection or escalation paths.
Governance improvement	Update risk assessment, testing, release gates, inventory, provider review, training or policy.
Effectiveness verification	Define the evidence and time period used to confirm that the action works and does not create an unacceptable new risk.

Every CAPA item shall have an owner, due date, priority, affected systems, verification method and closure evidence. Overdue high-severity CAPA shall be visible in management review.

Corrective action rule

A prompt change alone is not an adequate corrective action for an incident involving authorization, client isolation, tool permissions, approval bypass, secret exposure or unsafe downstream execution. The failed security or workflow boundary must be corrected outside the model.

28. Retesting and Verification

Before reactivation, KOMGIS shall test the failed control and the adjacent paths that could create the same impact. The test set shall be derived from the incident scenario, root cause and affected architecture rather than using only generic quality evaluation.

- reproduce the original failure in a controlled environment and confirm that the corrective control blocks or safely handles it;
- run regression tests for intended behavior so the fix does not silently disable required functionality;
- repeat authentication, object authorization and tenant-isolation tests when data access was involved;
- repeat RAG retrieval permission, metadata filter, cache and cross-client tests when retrieval was involved;
- repeat direct and indirect prompt-injection tests when untrusted content influenced behavior;
- verify tool allowlists, parameter constraints, per-call authorization, approval gates, loop limits and kill switches for agent incidents;
- verify schema, parser and downstream validation for structured-output incidents;
- verify log completeness, alert generation and incident correlation for detection gaps;
- verify credential rotation and absence of active copies for secret exposure;
- run representative human review and source-verification scenarios when oversight was a contributing cause;
- perform limited smoke testing of the staged production configuration after deployment, without reenabling unsafe action scope prematurely.

High and critical incidents require documented test evidence and review by a person who was not the sole implementer of the critical corrective control. Failed tenant-isolation, authorization, approval-gate or secret-protection tests block reactivation.

29. Reactivation Decision and Controlled Return to Service

Reactivation is a formal decision that the affected AI capability may return to operation under defined conditions. A service being technically available does not mean it is approved for reactivation.

The following gates are mandatory before reactivation of a capability disabled because of a High or Critical incident:

1. Containment is confirmed and no known harmful output, retrieval or action path remains active.
2. The affected system, client, data, output, action and time scope is documented to the level appropriate for the incident.
3. Affected outputs or external actions have been corrected, quarantined, reversed or assigned a documented treatment plan.
4. The root cause is understood to a level sufficient to justify the selected corrective control. Remaining uncertainty is explicitly recorded.
5. Required corrective controls are implemented and configuration, code, prompt/workflow and provider changes are versioned.
6. Incident-derived tests and relevant security, privacy and quality tests have passed.
7. Monitoring and alerting are active and enhanced where the incident revealed a detection gap.
8. Required client, privacy, legal, regulatory and provider reviews or notifications have been addressed or have an approved action plan.
9. Rollback, kill-switch and manual fallback paths are verified.
10. Residual risk is documented and accepted by the authorized owner.
11. The required reactivation approvers have recorded approval and any operating restrictions.

Severity	Minimum reactivation approval	Additional requirement
1 - Low	AI system owner or delegated technical owner	Normal release or issue closure.
2 - Medium	AI system owner + technical owner	Documented verification; security/privacy review when relevant.
3 - High	AI system owner + technical owner + AI governance; security/privacy/legal approval when their domain is involved	Staged reactivation, enhanced monitoring and formal incident link.
4 - Critical	Executive leadership + AI governance + technical owner + security; privacy/legal/compliance as applicable	Formal reactivation record, staged rollout, explicit residual-risk acceptance and executive approval.

A preferred staged return is: dark deployment or disabled-state verification; internal/test traffic; selected pilot tenant or controlled user group; limited production percentage or limited action scope; full operation after monitoring confirms stability. The exact stages may vary by architecture, but the reason for skipping a stage in a high or critical incident must be documented.

Reactivation approval may impose restrictions such as read-only mode, mandatory human review, reduced retrieval sources, disabled write tools, lower rate limits or enhanced logging. These restrictions remain in force until explicitly removed through change control.

30. Post-Incident Review and Lessons Learned

High and critical incidents require a documented post-incident review. Recurring or systemic medium incidents also require review. KOMGIS should complete the review within ten business days after stable containment or reactivation unless the incident owner documents a reason for a longer investigation.

