Enterprise AI Security Assessment (EAISA) Methodology

A practical public methodology for assessing real-world enterprise AI systems

Public Release v1.0 – Jul 26

Published by Paladin Security Limited

Date: *24/07/2026*

Document Control

Name	Enterprise AI Security Assessment (EAISA) Methodology
Document Version	1.0 – Jul 26
Status	Public Release
Publication Date	24 July 2026
Published by	Paladin Security Limited
Authors	Yogesh Deshpande, Technical Director, Paladin Security Phil Dobson, Managing Director, Paladin Security
Classification	Public
Contact	contact@paladinsecurity.co.nz

Table of Contents

1. Legal Notice

1.1 Copyright and Permitted Use

The official EAISA Methodology and its current version are published at https://paladinsecurity.co.nz/research/eaisa/. Modified versions must be clearly identified as adaptations and must not be presented as official Paladin Security publications.

1.2 Disclaimer

This document provides general information and guidance for enterprise AI security assessment. It does not constitute legal, regulatory, privacy, or compliance advice. Organisations should obtain advice appropriate to their systems, operating environments, and legal obligations before applying the methodology.

AI systems are context-dependent and may change as models, prompts, configurations, integrations, data sources, permissions, agents, tools, and vendor services evolve. Use of this methodology does not guarantee the security, safety, reliability, compliance, or future behaviour of an AI system.

1.3 Authorised Use

The assessment concepts, test areas, and examples in this document are intended solely for lawful, authorised and controlled security-assessment activities.

Testing must only be performed against systems for which explicit authorisation has been obtained. It should follow agreed rules of engagement, applicable laws, organisational policies, safety constraints, and data-handling requirements.

1.4 External Review

Acknowledgement of an external reviewer indicates that the reviewer provided feedback on the methodology. It does not imply endorsement by the reviewer or any employer,

[1] https://creativecommons.org/licenses/by/4.0/

organisation, government agency, professional body, or other entity with which the reviewer is affiliated, unless expressly stated.

1.5 Version Statement

This version reflects the methodology as published on 24 July 2026. It may be revised as enterprise AI systems, model capabilities, agentic architectures, attack techniques, and regulatory requirements evolve.

2. External Review and Acknowledgements

Paladin Security thanks the following reviewers for providing external feedback on this methodology:

- ***Nathaniel Goza**, Principal Information Security Consultant, Accident Compensation Corporation (ACC)*
- ***Lana Tosic**, Senior Security Consultant, Accident Compensation Corporation (ACC)*
- ***Mark Knowles**, General Manager Security Assurance, Department of Internal Affairs (DIA)*

Reviewer acknowledgement indicates that feedback was provided during the development of the methodology. It does not imply endorsement by a reviewer or their employer or organisation.

3. Executive Overview

Enterprise AI is no longer confined to isolated experiments. It is being connected to internal knowledge, sensitive data, identity services, SaaS platforms, cloud infrastructure, code repositories, and business workflows - sometimes with the ability to make decisions or act on a user's behalf.

This changes the security question. Testing the model remains important, but it does not show whether the wider system preserves access controls, protects sensitive information, or limits what an agent can do. Those outcomes depend on the architecture around the model and the authority granted to it.

EAISA addresses the practical gap between high-level AI security guidance and the evidence required to assess a production enterprise system. It draws on established work from OWASP, NIST, MITRE, CSA, and ISO and can be applied alongside the legal, regulatory and security requirements relevant to the organisation, jurisdiction, and deployment.

The methodology is built on a clear premise: *enterprise AI security is a system security problem.*

An assessment must establish:

- What the system can access, retrieve, combine and disclose
- Whether identity, permission, and data boundaries remain effective
- What actions it can perform and whose authority it uses
- How prompts, retrieved content, tools, and integrations can influence its behaviour
- What evidence demonstrates whether sensitive information was exposed
- How identified weaknesses should be rated when conventional scoring does not capture the full business context
- Whether the system is ready to operate with real enterprise data, users, and workflows

EAISA examines the complete enterprise AI attack surface, including governance, architecture, components, dependencies, identity, data handling, retrieval-augmented generation, prompt injection, agents, tools, plugins, MCP integrations, applications, APIs, cloud services, monitoring, incident response, resilience, and resource abuse.

It is intended for the people responsible for putting AI into production and standing behind the result: security leaders, AI product owners, engineering teams, governance and privacy teams, penetration testers, and AI red teams.

EAISA complements existing standards by turning their principles into a repeatable assessment process. Its purpose is to provide defensible evidence that an enterprise AI system is ready for production and can be trusted with the information, access, and authority it has been given.

4. Assessment Approach

EAISA applies five key principles to determine how an enterprise AI system behaves within its operational environment.

4.1 Assess risk in context

A model does not carry the same risk in every deployment. Its risk depends on the data and systems it can reach, the decisions it supports, the actions it can initiate, and the consequences of error, manipulation, or compromise.

Assessment scope and testing depth should therefore reflect the system's actual purpose, architecture, users, integrations, and business impact.

4.2 Preserve permission boundaries

An AI system should not provide access to source content or protected information beyond the permissions granted by the underlying systems. This applies to direct responses and to information exposed through retrieval, summarisation, citations, metadata, search results, and generated outputs.

Testing should cover effective access across users, groups, roles, tenants, projects, repositories, records, documents, and indexed content. Particular attention is required where copilots, knowledge assistants or retrieval-augmented generation systems combine information from several sources.

Permission testing must consider both enforcement at the source and any additional access decisions introduced by ingestion pipelines, indexes, caches, vector stores, and application logic.

Inference limitation: *An AI system may derive new conclusions from information a user is legitimately authorised to access. It is not practical to prohibit all such inference. Assessment should instead determine whether reasonably foreseeable inference or aggregation could reveal protected facts, defeat an intended access boundary, or create unacceptable harm. The expected boundary should be defined by the organisation's data-classification, privacy, and authorisation requirements.*

4.3 Account for discoverability and aggregation

AI can make existing information substantially easier to locate and combine. Data that appears harmless in isolation may expose sensitive details when aggregated across documents, conversations, or repositories.

Assessment should examine indirect disclosure through filenames, titles, snippets, metadata, citations, stale indexes, cached responses, embeddings, and cross-document summaries. A valid source permission does not remove the need to consider what the system can reveal by combining authorised material at speed and scale.

4.4 Treat agents as delegated execution paths

An AI agent that calls a tool, invokes an API or changes a business record is exercising delegated authority. Its security depends on more than the content of its response.

Assessment should establish:

- Which identity authorises each action
- Whether permissions are limited to the invoking user and intended task
- How tool selection and parameters are validated
- When human approval is required
- Whether actions are logged and attributable
- Whether harmful or unintended changes can be stopped or reversed
- How rate limits, transaction limits, and other safeguards constrain misuse
- Whether the agent can be manipulated into abusing a more privileged service, tool, or identity

The agent should receive no more authority than the task requires, and untrusted content must not be able to expand that authority.

4.5 Assess more than just prompt injection

Prompt injection, jailbreaks, and indirect instruction attacks are important, particularly when a system processes untrusted content or can invoke tools. They represent only part of the enterprise attack surface.

A complete assessment must also examine identity, authorisation, retrieval architecture, data handling, application and API security, cloud configuration, integration security, monitoring, human approval processes, and governance. Prompt resistance cannot compensate for excessive permissions, unsafe tool access, or weak operational controls.

5. Relationship to Existing Guidance and Methodology Limitations

EAISA complements established AI security and risk-management resources, including the OWASP AI Testing Guide, NIST AI RMF, MITRE ATLAS, CSA AI Controls Matrix, ISO/IEC 42001, and ISO/IEC 23894. It also draws on relevant national guidance, including publications from the UK NCSC, ASD's ACSC, and New Zealand's NCSC. EAISA provides a practical structure for applying these sources when assessing enterprise AI systems.

EAISA is jurisdiction-neutral and should be used alongside the legal, regulatory, contractual, and sector-specific requirements applicable to the organisation and deployment. Depending on the context, these may include the EU General Data Protection Regulation, the EU AI Act, the UK GDPR and Data Protection Act 2018, Australia's Privacy Act 1988, New Zealand's Privacy Act 2020, and relevant United States federal, state, and sector-specific requirements. National security guidance, such as the Australian ISM and NZISM, may also apply. These examples are illustrative rather than exhaustive.

