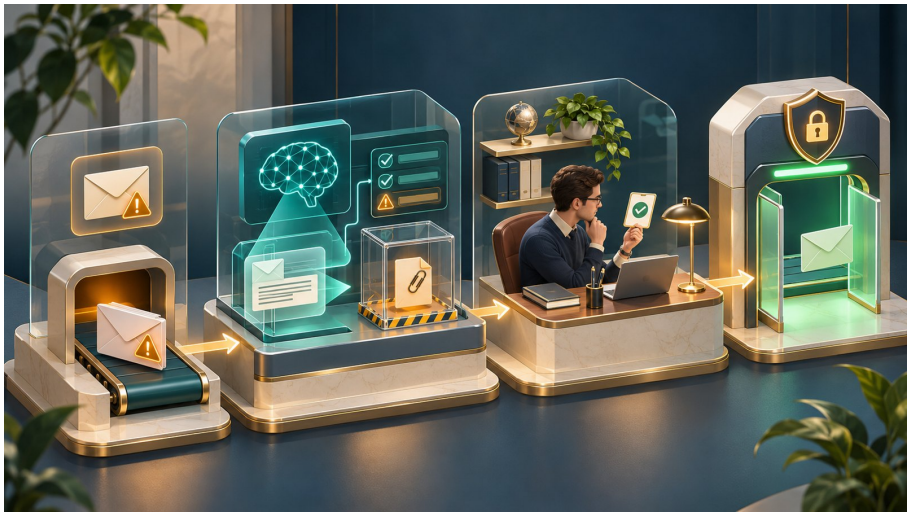
AIT TECHNOLOGY | BUSINESS IMPLEMENTATION SERIES

Using AI in cybersecurity: a controlled operating manual

Use AI to assist alert triage, investigate evidence and improve security operations while keeping data, permissions and response decisions under control.



AI-generated editorial illustration; fictional people and systems.

Tailored to your business

AIT Technology can customise these procedures and build bespoke software, AI workflows and integrations. Scope, testing, pricing and support are agreed before implementation.

Version 1.0 | 18 November 2026 | UK business context

Starting framework: approve company-specific scope and owners before adoption. No certification, guaranteed compliance or emergency response service is implied.

How to use this manual

Security/IT leads, managed service owners and managers considering AI-assisted defensive workflows.

Read the operating framework, assign department owners, select relevant sector scenarios, and adapt each procedure. The companion interactive PDF contains one reusable record per procedure. Save a separate copy for each case. Do not put passwords, API keys or recovery codes in these records.

Contents

Process at a glance

Operating framework

Department responsibilities

Business and sector scenarios

Step-by-step procedures

Records, approvals and maintenance

Custom implementation and sources

Process at a glance

01. Receive untrusted evidence	02. Minimise and validate	03. AI-assisted analysis	04. Human decision	05. Controlled action and audit

Use a named owner and recorded result at every handoff. Pause affected actions when safety, access or approval conditions are not met.

Operating framework

1. Use AI as an assistant to a defined defensive task

Useful candidates include summarising alerts, grouping related cases, drafting incident timelines, searching approved playbooks and explaining log fields to an analyst. These are potential uses to evaluate, not claims that a model will reliably detect attacks. Compare with existing rules and analyst practice.

AI should not independently decide that an account is safe, close a serious incident or authorise destructive response. Begin in read-only shadow mode. Keep conventional detection, backups and access controls active. AIT Technology can design a bespoke integration and review workflow around the tools you already use; licensing, coverage and response hours are agreed separately.

2. Treat attacker-controlled content as hostile input

An email, attachment, web page, support ticket or log field can contain instructions aimed at the model. An attacker may ask it to ignore policy, disclose secrets, mark an alert harmless or call a tool. The appearance of an instruction inside evidence does not grant it authority.

Keep evidence separate from trusted instructions and mark its origin. Restrict input types and size, minimise data and prevent automatic retrieval of arbitrary URLs or internal destinations. Use an authorised isolated analysis environment for suspicious attachments. A prompt saying "ignore malicious instructions" is useful guidance but cannot guarantee protection.

3. Put enforcement outside the model

The model receives only the evidence it needs and no credentials for broad security administration. A separate tool gateway validates action type, target, reviewer approval, rate limits and expiry. Keep default-deny permissions and a central stop control. Do not let generated text become a shell command, firewall rule or SQL statement without independent validation.

High-impact actions such as disabling users, isolating production devices, deleting messages, blocking network ranges or rotating shared secrets need the authorised response process and relevant business/safety review. A model recommendation is not an approval. Keep a reversible action and recovery path wherever practical.