- what happened and the final severity;
- how the incident was detected and whether detection was timely;
- what impact occurred and the final affected scope;
- which containment actions worked, caused delay or created secondary risk;
- the root, contributing, detection and governance causes;
- whether client, privacy, legal, regulatory or provider coordination was timely and accurate;

- whether manual fallback, backup, rollback or kill-switch capabilities were effective;
- which corrective actions apply to other systems or clients;
- which risk assessments, test suites, monitoring, training or policies must change;
- what evidence should be preserved as a reusable incident test scenario.

The review shall focus on improving controls and operating processes while maintaining accountability. It shall not discourage reporting or treat the first person who noticed an AI failure as the cause of the incident.

31. Records, Metrics and Trend Analysis

Incident records shall be retained according to KOMGIS retention, legal, contractual and audit requirements. Access shall be restricted because incident files may contain sensitive client data, security details, prompts, outputs, secrets or legal analysis.

KOMGIS should monitor incident-response metrics proportionate to its AI portfolio:

- number of AI incidents by severity, system and incident category;
- mean and median time to acknowledge, declare, contain, scope and reactivate;
- percentage of high and critical incidents meeting internal response targets;
- detection source: monitoring, reviewer, client, test, provider or audit;
- root-cause categories and recurring control failures;
- number and age of overdue high-severity CAPA items;
- percentage of incidents with complete model, prompt/workflow, source and tool traceability;
- number of incidents where client isolation, approval or write-tool controls failed;
- reopened incidents or recurrence within the effectiveness-verification period;
- reactivation rollback rate and incidents caused by emergency fixes.

Metrics shall not be used to suppress reporting or reward teams for artificially low incident counts. A rise in reported low-level events may indicate better detection and reporting rather than worse system quality. Management review shall interpret the context and the severity trend.

32. Exercises, Preparedness and Training

KOMGIS shall maintain incident readiness through system documentation, disable mechanisms, contact lists, training and exercises. High-risk AI systems should have incident scenarios derived from their architecture and risk assessment.

- maintain current AI system owner, technical owner, security, privacy/legal and client escalation contacts;
- verify that high-risk systems have a documented disable, rollback or safe-mode mechanism;
- maintain access to relevant logs, model/provider identifiers, retrieval source IDs and tool audit records;
- train reviewers and support personnel to recognize prompt injection, wrong-client references, unsupported citations, unsafe actions and approval failures;
- run at least annual tabletop exercises for material AI services and additional exercises after major autonomy, tool, RAG or provider changes;
- include cross-client RAG exposure, unauthorized agent action, model/provider behavior change and harmful client-facing output in the exercise library;
- record exercise findings and treat unimplemented material findings as CAPA or risk-treatment actions.

Exercises shall test decision-making and response coordination, not only whether a technical team can restart a service. At least one scenario should require client communication and privacy/legal notification assessment so those paths are practiced before a real incident.

33. Exceptions, Nonconformity and Enforcement

A team may not pre-approve an exception that permits high or critical AI incidents to bypass severity classification, evidence preservation, required notification review or reactivation approval. Temporary incident decisions that deviate from this plan must be documented with the reason, risk, decision owner and expiry or review point.

Nonconformities include failure to report a serious AI event, intentional deletion of evidence, unapproved reactivation, bypass of required client or specialist review, misleading incident communication, or failure to complete material corrective action.

Violations may result in suspension of AI access, feature or tool privileges, corrective training, project escalation or other action consistent with KOMGIS contractual and employment arrangements. Good-faith reporting and reasonable incident decisions made with available facts shall be supported even when later evidence changes the conclusion.

34. Plan Review and Continual Improvement

This plan shall be reviewed at least annually and additionally when:

- a High or Critical AI incident occurs;
- incident response identifies a material coordination, evidence or reactivation gap;
- KOMGIS introduces materially greater AI autonomy, new write tools or controlled execution;
- a major RAG, vector, graph retrieval or client-isolation architecture changes;
- a new model provider, hosting region or important subprocessor is introduced;
- client contractual or legal notification requirements materially change;
- monitoring, incident tooling or organizational roles materially change;
- management review or internal audit identifies a recurring incident-response weakness.

AI Governance / Technical Leadership shall coordinate the review with Security and Privacy / Legal as relevant. Updated versions shall preserve version history and identify material changes to severity, escalation, evidence, communication or reactivation requirements.