Detailed alignment with the OWASP AI Testing Guide and MITRE ATLAS is provided in *Appendix E*.

EAISA is not:

- A certification scheme or guarantee of AI safety
- Legal, regulatory, or privacy advice
- A complete responsible-AI, ethics, bias, or fairness assessment
- A substitute for secure architecture and software development
- Evidence that a system will remain secure after its models, prompts, data, permissions, tools, integrations, or operating environment change

Assessment results apply to the system, evidence, and conditions examined at the time. Material changes, newly identified attack techniques, or changes in provider behaviour may invalidate those results and require reassessment.

6. EAISA Enterprise AI Attack Surface Model

The EAISA Enterprise AI Attack Surface Model defines the areas that should be examined during an assessment and the relationships between them. It helps identify where data, instructions, identities, and delegated authority cross system boundaries or move between components with different levels of trust.

The layers are logical assessment domains, not a prescribed architecture or linear data flow. Their implementation and boundaries will vary between systems, and interactions may operate in both directions. Governance, privacy, legal obligations, monitoring, and incident response apply across the entire model rather than to a single layer.

Assessment should consider each layer individually and then examine how weaknesses combine across them. A control may appear effective within one component but fail when identity, retrieval, application logic, model behaviour, and tool execution are assessed as a connected path.

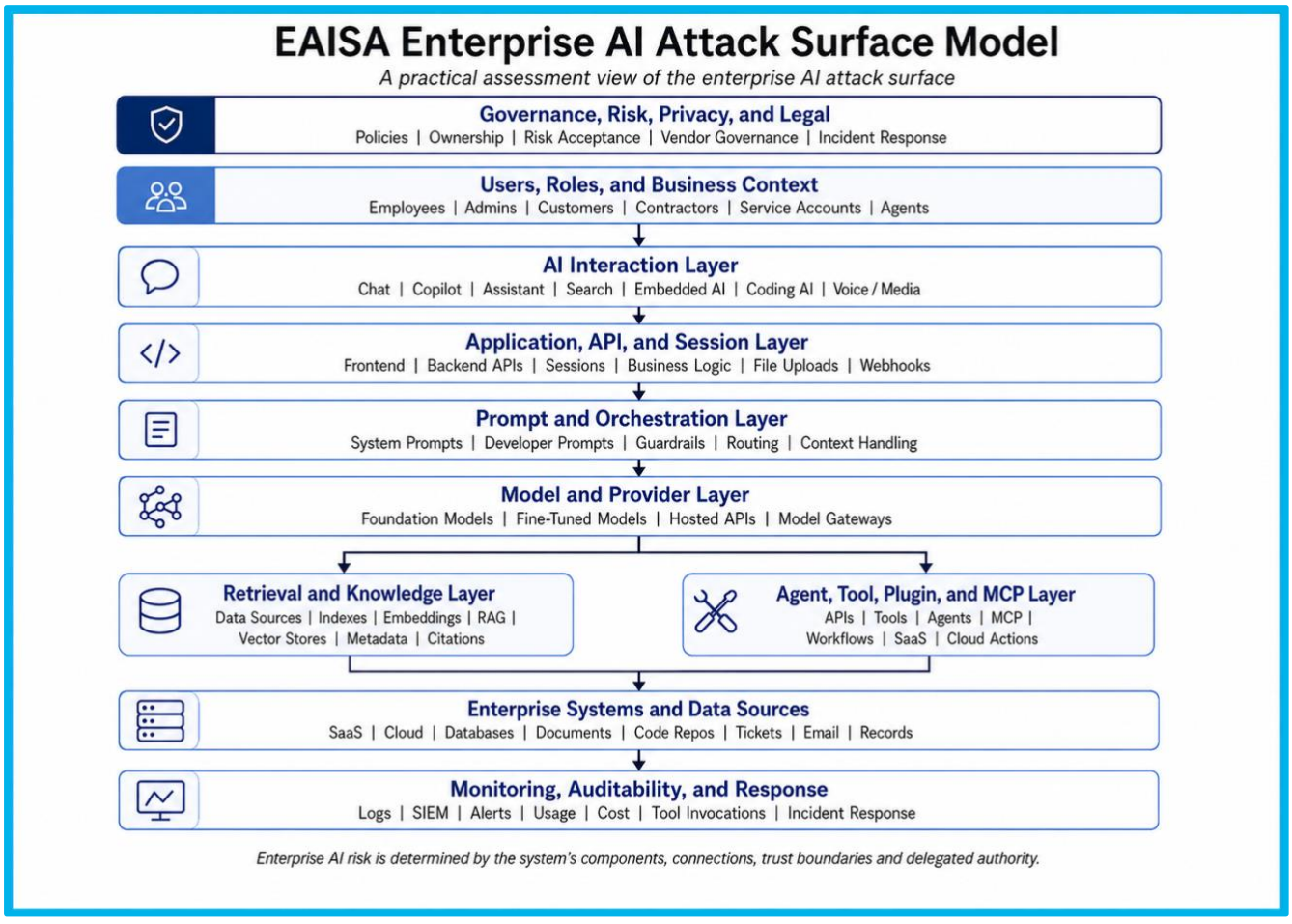

Figure 1: EAISA Attack surface model

Model Domains:

- *Governance, Risk, Privacy, and Legal:* Establishes ownership, policy, risk acceptance, regulatory obligations, and third-party oversight.
- *Users, Roles, and Business Context:* Defines who uses or administers the system, their authority and the business processes that depend on it.
- *AI Interaction:* Covers the interfaces through which users and other systems submit instructions and receive outputs.
- *Application, API, and Session:* Enforces authentication, authorisation, session security, input handling, and business logic around the AI capability.
- *Prompt and Orchestration:* Controls instruction hierarchy, context construction, routing, guardrails, and model or tool selection.
- *Model and Provider:* Covers the models, hosting arrangements, provider controls, and service terms on which the system depends.
- *Retrieval and Knowledge:* Governs ingestion, indexing, embeddings, retrieval, citations, and access to enterprise knowledge.
- *Agents, Tools, Plugins, and MCP:* Covers delegated actions performed through tools, APIs, workflows, and external services.
- *Enterprise Systems and Data Sources:* Represents the applications, repositories, and operational systems connected to the AI capability.
- *Monitoring, Auditability, and Response:* Provides evidence of system use, administrative changes, data access, tool actions, abuse, and security incidents.

These domains define the attack surface rather than a required architecture. Assessment procedures and evidence requirements are specified in the corresponding assessment modules.

7. Assessment Assurance Levels

EAISA uses assurance levels to define the scope and depth of an assessment. Selection should reflect the system's data sensitivity, access, autonomy, external exposure, business criticality, integration depth, and applicable legal or regulatory obligations.

Level	Name	Purpose
Level 1	AI Readiness Review	Reviews ownership, governance, inventory, intended use, architecture, and principal risks before detailed technical assessment.
Level 2	AI Configuration Assessment	Examines identity, platform settings, data handling, model and provider controls, connectors, agents, logging, and administrative security.
Level 3	AI Security Validation	Uses controlled hands-on testing to validate permission boundaries, retrieval behaviour, application and API controls, integrations, data exposure, and security logging.
Level 4	AI Adversarial Assessment	Tests realistic attack and abuse paths, including prompt injection, data extraction, retrieval manipulation, unauthorised tool use, and agent exploitation.
Level 5	Continuous AI Assurance	Repeats relevant validation and adversarial testing to identify configuration drift, access changes, model or prompt updates, new integrations, and security regressions.

Table 1: EAISA Assurance levels

Levels 1–4 provide progressively stronger evidence and may be combined where required. Level 5 adds ongoing assurance to the appropriate assessment depth; it does not replace the initial assessment. These levels describe assurance coverage, not organisational maturity or a guarantee of security.