4. Require traceable claims and measure failures

Ask for observed facts, source event references, uncertainty and suggested next checks in separate fields. An analyst should be able to trace a claim to an original event. A numerical confidence score generated by a model is not a calibrated probability of compromise. Missing logs must be reported as missing, not interpreted as no attack.

Evaluate benign and malicious samples, false positives, missed cases, cross-case information leakage and prompt-injection attempts. Measure analyst review time and factual accuracy. Retest after model, prompt, detection-source or action-tool changes. A successful demo on one phishing email does not establish detection effectiveness.

5. Protect security data and retain useful evidence

Security records can reveal personal data, secrets, business structure and attacker activity. Use an approved account/service, minimised inputs, appropriate retention and access controls. Redact credentials with a tested process; do not assume that asking the model to remove secrets prevents them reaching the provider.

Preserve original evidence in the approved evidence store and link derived summaries to it. Record collection time, source, transformations and access where required. AI summaries are working notes, not forensic proof. Specialist investigations may require additional evidence-handling controls and qualified support.

6. Introduce autonomy in explicit stages

Stage 1: read-only analysis on synthetic/historical authorised data. Stage 2: shadow analysis of approved live data, with no actions. Stage 3: analyst-approved bounded actions through a separate executor. Consider any further automation only after measured evidence, risk assessment and specific approval; do not infer it from a general wish for efficiency.

Keep fallback playbooks usable without AI. Pause the AI component on boundary failures, unexplained action, unsafe output or cost limits while preserving essential security monitoring. The response owner decides when to resume after a documented fix and retest.

Department responsibilities

Assign real names and deputies. A small company may combine roles, but must preserve accountability and approval boundaries.

Security / IT

Owner: Security operations owner

Routine: Review AI summaries against original events; maintain detection and tool boundaries.

Escalate when: Injected instructions, fabricated evidence or unauthorised action.

Evidence: Case references, reviewer decisions and action receipts.

Business operations

Owner: Service / site owner

Routine: Define the impact and safe window for containment actions.

Escalate when: Proposed action could interrupt production or safety-critical work.

Evidence: Containment and restart approvals.

Privacy and legal

Owner: Designated privacy lead

Routine: Approve categories of logs/messages and processing arrangements.

Escalate when: Sensitive evidence sent to an unapproved service or possible disclosure.

Evidence: Assessment, retention and reporting decisions.

HR and staff support

Owner: HR / support lead

Routine: Provide a clear reporting route and avoid treating model allegations as facts about staff.

Escalate when: A person may be wrongly accused or a personnel decision proposed.

Evidence: Human investigation and appropriate restricted case notes.

Leadership and suppliers

Owner: Risk owner / service manager

Routine: Approve scope, spend, coverage and supplier responsibilities.

Escalate when: No competent reviewer, service outage or proposal for wider autonomy.

Evidence: Service scope, acceptance evidence and authority matrix.

Business and sector scenarios

Planning scenarios, not assessments of named businesses. Add sector-specific safety, contractual and regulatory requirements before adoption.

Office / professional services

Priorities: Mailbox and identity alert triage.

Scenario: A malicious email asks the assistant to label itself safe.

Adaptation: Use read-only message access, preserve headers and require analyst verification of identity events.

Retail / e-commerce

Priorities: Website, admin and account anomalies.

Scenario: A model recommends blocking a large legitimate customer network.

Adaptation: Check transaction/context evidence, use narrow temporary measures only with approval and verify rollback.

Manufacturing / logistics

Priorities: Detection without disrupting physical operations.

Scenario: An isolation recommendation affects an operational controller.

Adaptation: Exclude uncontrolled OT actions; coordinate with engineers and safety owners using a tested incident plan.

Care / education / charities

Priorities: Protect sensitive user and beneficiary information.

Scenario: An alert includes confidential case content.

Adaptation: Minimise input, enforce access and retention, and involve the relevant privacy/safeguarding leads.

Software / managed services

Priorities: Tenant isolation, secrets and security-tool integrations.

Scenario: An assistant mixes evidence or executes an action across customers.

Adaptation: Enforce tenant boundary in storage, retrieval and execution; test it independently of model instructions.

Step-by-step procedures

01. AI-assisted suspicious-email triage

Owner: Security analyst

Start when: Reported suspicious message.

1. Collect the minimum approved metadata and text without opening an attachment on an ordinary workstation.

2. Label content untrusted; disable model action tools and arbitrary URL retrieval.

3. Request observed indicators, uncertainty and evidence references; do not ask only "safe or unsafe?".

4. Analyst checks sender/authentication context and relevant identity/security events using approved tools.

5. Record a human disposition and trigger conventional response when indicated.

Evidence: Original message reference, derived summary, analyst checks and decision.