Appendix A. AI Incident Severity Matrix

Severity	Impact test	Examples	Immediate response	Minimum leadership
Severity 1 - Low	No material client impact; non-sensitive; isolated; caught before use; easily reversible.	Minor formatting or non-sensitive internal error; citation display issue with correct source evidence; isolated model response rejected before use.	Correct/quarantine; recurring-pattern review; normal owner follow-up.	System/technical owner.
Severity 2 - Medium	Limited and reversible impact; repeated quality problem; internal or limited workflow effect; no evidence of significant exposure.	Wrong internal summary; incorrect extraction caught in review; repeated unsupported answer; limited monitoring gap.	Contain workflow; review affected outputs; document cause/correction; monitor recurrence.	System owner + technical owner; specialist review as relevant.
Severity 3 - High	Client-facing or sensitive impact; unauthorized retrieval/action; approval failure; confidential exposure; material affected-person or workflow effect.	Wrong-client retrieval; confidential data in output; unapproved email/tool action; material hallucination sent to client; sensitive workflow failure.	Immediate containment; incident command; evidence preservation; scope; notification review; formal CAPA and reactivation approval.	AI governance + system/technical owners + relevant security/privacy/legal/client owner.
Severity 4 - Critical	Large-scale/systemic exposure; severe personal-data event; dangerous autonomous action; serious harm; major regulatory/contractual risk.	Systemic cross-client leak; privileged agent performs uncontrolled destructive actions; severe high-impact automated decision/action; major data exposure.	Stop system/capability; executive response; urgent harm reduction and scope; required notifications; formal RCA/CAPA; staged reactivation.	Executive leadership + AI governance + security + technical owner + privacy/legal as applicable.

Classification rule: use the highest credible impact supported by current facts. When a suspected cross-client exposure, sensitive data disclosure or unauthorized AI action may still be ongoing and scope is unknown, classify provisionally as High or Critical until evidence narrows the risk.

Appendix B. First 60 Minutes Checklist

Time	Minimum response actions
0-5 minutes	Stop using the affected output/action; report through the incident path; capture the visible output, system and time; determine whether harmful activity is still ongoing.
5-15 minutes	Create incident record; assign provisional severity; identify system/technical owner; apply immediate tool, workflow, retrieval or feature containment; preserve first evidence.
15-30 minutes	Assign incident owner/commander for High/Critical; notify required security/privacy/legal/client owners; capture model, prompt/workflow, tenant, request and tool IDs; stop queues or batch actions if relevant.
30-45 minutes	Build initial factual timeline; identify likely affected time window and shared components; start affected-output/action query; confirm containment effectiveness; open provider case when relevant.
45-60 minutes	Publish first internal status: severity, known facts, suspected scope, containment, highest-priority unknowns, owners and next update time; decide whether client or formal notification review is immediately required.

Do not wait for perfect certainty

The first hour is for stopping impact, preserving evidence, assigning ownership and establishing a reliable fact pattern. Root cause may remain unknown. The incident record should say what is unknown rather than delaying declaration or containment.

Appendix C. Evidence Preservation Checklist

#	Evidence item	Complete / location
1	Incident ID, severity, detection time, declaration time and timezone.	☐ _____
2	AI system, environment, deployment/release and feature-flag state.	☐ _____
3	Model provider, model/model alias, region and material model parameters.	☐ _____
4	Prompt, system instruction, workflow, agent graph and orchestration version.	☐ _____
5	User/service identity, tenant/workspace/project and relevant roles/permissions.	☐ _____
6	Request, trace, conversation, job, queue and provider response IDs.	☐ _____
7	Relevant user input, uploaded file or untrusted content source.	☐ _____
8	Retrieved document IDs, chunk IDs, metadata filters, scores and citations.	☐ _____
9	Raw model output, parsed/validated output and final user-visible output.	☐ _____
10	Tool request, validated parameters, authorization decision, approval state and action result.	☐ _____
11	Human reviewer, edits, rejection/approval and override evidence.	☐ _____
12	Application, retrieval, model gateway, tool, security and infrastructure logs.	☐ _____
13	Model/provider notices, status events or support cases.	☐ _____
14	Deployment commit, configuration, permission and data-source change history.	☐ _____
15	Screenshots and client/support reports, preserving original timestamps where possible.	☐ _____
16	Containment commands, feature-flag changes, credential revocations and system state after each action.	☐ _____
17	Evidence access list, export method and integrity hash where required.	☐ _____
18	Legal hold, retention extension or privileged handling instruction where applicable.	☐ _____