7.1 Selecting an Assurance Level

The selected level should be proportionate to the harm that could result from unauthorised access, disclosure, manipulation, incorrect output, or unintended action. Multiple levels may apply where a system contains components with different risk profiles.

AI System Profile	Typical Characteristics	Indicative EAISA Assurance Level
Low-risk internal AI assistant	Internal use only, no sensitive enterprise data, no connected repositories, and no ability to perform actions	Level 1 or 2
Internal AI assistant with limited data access	Defined internal users, restricted data sources, and read-only operation	Level 2

AI System Profile	Typical Characteristics	Indicative EAISA Assurance Level
Enterprise copilot, search or RAG system	Retrieves information from business repositories, indexes, vector stores, or connected SaaS platforms	Level 3
System handling sensitive or regulated information	Processes personal, financial, health, legal, government, or commercially sensitive information	Level 3 or 4
Agentic or tool-enabled system	Invokes APIs, tools, plugins, MCP servers, or workflows that read or modify enterprise systems	Level 4
Externally accessible AI system	Accepts input from customers, partners, public users, or other untrusted sources	Level 4
High-impact AI workflow	Influences decisions or actions affecting finance, employment, legal rights, health, safety, security, or operations	Level 4
Frequently changing production system	Models, prompts, permissions, tools, connectors, or data sources change regularly	Level 5, incorporating the appropriate Level 2–4 activities
Mission-critical or regulated platform	Combines sensitive data, broad access, material business dependency, high-impact decisions, or delegated actions	Level 4 with Level 5 continuous assurance

Table 2: Indicative assurance-level selection

The final level should be agreed through scoping, threat modelling, and business impact analysis. Testing must also account for operational safety, system constraints, and the approved rules of engagement.

8. Testing Safety Classes

EAISA safety classes define the permitted level of interaction and potential operational impact during testing. They are independent of assurance levels: the assurance level determines what evidence is required, while the safety class determines how that evidence may be obtained.

Safety Class	Permitted Testing	Minimum rules-of-engagement requirements
Class A - Passive Review	Review of architecture, configuration, policies, logs, inventories, and documentation without interacting with the AI system.	Confirm scope, information-handling requirements, approved evidence sources, and access restrictions.
Class B - Controlled Query Testing	Approved prompts, retrieval queries, summarisation tests, and other read-only interactions using authorised accounts and data.	Define test identities, permitted data, query limits, logging requirements, and procedures for handling unexpected sensitive output.
Class C - Controlled Action Testing	Testing of agents, tools, plugins, APIs, and workflows using test records, mock services, non-production environments, or approved production-safe objects.	Specify permitted actions and targets, transaction limits, approval points, cleanup requirements, rollback procedures, and named system owners.
Class D - High-Risk Adversarial Testing	Attempts to bypass controls, manipulate agents, poison retrieval, chain weaknesses, or trigger complex tool behaviour.	Require explicit written authorisation, detailed test cases, isolated environments where practical, monitoring, stop conditions, escalation contacts, and tested recovery procedures.

Table 3: EAISA testing safety classes

8.1 Excluded Activities

Authorisation should identify the precise action, target, timing, safeguards, and accountable approver. General permission to conduct an assessment is not sufficient authorisation for an otherwise excluded activity.

Unless separately identified and explicitly authorised in the rules of engagement, testing must not:

- Modify or delete production data
- Contact real customers, employees, suppliers, or other unintended recipients
- Initiate live financial, legal, or operational transactions
- Disrupt production services or business processes

- Exceed agreed usage, cost, or rate limits
- Introduce persistent malicious content into production knowledge sources
- Access data outside the approved scope
- Disable security, safety, or monitoring controls
- Perform actions that cannot be reliably stopped or reversed

9. Assessment Lifecycle

The EAISA assessment lifecycle defines how an assessment is planned, conducted, and concluded. Detailed test activities and evidence requirements are specified in the assessment modules in *Appendix A*. The phases may overlap or be revisited as new components, dependencies or attack paths are identified.

9.1 Scope and Safety Planning

The assessment begins by confirming its objectives, systems, environments, user roles, and permitted testing activities. This phase establishes the assurance level, testing safety class, rules of engagement, data-handling requirements, prohibited actions, escalation contacts, and stop conditions.

9.2 System Discovery and Inventory

The assessment team develops a working view of the system's architecture, data flows, trust boundaries, identities, models, prompts, retrieval sources, agents, tools, integrations, and external dependencies. This phase produces the component inventory and AI Bill of Materials needed to identify what is in scope and how the system operates.

9.3 Threat Modelling and Test Planning

Threat modelling identifies credible abuse paths based on the system's access, authority, data sensitivity, exposure, and business function. The resulting scenarios are prioritised by potential impact and used to select the relevant assessment modules, test cases, and evidence requirements. The *AI Threat Scenario Catalogue in Appendix C* may be used to support this process.

9.4 Assessment Execution

The selected *Appendix A* modules are applied using the approved accounts, environments, and safety constraints. Testing may combine governance and configuration review, technical validation, and controlled adversarial activity, depending on the agreed assurance level. Evidence is recorded throughout and handled according to the assessment's security and privacy requirements.

9.5 Analysis and Reporting

Observed behaviour is evaluated against the system's intended security boundaries, applicable requirements, and business context. Findings should identify the affected component or workflow, supporting evidence, realistic impact, and proportionate remediation without unnecessarily reproducing sensitive information.

9.6 Remediation, Retesting and Ongoing Assurance

Remediation is prioritised according to risk, dependencies, and operational constraints. Retesting determines whether the corrective action addresses the original weakness without introducing new exposure. Systems subject to frequent changes or continuous assurance should repeat relevant discovery, threat modelling, and validation activities when models, prompts, data sources, permissions, tools, or integrations change.

10. Evidence Confidence Levels

Model outputs may vary between attempts, while system behaviour can change with user context, retrieved content, model versions, configuration, and provider updates. Findings must therefore record both what occurred and how confidently the behaviour has been established.

Level	Name	Meaning
E1	Observed	The behaviour was captured in at least one valid test but has not yet been reproduced or corroborated.
E2	Reproduced	The behaviour was reproduced under the same documented conditions.
E3	Corroborated	Additional evidence such as logs, configuration, code, retrieved content, or comparison across roles and sessions supports the observation and its cause.
E4	Systemic	The underlying design or control weakness has been established, and its affected users, data, components, or workflows have been reasonably determined.
E5	Impact Demonstrated	An authorised end-to-end scenario confirms that the weakness can produce a defined business or security impact.

Table 4: EAISA evidence confidence levels

The evidence level describes confidence in a finding, not its severity. E5 testing may be unsafe or unnecessary where impact can be established through lower-risk evidence, and the absence of an end-to-end demonstration should not be treated as proof that no risk exists.

Evidence should record the relevant user role, model, and version where available, system configuration, prompt and context, data state, tool actions, timestamps, and test conditions. Sensitive content must be minimised, redacted, and stored according to the agreed evidence-handling requirements.

11. AI-Specific Risk Rating Supplement

Established scoring methods should continue to be used where they adequately represent an application, API, cloud, or infrastructure vulnerability. Where AI-specific behaviour is not captured well by those methods, EAISA supplements the rating with an assessment of likelihood, impact, and validated controls.

Risk ratings must reflect the demonstrated behaviour and deployment context. They should not be increased merely because a system uses AI or reduced merely because exploitation is probabilistic.

11.1 Likelihood Factors

Likelihood considers whether a threat actor could reproduce and use the behaviour under realistic conditions:

- Required identity, role, access, and proximity
- Exposure to external users or untrusted content
- Reproducibility and consistency of the behaviour
- Number of attempts or interactions required
- Need for specialist knowledge, user participation, or insider information
- Complexity and dependence on other weaknesses
- Ability to automate or scale the attack
- Persistence of the condition across sessions or system changes

11.2 Impact Factors

Impact considers the credible consequences of successful exploitation:

- Sensitivity and volume of information exposed
- Number and type of affected users, tenants, repositories, or systems
- Authority and consequence of actions the system can perform
- Effect on data, model, retrieval, or business-process integrity
- Influence on consequential human or automated decisions
- Effect on downstream applications, operational systems, or safety
- Availability, resource-consumption, and financial impact
- Legal, regulatory, contractual, and privacy consequences

- Ability to reverse the action and recover safely

11.3 Risk Modifiers

The final rating should account for controls that materially change the likelihood or impact. These may include approval gates, transaction limits, source-system authorisation, isolation, monitoring, rollback capability, and effective incident response.