Completion / handoff: Disposition has traceable evidence; model output alone cannot close the case.

02. AI-assisted alert correlation

Owner: Security operations lead

Start when: Multiple related security alerts.

1. Limit input to the authorised tenant/case and relevant time window.

2. Ask the model to group observations and cite original event IDs.

3. Validate the timeline, time zones, identities and gaps.

4. Check alternative explanations and missing sources.

5. Escalate according to impact and the normal incident plan, not a model confidence number.

Evidence: Event references, checked timeline and escalation rationale.

Completion / handoff: Analyst accepts the correlation with uncertainty recorded.

03. Approve a response action

Owner: Authorised incident lead + business owner where needed

Start when: AI suggests account/device/network containment.

1. Verify target identity, supporting evidence and business impact independently.

2. Choose the narrowest approved action with expiry/rollback.

3. Record explicit approval outside the model output.

4. Executor checks permission, target, approval and limits; return an actual receipt.

5. Verify the effect and monitor for unintended disruption.

Evidence: Approval, executor receipt and post-action test.

Completion / handoff: Authorised action verified or reversed; uncertain execution reconciled.

04. Prompt-injection attempt

Owner: Security lead + AI service owner

Start when: Evidence tries to redirect instructions, reveal secrets or invoke tools.

1. Do not follow the embedded commands; preserve a minimal restricted example.

2. Pause action-capable processing for the affected route if a boundary may have failed.

3. Check whether any data was disclosed or tool action occurred; use the incident process when needed.

4. Fix input, retrieval or executor boundaries, not only wording.

5. Retest with the original and varied adversarial examples before owner-approved resume.

Evidence: Attack sample reference, actual tool/data effects and regression tests.

Completion / handoff: Boundary holds in relevant tests and affected data/actions are assessed.

05. Incorrect AI incident summary

Owner: Incident lead

Start when: Summary invents facts, omits impact or assigns blame.

1. Mark the draft unverified and stop onward distribution.

2. Compare each material claim with original evidence.

3. Correct the timeline and affected scope; separate assumptions.

4. Review any decisions already based on the incorrect summary.

5. Update the evaluation set and review process.

Evidence: Corrected version, source references and decision review.

Completion / handoff: Approved human account replaces the draft and affected decisions are addressed.

06. Disable the AI component safely

Owner: AI/security service owner

Start when: Boundary failure, unexpected spend, unsafe behaviour or provider outage.

1. Disable model jobs/action access at the independent controller.

2. Keep conventional security monitoring and incident channels running.

3. Preserve approved logs and queued evidence without uncontrolled retries.

4. Assign human handling for urgent cases and evaluate the failure.

5. Require an authorised restart after the relevant tests and budget/security checks.

Evidence: Stop confirmation, fallback coverage and restart decision.

Completion / handoff: No further AI actions/spend beyond already in-flight bounded work; essential detection continues.

Records, approvals and maintenance

Use the companion forms to record company/case reference, owner and time zone, scope/trigger, verified facts, actions, approvals, completion and next review. Preserve original evidence in a restricted approved location; use references rather than copying sensitive material unnecessarily.

Company adoption checklist

- Define legal entities, sites, systems, information categories and exclusions.

- Assign accountable owners, deputies and escalation contacts, including agreed coverage hours.

- Agree retention, access, response objectives, budgets and stop conditions.

- Test representative workflows, exceptions, recovery and approval boundaries.

- Approve a version, train users, record evidence and set the next review date.

Urgent, significant and routine work

Urgent: suspected active compromise, sensitive-data exposure, fraud, safety risk or critical loss of service. Contact the named lead immediately through a trusted route. Significant: contained material disruption or weakness, with a coordinated owner and deadline. Routine: low-impact requests tracked to closure. Define actual response times and reclassify if impact changes.

The manual and AIT contact form are not emergency channels. Do not delay bank, safety or incident escalation while completing a form.

Custom implementation with AIT Technology

We can adapt this manual to your organisation and departments, map existing processes and build bespoke software, controlled AI workflows, integrations, approval queues and practical dashboards. Begin with discovery, agree scope and tests, implement reversibly, validate with users, and agree ongoing ownership and support separately.

Sources and limitations

Sources checked 18 November 2026. This is original AIT operational guidance. Suggested roles, frequencies, templates and workflows must be tailored; they are not verbatim official requirements. Linked guidance may change. Regulated or safety-critical use requires appropriate specialist assessment.

NCSC: guidelines for secure AI system development

https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

NCSC: incident management

https://www.ncsc.gov.uk/collection/incident-management

ICO: guidance on AI and data protection

https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/

NCSC: secure AI development - further guidance

https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development/guidelines

Discuss a tailored implementation: https://aittech.co.uk/contact/