Appendix D. AI Incident Record Template

Field	Incident record
Incident ID	_____
Incident title	_____
Date/time detected + timezone	_____
Date/time declared	_____
Current severity	1 / 2 / 3 / 4
Incident commander	_____
Incident owner	_____
AI system / feature	_____
Environment / release	_____
Model / provider	_____
Prompt/workflow/agent version	_____
Client / tenant scope	_____
Detection source	Monitoring / reviewer / client / support / test / provider / other
Current containment	_____

Field	Incident record
Confirmed facts	_____
Working hypotheses	_____
Unknowns	_____
Affected time window	_____
Affected outputs / actions	_____
Data categories involved	_____
External recipients / affected persons	_____
Security review owner	_____
Privacy/legal review owner	_____
Client communication owner	_____
Provider case / reference	_____
Next status update	_____

Required incident record sections: factual timeline; evidence index; action tracker; scope statement; communication log; root-cause analysis; CAPA; test evidence; residual risk; reactivation approval; post-incident review and closure decision.

Appendix E. Communication and Notification Decision Matrix

Audience	Trigger	Minimum content	Timing rule
Internal system/technical owners	Any declared incident	System, event, severity, containment, evidence owner, immediate action.	First response target for severity.
Security	Unauthorized access, prompt injection, secret exposure, tool abuse, cross-tenant risk, vulnerability or suspicious activity	Attack/control facts, logs, credentials, containment, reproduction restrictions.	Immediate for High/Critical; early for relevant Medium.
Privacy / data protection	Personal data, client data exposure, wrong-client retrieval, sensitive logs or provider data processing may be involved	Data categories, role, clients/persons, scope, likely impact, containment and provider path.	Early enough to assess applicable breach/notification deadlines.
Legal / compliance	Contractual, regulated-domain, serious-incident, major client or public impact may apply	Confirmed facts, severity, scope, actions, preserved evidence and communication plan.	Early for High/Critical or when legal obligations may be triggered.
Client	Client service/data/output/action affected or contract requires notice	Affected service, containment, known impact, client action, next update; no speculation or other-client details.	According to impact, contract and approved incident cadence.
Model/cloud provider	Provider behavior, processing, retention, region, outage or security may be relevant	Service/model identifiers, request IDs, timestamps, reproducible technical facts and requested support.	As soon as provider evidence or containment support is needed.
Regulator / authority	Qualified review determines reporting may be required	Approved factual report, role, system, impact, containment, corrective status and required supporting information.	Within applicable statutory deadline; tracked by Legal/Privacy/Compliance.

Audience	Trigger	Minimum content	Timing rule
Public / media	Material public interest or leadership decision	Approved concise facts, user/client action and contact path.	Executive + Legal/Compliance approval required.

Appendix F. AI Incident Playbooks

F.1 Unsupported or harmful client-facing output

1. Stop delivery and mark affected output suspect.
2. Identify model, prompt/workflow version, retrieved sources and affected time window.
3. Query similar outputs and reviewer rejection/edit patterns.
4. Correct or withdraw external content where required.
5. Investigate source grounding, citation correctness, validation and oversight.
6. Add incident-derived evaluation cases and approval/source-verification controls before reactivation.

F.2 Wrong-client RAG retrieval or cross-tenant exposure

1. Disable affected retrieval path or feature immediately.
2. Preserve tenant/user context, filters, source/chunk IDs and cache state.
3. Test shared retriever, metadata mapping, namespace and object authorization in a controlled environment.
4. Identify every tenant, query and output in the affected time window.
5. Coordinate privacy/security/client review; do not mix client evidence.
6. Correct authorization before retrieval, invalidate unsafe cache/index state and pass cross-tenant tests before staged reactivation.

F.3 Direct or indirect prompt injection

1. Disable affected tool/export path or untrusted content source.
2. Preserve malicious input, document, webpage, email, hidden text/OCR and resulting model/tool traces.
3. Trace trusted/untrusted prompt separation and server-side authorization decisions.
4. Identify data retrieved or tools requested/executed.
5. Remove or quarantine malicious content after evidence preservation; correct tool and retrieval boundaries.
6. Run direct/indirect injection regression and monitoring tests before reactivation.

F.4 Unauthorized AI agent action

1. Disable write tool or stop agent workers/queues.
2. Revoke or narrow tool credentials and freeze pending tasks.
3. Preserve plan/state, tool request, parameters, authorization, approval and external response IDs.
4. Reconcile and reverse external side effects where possible.
5. Investigate tool allowlist, per-call authorization, approval, retries, loops and alternate tools.
6. Verify kill switch, approval blocking and action limits before staged reactivation.