Only controls that were verified or supported by reliable evidence should reduce the rating. Planned controls and untested assumptions should be recorded as remediation rather than treated as existing protection.

Evidence confidence should be reported separately using the EAISA evidence levels. Limited evidence may reduce confidence in the assessment, but it does not necessarily reduce the underlying risk.

11.4 Severity Levels

Severity	Description
Critical	A feasible abuse path could cause widespread or catastrophic impact, such as compromise of highly sensitive information at scale, privileged control across systems or tenants, material operational disruption, or consequences affecting safety or mission-critical functions.
High	Successful exploitation could cause significant data exposure, cross-user or cross-tenant access, misuse of privileged actions, compromise of an important business process, or substantial downstream impact.
Medium	Exploitation could cause material but contained harm and depends on conditions such as specific access, user interaction, repeated attempts, or chaining with another weakness.
Low	The weakness has limited direct impact or practical exploitability but weakens a security boundary or increases exposure to future abuse.
Informational	No direct exploitable weakness was established, but the observation identifies an opportunity to improve governance, visibility, maintainability, or assurance.

Table 5: EAISA severity levels

Each rating should include a concise rationale covering the credible attack path, required conditions, affected scope, business impact, validated controls, and evidence confidence.

12. Go-Live Readiness Criteria

A high-impact enterprise AI system should not enter production until its material risks are understood, its controls have been validated, and accountable owners have approved the remaining exposure.

At minimum:

- No Critical finding remains unresolved.
- High findings have been remediated, reduced through validated controls, or formally accepted by an authorised risk owner.
- Permission boundaries have been tested using representative roles, data sources, and access states.
- Data handling, privacy obligations, retention, and provider use have been reviewed and approved.
- The system owner, intended use, models, providers, data sources, agents, tools, and integrations are documented.
- Actions performed by agents and tools are attributable to a user or approved workload identity.
- Consequential actions are subject to appropriate approval, transaction limits, and other safeguards.
- Audit records are sufficient to investigate retrieval, tool use, agent actions, and administrative changes without collecting unnecessary sensitive content.
- Monitoring, escalation, and AI-specific incident response procedures are operational.
- Emergency controls can disable or isolate affected models, agents, tools, and connectors without creating unsafe business outcomes.
- Usage limits, rate controls, and cost alerts reflect the expected workload and potential for abuse.
- Production change controls cover models, prompts, permissions, retrieval sources, indexes, agents, tools, and integrations.
- Material remediations have been retested, and accepted residual risks are recorded.

Go-live approval is a risk decision, not a guarantee of future safety. The decision should identify the accountable approvers, supporting evidence, accepted risks, and conditions requiring reassessment.

13. Conclusion

Enterprise AI can create real value, but the access and authority granted to it must be matched by evidence that its controls work.

EAISA gives organisations a practical way to obtain that evidence before weaknesses in identity, data access, retrieval, or delegated actions become business incidents. It helps teams make informed decisions about deployment, risk acceptance, and remediation based on observed system behaviour rather than assumptions.

Trust in an enterprise AI system is not established once. It must be maintained as the system, its data and its connections change.

Appendix A - Enterprise AI Security Assessment (EAISA) Matrix

The EAISA Assessment Matrix contains the modules used to plan and conduct an enterprise AI security assessment. Each module defines an assurance objective, illustrative assessment activities, expected evidence, and potential indicators of weakness.

The matrix is technology-agnostic. Modules should be selected according to the system archetype, architecture, data sensitivity, access, delegated authority, threat model, assurance level, and testing safety class. Applying every module to every system is neither required nor appropriate.

A.1 Using the Matrix

Assessment teams should document which modules are included, excluded, or modified during scoping. Module selection should reflect the system's business purpose, users, data sources, models, retrieval functions, agents, tools, integrations, deployment environment, and applicable obligations.

Assessment activities must remain within the agreed rules of engagement. Examples in the matrix are illustrative and may require adaptation, substitution, or omission where testing could affect sensitive data, production workflows, external parties, cost, or operational safety.

Results should be supported by evidence that records the relevant identity, system state, configuration, and test conditions. Findings should use the evidence confidence and risk-rating guidance defined in the main methodology.

A.2 Matrix Structure

Each assessment module uses the following structure:

Field	Description
Module ID	Unique assessment module identifier.
Assessment Area	The security domain being assessed.
Assurance Objective	The assurance objective for the module.
Illustrative Assessment Activities	Representative activities that may be performed.
Expected Evidence	Evidence expected to support observations or findings.
Potential Indicators of Weakness	Common issue types that may be identified.

For every module, also apply these interpretation rules:

The listed activities do not constitute a mandatory checklist. Equivalent procedures may be used where they provide sufficient evidence for the assurance objective. Potential indicators of weakness must be evaluated in context and do not automatically constitute reportable findings.

A.3 Governance, Accountability, and Administration

This domain assesses whether the AI system has defined ownership, approved uses, reliable inventory, controlled change, and appropriately restricted administration.

A.3.1 AI-GOV-01 - AI Ownership, Accountability, and Risk Acceptance

Field	Details
Assessment Area	Ownership and accountability
Assurance Objective	Establish whether accountable owners are assigned for the system, its data, technical operation, security, privacy, and business outcomes.
Illustrative Assessment Activities	Identify the business and technical owners; confirm responsibility for security, data protection, and operational decisions; review approval and escalation paths; determine who may accept residual risk; verify that ownership remains current.
Expected Evidence	Ownership records, responsibility matrix, risk register, approval records, governance terms of reference, and escalation procedures.
Potential Indicators of Weakness	No accountable system owner; conflicting responsibilities; unassigned security or privacy decisions; residual risks accepted without appropriate authority; ownership records no longer reflect the deployed system.

A.3.2 AI-GOV-02 - Approved Use and User Responsibilities

Field	Details
Assessment Area	Acceptable use and user governance
Assurance Objective	Determine whether approved, restricted, and prohibited uses are defined and communicated to relevant users.
Illustrative Assessment Activities	Review acceptable-use requirements; compare documented uses with the deployed capability; assess guidance for entering sensitive information and relying on generated output; review user notices, training, and reporting channels.
Expected Evidence	Acceptable-use policy, approved-use register, user guidance, training material, interface notices, and exception records.

Field	Details
Potential Indicators of Weakness	No defined use boundaries; sensitive data entered without guidance or approval; users are not told when verification or human review is required; deployed use differs materially from the approved purpose; no process exists for reporting unsafe behaviour.

A.3.3 AI-GOV-03 - System Inventory and Risk Classification

Field	Details
Assessment Area	AI asset management
Assurance Objective	Determine whether deployed and approved are inventoried, owned, and classified according to their access, authority, and potential impact.
Illustrative Assessment Activities	Compare the AI inventory with known applications, SaaS features, and cloud resources; verify coverage of models, agents, tools, and connectors; examine classification criteria; sample entries for ownership, lifecycle status and review dates.
Expected Evidence	AI system register, SaaS inventory, agent and connector registers, risk-classification criteria, ownership records, and review history.
Potential Indicators of Weakness	Unrecorded AI systems or features; undocumented agents or connectors; missing ownership; inconsistent classification; retired systems remain connected; inventory does not reflect production deployment.

This module assesses the governance process and coverage of the organisational inventory.

A.3.4 AI-GOV-04 - Change and Release Management

Field	Details
Assessment Area	Change control
Assurance Objective	Establish whether changes that may alter system behaviour, access, or risk are reviewed, tested, approved, deployed, and reversible.
Illustrative Assessment Activities	Review change controls for models, prompts, guardrails, permissions, agents, tools, connectors, retrieval sources, and indexes; inspect representative releases; examine testing and approval evidence; verify rollback and emergency-change procedures; determine how vendor-initiated changes are identified and assessed.
Expected Evidence	Change records, source, and configuration history, test results, approval records, deployment logs, release notes, rollback procedures, and vendor-change notifications.