F.5 Secret or credential exposed to model/output/log

1. Rotate or revoke the secret immediately.

2. Invalidate sessions/tokens and review usage logs.
3. Determine every prompt, provider request, log, file, cache or index containing the secret.
4. Open provider case when provider-side processing/retention is relevant.
5. Remove active copies after evidence preservation.
6. Add secret scanning, prompt/log controls and verification before closure.

F.6 Unreviewed external communication sent by AI

1. Disable sending tool/action and stop queued messages.
2. Preserve draft, AI output, approval state, tool call and recipient list.
3. Identify and contact recipients when correction/withdrawal is required.
4. Investigate why approval did not block execution and whether alternate action paths exist.
5. Add server-side approval gate tests and action audit alerts.
6. Reactivate sending only after approval blocking and kill switch are verified.

F.7 Bias, unfair treatment or human-oversight failure

1. Pause affected decision-support or automation; require manual review.
2. Preserve affected cases, inputs, outputs, reviewer decisions and relevant evaluation data.
3. Identify affected groups/persons and downstream decisions or actions.
4. Assess data, model performance, workflow thresholds, UI evidence, reviewer workload and automation bias.
5. Correct affected cases and coordinate client/legal/privacy review where relevant.
6. Define representative evaluation, oversight and appeal/override controls before reactivation.

Appendix G. Root Cause Taxonomy

Root cause category	Examples
Architecture / design	Security or workflow boundary missing; excessive context; unsafe trust model; no safe fallback; unclear oversight design.
Authorization / client isolation	Missing object check; wrong tenant context; filter applied after retrieval; cache key leakage; shared credential scope.
Data / RAG / index	Wrong metadata; poisoned source; stale permissions; deletion not propagated; OCR/chunking error; graph expansion beyond scope.
Prompt / workflow	Unversioned change; trusted/untrusted mixing; ambiguous instruction; unsafe fallback; workflow state bypass.
Model / provider	Model alias/version behavior change; provider outage; context or response behavior change; region/processing issue.
Agent / tool	Excessive tool scope; missing per-call authorization; invalid parameter constraints; retry/loop abuse; approval bypass; unsafe memory/state.
Output / downstream validation	Malformed output accepted; schema fallback; model HTML/SQL/code executed unsafely; wrong field mapping; no confidence/source check.
Human oversight / UX	Reviewer lacked evidence, authority or training; interface encouraged automation bias; alert fatigue; workload made review ineffective.
Monitoring / logging	Missing correlation IDs; insufficient source/tool logs; alert threshold gap; retention too short; logs inaccessible during response.
Change / release management	Material model, provider, prompt, tool, permission or data-source change released without required risk/test review.

Root cause category	Examples
Operations / support	Unsafe manual change; over-privileged support access; incident evidence overwritten; unclear ownership or escalation path.
Governance / training	Risk not identified; system not inventoried; policy gap; provider review incomplete; training or competence insufficient.

Appendix H. Reactivation Checklist and Approval Record

#	Reactivation gate	Complete / evidence
1	Ongoing harmful output, retrieval or action path is contained.	☐ _____
2	Affected system, version, time window, client/data scope and affected outputs/actions are documented.	☐ _____
3	Affected outputs/actions are corrected, reversed, quarantined or assigned a treatment plan.	☐ _____
4	Root and contributing causes are understood to the level required by severity.	☐ _____
5	Corrective controls are implemented and versioned.	☐ _____
6	Original incident scenario no longer succeeds in controlled testing.	☐ _____
7	Relevant tenant-isolation, authorization, RAG, prompt-injection, agent-tool, approval and validation tests pass.	☐ _____
8	Monitoring, alerting and traceability are active and any detection gaps are addressed.	☐ _____
9	Credentials, provider configuration and data exposure have been addressed where relevant.	☐ _____
10	Required client, privacy, legal, regulatory and provider reviews/notifications are complete or tracked.	☐ _____
11	Rollback, kill switch and manual fallback are verified.	☐ _____
12	Residual risk and operating restrictions are documented.	☐ _____
13	Reactivation stage and monitoring period are defined.	☐ _____
14	Required approvers have approved reactivation.	☐ _____

Approval field	Record
Incident ID	_____
System / feature	_____
Severity	_____
Reactivation stage	Internal / pilot / limited / full
Operating restrictions	_____
Enhanced monitoring until	_____
Residual risk summary	_____
AI system owner approval	Name / date / signature: _____
Technical owner approval	Name / date / signature: _____
AI Governance approval	Name / date / signature: _____
Security approval, if required	Name / date / signature: _____
Privacy / Legal approval, if required	Name / date / signature: _____
Executive approval, if required	Name / date / signature: _____