Field	Details
Potential Indicators of Weakness	Production prompts changed without review; agents or connectors published outside change control; security testing omitted after a material change; model updates accepted without impact assessment; no reliable rollback path; emergency changes remain in place without retrospective review.

A.3.5 AI-ADMIN-01 - Administrative Access and Privileged Change

Field	Details
Assessment Area	Privileged administration
Assurance Objective	Determine whether administrative access is limited, attributable, and monitored according to the sensitivity of the capability being managed.
Illustrative Assessment Activities	Review administrative roles and assignments; identify who can enable models, publish agents, add connectors, modify prompts, access transcripts, change retention, or disable logging; assess separation between ordinary and administrative use; sample privileged actions and access reviews.
Expected Evidence	Role definitions, privileged account assignments, authentication policies, access reviews, approval records, audit logs, and administrative change records.
Potential Indicators of Weakness	Excessive or standing administrative access; shared administrator accounts; privileged actions cannot be attributed; sensitive transcripts are broadly accessible; users can publish organisation-wide agents without approval; logging or retention can be disabled without detection.

Where supported and proportionate, the assessment may also consider separate administrative identities, strong authentication, time-bound privilege, and approval for sensitive changes. These should not be presented as universal product requirements where the platform does not support them.

A.4 Data Security

This domain assesses how information is collected, transferred, processed, stored, and deleted across the AI system. Privacy obligations are addressed separately; these modules focus on the technical controls protecting enterprise data.

A.4.1 AI-DATA-01 - Sensitive Data Handling

Field	Details
Assessment Area	Sensitive data protection
Assurance Objective	Determine whether sensitive information is identified and protected throughout AI processing according to its classification and approved use.
Illustrative Assessment Activities	Trace sensitive data through prompts, uploaded files, retrieved context, model requests, responses, tool outputs, and logs; review access and transfer controls; examine redaction, masking, or filtering where used; verify encryption and key management for relevant data flows and stores.
Expected Evidence	Data-flow diagrams, classification records, provider configurations, access controls, encryption settings, redacted request, and response samples, tool logs, and data-handling procedures.
Potential Indicators of Weakness	Sensitive information sent to an unapproved model or provider; data exposed to users without the required access; classification lost during processing; secrets included in prompts or logs; sensitive content stored or transmitted without appropriate protection.

A.4.2 AI-DATA-02 - Data Minimisation

Field	Details
Assessment Area	Data minimisation
Assurance Objective	Establish whether the system processes and retains only the information reasonably required for its approved purpose.
Illustrative Assessment Activities	Inspect prompt payloads, retrieved context, API fields, uploaded content, and logging; determine whether full records or documents are used where limited fields or excerpts would be sufficient; review whether data is duplicated across components without a defined need.
Expected Evidence	Data mappings, prompt and API samples, retrieval configuration, logging design, field-selection logic, and documented processing purposes.
Potential Indicators of Weakness	Entire records sent when limited fields are sufficient; excessive document context provided to the model; sensitive attributes included without a defined purpose; duplicated data retained across logs, caches and provider services.

Data minimisation should be assessed against the system’s purpose and accuracy requirements. Reducing context is not beneficial if it causes unsafe or materially misleading behaviour.

A.4.3 AI-DATA-03 - Prompt, Response, and Content Storage

Field	Details
Assessment Area	AI data storage
Assurance Objective	Determine whether prompts, responses, transcripts, uploaded files, and tool outputs are stored with appropriate access, isolation, and security controls.
Illustrative Assessment Activities	Identify each storage location; review access by users, administrators, support personnel, and providers; examine tenant and environment isolation; assess exports, backups, caches, and temporary files; verify whether stored content is encrypted and auditable.
Expected Evidence	Storage architecture, access-control records, encryption configuration, administrator permissions, transcript settings, backup design, audit logs, and redacted content samples.
Potential Indicators of Weakness	Broad access to transcripts or uploads; production conversations accessible from non-production environments; sensitive tool output stored in general-purpose logs; unprotected exports or temporary files; administrator access is not recorded.

A.4.4 AI-DATA-04 - Embedding and Vector Data Protection

Field	Details
Assessment Area	Embedding and vector store security
Assurance Objective	Determine whether embeddings, associated metadata, and vector stores preserve required access and isolation boundaries.
Illustrative Assessment Activities	Review vector-store identities and permissions; examine tenant, workspace, and environment segregation; inspect metadata and source references stored with vectors; assess query interfaces and bulk-export capability; verify how source updates and deletion are propagated.
Expected Evidence	Vector-store architecture, access policies, identity configuration, metadata samples, ingestion and deletion workflows, query logs, and segregation test results.
Potential Indicators of Weakness	Shared index exposes data across permission domains; vector queries bypass application authorisation; metadata reveals protected information; bulk export is excessively privileged; embeddings or references remain available beyond the defined deletion period.

Embeddings should not automatically be treated as anonymous or non-sensitive. Their classification should reflect the source material, associated metadata, and realistic extraction, or inference risks in the deployed environment.

A.4.5 AI-DATA-05 - Data Retention, Deletion, and Lifecycle Control

Field	Details
Assessment Area	Data lifecycle management
Assurance Objective	Establish whether AI-related data is retained, updated, and deleted according to defined business, security, privacy, and legal requirements.
Illustrative Assessment Activities	Review retention periods for prompts, responses, files, logs, caches, indexes, embeddings, and tool outputs; test approved deletion and reclassification scenarios; examine propagation to derived stores and provider-managed services; verify exception and legal-hold processes where applicable.
Expected Evidence	Retention schedules, platform settings, deletion procedures, provider terms, test results, source and index timestamps, exception records, and lifecycle monitoring.
Potential Indicators of Weakness	Deleted or reclassified content remains retrievable beyond the approved propagation period; transcripts retained without a defined purpose; source deletion does not reach indexes or caches; provider retention conflicts with organisational requirements; deletion failures are not detected.

A deletion test should account for documented propagation delays, immutable backups, and legal holds. The assessment should distinguish an expected lifecycle delay from content that remains available contrary to policy.

A.5 Identity, Authorisation, and Boundary Isolation

This domain assesses whether AI-mediated access preserves the identity, permissions, and isolation controls defined by the connected systems. It covers direct retrieval, generated responses, citations, metadata, conversation state, and delegated actions.

A.5.1 AI-IDAM-01 - Role-Based Permission Boundary Testing

Field	Details
Assessment Area	Role-based authorisation
Assurance Objective	Determine whether users can obtain only the information and capabilities permitted by their assigned roles.
Illustrative Assessment Activities	Use representative accounts with different roles; submit equivalent queries across those accounts; compare direct source access with AI search, retrieval, summarisation, and citation behaviour; examine whether backend services independently enforce authorisation.

Field	Details
Expected Evidence	Test-account matrix, role and group assignments, source permissions, prompt and response records, retrieval logs, direct-access results, and API authorisation evidence.
Potential Indicators of Weakness	A standard user receives content restricted to a privileged role; summaries disclose protected facts from inaccessible sources; citations or metadata reveal restricted records; authorisation is enforced only by the prompt or user interface.

Testing should distinguish unauthorised disclosure from reasonable inference based solely on information the user is permitted to access, as described in *Section 4.2*.

A.5.2 AI-IDAM-02 - Cross-User and Cross-Group Isolation

Field	Details
Assessment Area	User and group isolation
Assurance Objective	Establish whether private or group-restricted content remains isolated between users and collaboration boundaries.
Illustrative Assessment Activities	Create approved test content for individual users and groups; attempt discovery from accounts outside each boundary; examine conversations, uploaded files, memories, cached results and generated artefacts; compare AI-mediated results with direct source access.
Expected Evidence	User and group permissions, controlled test-content records, prompt and response evidence, conversation or memory configuration, source-access results, and relevant logs.
Potential Indicators of Weakness	One user can access another user's private conversation or files; group-restricted information appears in another group's output; cached or remembered content crosses user boundaries; shared links or generated artefacts bypass intended restrictions.

A.5.3 AI-IDAM-03 - Delegated Identity and Execution Context

Field	Details
Assessment Area	Delegated identity
Assurance Objective	Determine whether agents and tools use an appropriate identity and remain within the authority approved for the invoking user and task.
Illustrative Assessment Activities	Identify whether actions execute using delegated user tokens, workload identities, or shared service accounts; inspect token scopes and claims; compare actions available to users with different privileges; verify how

Field	Details
	user context is passed to tools and downstream APIs; examine attribution in audit records.
Expected Evidence	Identity-flow diagrams, token claims, application permissions, service-account roles, user-role matrix, tool configuration, invocation logs, and action results.
Potential Indicators of Weakness	A low-privileged user invokes an action through a more privileged service identity; user context is lost between the AI application and tool; shared credentials prevent attribution; downstream services trust user-supplied identity attributes; delegated permissions exceed the approved task.

A service identity may legitimately hold permissions that individual users do not. In that design, application controls must independently determine whether each requested action is authorised and must record the responsible user and execution identity.

A.5.4 AI-IDAM-04 - Permission Change and Revocation

Field	Details
Assessment Area	Authorisation lifecycle
Assurance Objective	Establish whether access changes and revocations take effect across the AI system within the approved period.
Illustrative Assessment Activities	Grant and remove access using approved test accounts; test retrieval before and after the change; examine active sessions, cached responses, indexes and connector synchronisation; test deletion or reclassification of controlled content; review propagation and failure handling.
Expected Evidence	Permission-change records, timestamps, session and token configuration, connector synchronisation logs, index update records, before-and-after results, and documented propagation targets.
Potential Indicators of Weakness	Revoked users retain access beyond the approved period; active sessions bypass a required revocation; indexes continue returning protected content; failed permission synchronisation is not detected; reclassified content remains exposed.

Immediate propagation may not be technically achievable in every architecture. The assessment should compare observed behaviour with the documented security requirement and determine what exposure exists during any delay.

A.5.5 AI-TENANT-01 - Tenant, Workspace, and Environment Isolation

Field	Details
Assessment Area	Isolation between security domains
Assurance Objective	Determine whether the system preserves isolation between tenants, workspaces, projects, business units, and deployment environments.
Illustrative Assessment Activities	Test approved cross-boundary retrieval and API scenarios; review partitioning and server-side tenant enforcement; examine whether environments share indexes, vector stores, model credentials, service identities, or data sources; assess agent and tool boundaries; verify that client-supplied scope identifiers cannot override the authenticated context.
Expected Evidence	Architecture and data-flow diagrams, tenant and workspace configuration, server-side authorisation logic, index and vector-store design, service-identity records, environment configuration, test evidence, and audit logs.
Potential Indicators of Weakness	Content crosses tenant or workspace boundaries; non-production identities can access production information; a shared service identity bypasses domain separation; user-controlled tenant identifiers influence access without server-side validation; shared indexes do not enforce the required isolation.

Shared infrastructure is not inherently insecure. The assessment should determine whether logical isolation is consistently enforced at every relevant query, storage, and execution boundary.

A.6 Retrieval and Knowledge Security

This domain assesses whether enterprise knowledge is ingested, indexed, and retrieved without weakening source permissions, integrity, or traceability. It applies to RAG systems, enterprise search, copilots, and other AI features that use external information to construct responses.

A.6.1 AI-RAG-01 - Source Permission Enforcement

Field	Details
Assessment Area	Retrieval authorisation
Assurance Objective	Determine whether retrieval consistently enforces the effective permissions of the connected source systems.
Illustrative Assessment Activities	Compare source permissions with retrieval results using representative roles; review how identities and access rules are propagated or evaluated; test restricted documents and repositories; examine

Field	Details
	enforcement at query, retrieval and response stages; verify failure behaviour when permission information is unavailable.
Expected Evidence	Source permissions, retrieval architecture, identity claims, index configuration, filter logic, query logs, role-based test results and direct-access comparisons.
Potential Indicators of Weakness	Restricted content is returned to an unauthorised user; inherited or conditional permissions are omitted; access filters rely on user-controlled values; retrieval proceeds when authorisation data is missing or stale; enforcement occurs only in the user interface.

The architecture does not have to copy source ACLs into the index. It must provide equivalent and reliable enforcement before protected content reaches the model or user.

A.6.2 AI-RAG-02 - Retrievable-Unit Permission Integrity

Field	Details
Assessment Area	Document and chunk authorisation
Assurance Objective	Establish whether permissions remain effective at the smallest unit of content that can be retrieved or supplied to the model.
Illustrative Assessment Activities	Review document parsing and chunking; test documents containing sections, attachments, comments, or fields with different permissions; examine parent-child retrieval and neighbouring-chunk expansion; verify that generated context does not combine authorised and unauthorised portions.
Expected Evidence	Chunking and parsing configuration, source structure, permission records, retrieved-context samples, query traces, and role-comparison results.
Potential Indicators of Weakness	A permitted chunk contains text from a restricted section; attachment or comment permissions are lost during ingestion; neighbouring chunks expose protected content; access is enforced at document level when finer restrictions exist within the source.

Chunk-level ACLs are not required in every implementation. The requirement is effective authorisation at the smallest retrievable unit supported by the system.

A.6.3 AI-RAG-03 - Metadata and Reference Disclosure

Field	Details
Assessment Area	Retrieval metadata

Field	Details
Assurance Objective	Determine whether metadata, search results, and source references disclose protected information independently of document content.
Illustrative Assessment Activities	Search for controlled project names, filenames, titles, authors, labels, paths, tags and source identifiers using accounts with different permissions; inspect previews, snippets, result counts, citations and error messages; compare displayed metadata with direct source visibility.
Expected Evidence	Source metadata, permission records, search and response results, citation output, direct-access comparisons, and screenshots.
Potential Indicators of Weakness	Restricted titles or filenames are visible; citations reveal confidential matters; result counts confirm the existence of protected records; source paths expose tenant, customer or project information; snippets contain text from inaccessible content.

A.6.4 AI-RAG-04 - Canary Content Validation

Field	Details
Assessment Area	Controlled retrieval validation
Assurance Objective	Use controlled content to determine whether retrieval, permission, and isolation controls behave as intended.
Illustrative Assessment Activities	Create synthetic documents containing unique markers; apply different users, groups, and classifications; test direct search, semantic retrieval, summarisation, citations and aggregation; change or revoke access and repeat relevant tests within the approved propagation period.
Expected Evidence	Canary-content register, unique markers, source permissions, creation and change timestamps, prompts, retrieved context, responses, and role-comparison results.
Potential Indicators of Weakness	Canary content crosses an access boundary; restricted markers appear in a response or citation; revoked content remains available beyond the approved period; content from one workspace or tenant is returned in another.

Canary material should contain synthetic information, be clearly owned and be removed after testing. It should not introduce real secrets or content that could affect production users.

A.6.5 AI-RAG-05 - Ingestion Integrity and Retrieval Poisoning

Field	Details
Assessment Area	Knowledge-base integrity
Assurance Objective	Determine whether unauthorised or manipulated content can enter the retrieval corpus and materially influence system output.
Illustrative Assessment Activities	Identify who and what can create or modify indexed content; review source approval and provenance controls; introduce approved benign test content; examine ingestion from shared, external, or user-controlled sources; test whether manipulated metadata or ranking signals increase the content's retrieval priority.
Expected Evidence	Ingestion architecture, source inventory, write permissions, provenance records, approval workflows, index logs, ranking configuration, and controlled test results.
Potential Indicators of Weakness	Untrusted users can publish content into an authoritative corpus; source provenance is not retained; manipulated metadata influences ranking; deleted or superseded content remains authoritative; ingestion changes are not attributable or monitored.