Appendix I. Incident Response RACI and Escalation Matrix

Activity	Executive	Incident cmdr	Technical	Security	Privacy	Legal	Client owner
Declare incident and provisional severity	C	R/A	C	C	C	I	I
Immediate technical containment	I	A	R	C	C	I	I
Preserve technical evidence	I	A	R	R	C	I	I
Security investigation	I	C	C	R/A	C	I	I
Privacy/data scope assessment	I	C	C	C	R/A	C	I
Legal/regulatory assessment	I	C	I	C	C	R/A	I
Client comms	I	A	C	C	C	C	R
Root cause and CAPA	I	A	R	C	C	C	C
Incident-derived retesting	I	A	R	C	C	I	I
High reactivation approval	I	A	R	C	C	C	C
Critical reactivation approval	A	R	R	C	C	C	I
Post-incident review	I	A/R	R	C	C	C	C

R = Responsible for performing or coordinating the work. A = Accountable for the decision or completed outcome. C = Consulted. I = Informed. The incident commander may adapt the RACI for a specific incident, but required security, privacy, legal or executive approvals under this plan remain in force.

Appendix J. Required Incident Records

Record	Minimum requirement
Incident record	Required for declared Medium, High and Critical incidents; Low incidents when recurring or material pattern exists.
Severity decision and change log	Required for all declared incidents.
Factual timeline	Required for Medium, High and Critical; formal detailed timeline for High/Critical.
Evidence index	Required for Medium, High and Critical.
Containment and action log	Required for Medium, High and Critical.
Affected scope statement	Required for High/Critical; Medium when client/data impact exists.
Communication log	Required when client, provider, regulator or formal internal communication occurs.
Privacy/legal notification assessment	Required when personal data, contract or regulatory implications may exist.
Root-cause analysis	Formal for High/Critical; Medium when systemic, sensitive or recurring.
CAPA record	Formal for High/Critical; Medium when systemic, sensitive or recurring.
Retest / verification record	Required before reactivation of incident-disabled capability.
Reactivation approval	Required for High/Critical and any incident where a controlled AI feature was formally disabled pending investigation.
Post-incident review	Required for High/Critical and recurring/systemic Medium incidents.
Closure decision	Required for all declared incidents; records residual actions, open CAPA and final owner approval.

Appendix K. Alignment and Reference Documents

This table is a high-level design alignment. It is not a certification or legal-compliance statement and does not replace system-specific applicability analysis.

Plan area	ISO/IEC 42001 management-system theme	EU AI Act theme
Incident preparedness and ownership	ISO/IEC 42001 management-system roles, operation and documented information	Organizational accountability, quality management and post-deployment responsibilities for applicable AI systems.
Detection, monitoring and reporting	Performance evaluation, monitoring and operational control	Post-market monitoring, logging/traceability and serious-incident handling where applicable.
Containment and evidence preservation	Operational control and risk treatment	Cybersecurity, robustness, data governance, traceability and mitigation of system impact.
Investigation and impact assessment	AI risk assessment and system impact assessment	Risk management, affected persons and assessment of system consequences.
Communication and notification	Communication and documented information	Transparency, provider/deployer coordination and reporting obligations according to applicable role and law.
Corrective action and reactivation	Nonconformity, corrective action and continual improvement	Corrective measures, system updates, risk control and post-incident improvement for applicable systems.

Reference Documents Used to Develop This Plan

- KOMGIS Responsible AI and AI Management Policy - company-wide AI governance, risk, monitoring, incident and reactivation requirements.
- KOMGIS AI Security Policy - protection of AI systems against prompt injection, data leaks, unauthorized access, unsafe retrieval and tool use.
- KOMGIS Data Privacy Policy - personal and client data handling, breach coordination, provider, retention and deletion controls.
- ISO/IEC 42001:2023 - Information technology - Artificial intelligence - Management system.
- Regulation (EU) 2024/1689 of the European Parliament and of the Council - Artificial Intelligence Act.
- Applicable client contracts, data processing agreements, system requirements and KOMGIS security/privacy incident procedures.

Plan interpretation questions shall be escalated to Security / AI Governance / Technical Leadership and, where privacy, legal or regulatory interpretation is required, to qualified Privacy / Legal / Compliance review.