A.6.6 AI-RAG-06 - Indirect Instruction Injection

Field	Details
Assessment Area	Retrieved-instruction handling
Assurance Objective	Determine whether instructions embedded in retrieved content can override trusted instructions, alter security-relevant behaviour, or influence tool execution.
Illustrative Assessment Activities	Place approved benign instructions in retrievable documents or metadata; test conflicting and concealed instructions; examine whether retrieved content can suppress warnings, redirect retrieval, request disclosure, or influence tool parameters; review how untrusted content is separated and labelled within the model context.
Expected Evidence	Controlled source content, prompt construction, retrieved context, model responses, tool-call records, orchestration controls, and guardrail telemetry.
Potential Indicators of Weakness	The model treats retrieved content as authoritative instruction; a document changes tool selection or parameters; embedded instructions cause protected information to be disclosed; retrieved content overrides approval or policy controls.

Following an instruction in retrieved content is not automatically a vulnerability. A finding requires a violation of the intended instruction hierarchy, security boundary or business rule.

A.6.7 AI-RAG-07 - Grounding, Citation and Source Traceability

Field	Details
Assessment Area	Retrieval grounding and traceability
Assurance Objective	Establish whether retrieval-based responses can be traced to authorised sources that support the claims presented.
Illustrative Assessment Activities	Ask questions requiring source support; compare material claims with the retrieved and cited content; verify that cited sources are accessible to the user; test superseded and conflicting sources; examine how the system communicates missing evidence or uncertainty.
Expected Evidence	Prompt and response records, retrieved context, cited sources, source versions, user permissions, claim-to-source comparisons, and configuration governing citations or grounding.
Potential Indicators of Weakness	Citations do not support the stated claim; a citation points to inaccessible content; outdated sources are presented as current; conflicting evidence is concealed; the system presents an unsupported answer as grounded.

Citation capability is not mandatory for every AI system. Where citations are not part of the design, the assessment should determine whether the level of traceability is appropriate for the system's purpose and impact.

A.7 Prompt and Orchestration Security

This domain assesses whether the system preserves its intended instruction hierarchy and whether untrusted input can alter security-relevant behaviour. Prompt testing should focus on observable impact rather than whether a model can be persuaded to produce unusual text.

A.7.1 AI-ORCH-01 - Deterministic Security Policy Enforcement

Field	Details
Assessment Area	Policy enforcement architecture
Assurance Objective	Determine whether security-critical decisions are enforced by trusted application, policy, identity, or tool layers rather than depending solely on model instructions.

Field	Details
Illustrative Assessment Activities	Identify where access control, data filtering, tool authorisation, approval, and transaction limits are enforced; review whether downstream APIs independently validate identity and authority; attempt approved prompt manipulation against controls represented only in instructions; examine failure behaviour when the model produces an unexpected decision.
Expected Evidence	Architecture diagrams, policy-engine configuration, authorisation logic, tool and API controls, request and response samples, test results, and enforcement logs.
Potential Indicators of Weakness	The model is asked to decide whether a user may access protected data; tool access depends only on a system prompt; downstream APIs trust model-generated authorisation decisions; prompt manipulation bypasses an approval or transaction limit.

Models may assist with classification or policy interpretation, but the final enforcement of security-critical decisions should occur in a component that fails predictably and can be tested independently.

A.7.2 AI-PROMPT-01 - Direct Prompt Injection

Field	Details
Assessment Area	User-controlled instruction handling
Assurance Objective	Determine whether user-supplied instructions can override trusted instructions or cross a defined security boundary.
Illustrative Assessment Activities	Submit approved conflicting, role-changing and instruction-override prompts; test attempts to access restricted information, invoke unavailable capabilities, or alter output constraints; vary phrasing, encoding, and conversation context where proportionate; compare model responses with downstream enforcement.
Expected Evidence	Test prompts, complete conversation context, responses, user role, system configuration, enforcement logs, and any resulting retrieval or tool activity.
Potential Indicators of Weakness	User instructions cause protected information to be disclosed; unavailable tools become accessible; the model creates parameters that bypass downstream restrictions; a prompt changes security-relevant behaviour without independent enforcement.

A model disregarding an instruction or producing policy-inconsistent text is not automatically a security vulnerability. The assessment should identify the security, privacy, or business boundary that was crossed and the resulting impact.

A.7.3 AI-PROMPT-02 - Internal Instruction and Configuration Exposure

Field	Details
Assessment Area	Prompt and orchestration information exposure
Assurance Objective	Determine whether disclosure of internal instructions or orchestration details exposes sensitive information or enables a credible attack path.
Illustrative Assessment Activities	Test whether the system reveals system or developer instructions, tool descriptions, internal identifiers, routing logic, or guardrail configuration; compare returned information with the deployed configuration; assess whether disclosed details contain secrets, protected business logic, or information useful for bypassing controls.
Expected Evidence	Prompt and response records, redacted configuration extracts, tool definitions, orchestration records, and analysis of how the disclosed information could be used.
Potential Indicators of Weakness	Internal instructions contain credentials or sensitive data; exposed tool schemas reveal privileged functions; disclosed routing details enable control bypass; confidential business rules are returned to unauthorised users.

System prompts should not contain secrets. Their disclosure should be reported only where the content is sensitive or materially assists exploitation; prompt confidentiality should not be relied upon as a security boundary.

A.7.4 AI-PROMPT-03 - Multi-Turn Context Manipulation

Field	Details
Assessment Area	Conversational state security
Assurance Objective	Establish whether accumulated conversation state can weaken security controls or cause the system to accept untrusted authority, instructions, or assumptions.
Illustrative Assessment Activities	Conduct controlled multi-turn conversations using staged requests, false authority claims, conflicting instructions and context accumulation; test whether earlier denied actions become available later; examine session resets, context limits, and state carried between conversations.
Expected Evidence	Complete transcripts, timestamps, session identifiers, user role, context-management configuration, enforcement logs, and resulting system or tool activity.
Potential Indicators of Weakness	The system accepts an unverified claim of authority; a prohibited action becomes available after staged requests; security-relevant state crosses users or sessions; earlier untrusted content continues to influence actions after it should have been cleared.

Variation between model responses should be recorded using the EAISA evidence confidence levels. A single inconsistent response may justify investigation but does not by itself establish a systemic control failure.

A.7.5 AI-PROMPT-04 - Instruction Provenance and Context Separation

Field	Details
Assessment Area	Instruction hierarchy and context construction
Assurance Objective	Determine whether the system distinguishes trusted instructions from user input, retrieved content, tool output, and other untrusted context.
Illustrative Assessment Activities	Review how system, developer, user, retrieval and tool content are assembled; identify the source and trust level of each context element; test conflicting instructions across those sources; examine delimiting, content labelling, context isolation, and tool-output handling; verify that untrusted content cannot redefine security policy.
Expected Evidence	Prompt templates, orchestration logic, context samples, tool definitions, trust-boundary diagrams, test conversations, and tool-call records.
Potential Indicators of Weakness	User input is concatenated into trusted instructions; retrieved text is presented as system instruction; tool output can redefine subsequent actions; content provenance is lost during orchestration; untrusted context changes policy or authorisation decisions.

Delimiters and instruction wording may reduce accidental confusion but are not sufficient controls for high-impact actions. Security boundaries should remain enforced outside the model.

A.8 Output and Downstream Handling

This domain assesses how AI-generated content is presented to users and consumed by other systems. Model output should be treated as untrusted data unless it has passed controls appropriate to its destination and intended use.

A.8.1 AI-OUTPUT-01 - Safe Rendering and Content Presentation

Field	Details
Assessment Area	User-facing output rendering
Assurance Objective	Determine whether generated content is rendered without enabling script execution, unsafe resource loading, deceptive navigation, or other client-side behaviour.

Field	Details
Illustrative Assessment Activities	Test approved HTML, Markdown, links, images, attachments, filenames, and document content; examine context-aware encoding and sanitisation; review permitted URL schemes and external resource loading; assess rendering in browsers, email, chat, tickets, and generated documents.
Expected Evidence	Raw model output, rendered output, sanitisation configuration, content-security controls, URL-handling logic, screenshots, and browser or application logs.
Potential Indicators of Weakness	Generated markup executes in the user's session; unsafe URI schemes are accepted; remote resources are loaded without appropriate control; misleading links conceal their destination; generated filenames or document content trigger unsafe downstream behaviour.

HTML, Markdown and links are not inherently unsafe. A finding requires ineffective handling in the context where the output is rendered or used.

A.8.2 AI-OUTPUT-02 - Decision and Operational Integrity

Field	Details
Assessment Area	Reliance on generated output
Assurance Objective	Determine whether unsupported, incomplete, or misleading output could materially affect a business, security, or operational decision.
Illustrative Assessment Activities	Identify workflows that rely on generated recommendations or summaries; compare representative outputs with authoritative information; examine uncertainty, material omissions, and conflicting evidence; review human-verification requirements and controls preventing unverified output from becoming an authoritative record or action.
Expected Evidence	Workflow diagrams, representative prompts, and outputs, authoritative source records, review procedures, approval evidence, and downstream action logs.
Potential Indicators of Weakness	Unverified output automatically drives a consequential decision; material uncertainty is presented as fact; a summary omits information required for safe action; operators are instructed to rely on generated content without appropriate review.

General model inaccuracy is not automatically a security finding. It becomes relevant to EAISA where the deployed workflow creates a credible integrity, safety, compliance, or business risk.

A.8.3 AI-OUTPUT-03 - Sensitive Information in Generated Output

Field	Details
Assessment Area	Output data exposure
Assurance Objective	Establish whether generated responses limit sensitive information to what the user is authorised and has a legitimate need to receive.
Illustrative Assessment Activities	Request broad summaries and comparisons using representative roles; test aggregation across authorised sources; inspect generated tables, exports and citations; assess whether redaction, masking or output limits operate as intended; compare output detail with the approved use case.
Expected Evidence	Redacted prompt and response records, source permissions, user-role evidence, data-classification records, output-control configuration and aggregation test results.
Potential Indicators of Weakness	Output discloses information from an inaccessible source; broad aggregation reveals protected facts; sensitive fields are returned where masked values would meet the purpose; generated exports contain more information than the interface displays or the user requires.

A user may legitimately infer new information from authorised sources. Assessment should focus on whether the output defeats an intended data boundary, reveals protected facts or creates unacceptable aggregation risk.

A.8.4 AI-OUTPUT-04 - Output-to-Workflow Injection

Field	Details
Assessment Area	Machine-consumed output and workflow safety
Assurance Objective	Determine whether AI-generated output can alter downstream commands, parameters, records or workflow behaviour outside the intended task.
Illustrative Assessment Activities	Map where output is passed into APIs, tools, tickets, emails, code, queries, commands or automation; test approved structured and free-text payloads; review schema and semantic validation; examine escaping, allowlists, approval controls and transaction limits; verify how malformed or unexpected output is handled.
Expected Evidence	Integration diagrams, output schemas, validation logic, workflow and tool logs, request and response samples, approval records and controlled test results.
Potential Indicators of Weakness	Free-form output is executed as a command; generated fields alter workflow routing or record ownership; structured output passes schema validation but contains unauthorised values; downstream systems trust

Field	Details
	generated URLs, code or queries without validation; malformed output leaves a transaction in an unsafe state.

Schema validation confirms structure, not authority or business intent. Downstream components must also validate the permitted operation, target, values and invoking identity.

A.9 Agent, Tool, Plugin, MCP, and Memory Security

This domain assesses systems that can retain state, select capabilities or act through tools and external services. Testing should follow delegated authority from the initiating user through the agent, orchestration layer, integration protocol and target system.

A.9.1 AI-AGENT-01 - Agent Lifecycle and Publication Control

Field	Details
Assessment Area	Agent governance and deployment
Assurance Objective	Determine whether AI agents are owned, reviewed and released through controls proportionate to their access and capabilities.
Illustrative Assessment Activities	Review who can create, modify, publish and share agents; sample agent approvals and versions; compare published capabilities with approved purposes; examine retirement and disablement processes.
Expected Evidence	Agent register, ownership records, publication settings, version history, approvals, capability definitions and retirement records.
Potential Indicators of Weakness	Users can publish broadly accessible agents without review; an agent has no accountable owner; production capabilities differ from those approved; retired agents or versions remain usable.

This module assesses the agent lifecycle. Organisation-wide inventory coverage remains within *AI-GOV-03*.

A.9.2 AI-AGENT-02 - Tool Scope and Least Privilege

Field	Details
Assessment Area	Tool permissions
Assurance Objective	Establish whether each agent, tool and integration receives only the permissions required for its approved purpose.

Field	Details
Illustrative Assessment Activities	Review API scopes, application roles, service identities and accessible operations; compare read and write capabilities with the intended use; test representative users; examine whether tools can reach unrelated tenants, repositories or environments.
Expected Evidence	Tool configuration, OAuth scopes, application permissions, service-identity roles, API policies, user-role matrix and controlled action results.
Potential Indicators of Weakness	A read-only use case receives write or delete access; one service identity spans unrelated environments; an agent can invoke functions outside its purpose; unused privileged scopes remain enabled.

A.9.3 AI-AGENT-03 - Action Authorisation and Approval

Field	Details
Assessment Area	Agent action control
Assurance Objective	Determine whether every action is authorised for the responsible user and whether consequential actions receive appropriate confirmation or approval.
Illustrative Assessment Activities	Test approved create, update, send, approve and delete operations using controlled objects; verify authorisation at the target system; inspect approval prompts for action, target and consequence; test cancellation, changed parameters and expired approvals.
Expected Evidence	Identity and authorisation records, approval screens, tool calls, before-and-after state, transaction logs and user-role evidence.
Potential Indicators of Weakness	A user approves an action they could not perform directly; approval omits the target or material parameters; one approval authorises a different or repeated action; downstream services do not independently enforce authority.

Approval should be based on consequence and reversibility. Requiring confirmation for every low-risk action may encourage habitual approval without improving security.

A.9.4 AI-AGENT-04 - Tool Selection and Argument Integrity

Field	Details
Assessment Area	Tool invocation integrity
Assurance Objective	Determine whether untrusted input can cause selection of an unintended tool or manipulation of its arguments.

Field	Details
Illustrative Assessment Activities	Test approved user and retrieved content that may influence tool selection; inspect generated arguments; vary object identifiers, recipients, paths and action types; review server-side validation, allowlists and binding between the approved action and executed request.
Expected Evidence	Tool definitions, argument schemas, invocation logs, validation logic, approval records, source input and controlled action results.
Potential Indicators of Weakness	Retrieved content selects an unintended tool; hidden or sensitive arguments are controlled through free text; identifiers can be replaced after approval; the tool accepts values outside the authorised task.

A valid schema confirms format, not authorisation or business intent. The receiving tool must validate the operation, target and values.

A.9.5 AI-AGENT-05 - Confused Deputy and Cross-User Action Abuse

Field	Details
Assessment Area	Delegated authority isolation
Assurance Objective	Determine whether an agent can be used to exercise another user's or service's authority outside the approved delegation.
Illustrative Assessment Activities	Compare actions across users with different permissions; inspect how user context is bound to each invocation; test controlled references to another user's objects; review shared service identities and downstream authorisation; verify audit attribution.
Expected Evidence	Identity-flow diagrams, token claims, service-account permissions, tool logs, target-system audit records and cross-role test results.
Potential Indicators of Weakness	A low-privileged user triggers a privileged service action; one user can act on another user's records; the target system receives no reliable user context; actions are attributed only to a shared service account.

A.9.6 AI-AGENT-06 - Execution Bounds, Idempotency and Recovery

Field	Details
Assessment Area	Agent execution resilience
Assurance Objective	Establish whether agent execution is bounded and can be stopped, retried or reversed without uncontrolled repetition or inconsistent state.

Field	Details
Illustrative Assessment Activities	Review maximum steps, timeouts, cancellation, concurrency and usage limits; test controlled repeated or failed operations; examine retry and idempotency controls; verify rollback or reconciliation for partial completion.
Expected Evidence	Agent configuration, execution traces, rate and quota settings, transaction identifiers, timeout records, recovery procedures and controlled test results.
Potential Indicators of Weakness	An agent enters an unbounded loop; retries duplicate a transaction or message; cancellation does not stop downstream activity; partial failure leaves records in an unsafe or inconsistent state; no practical recovery method exists.

A.9.7 AI-AGENT-07 - Multi-Step Planning and Task Control

Field	Details
Assessment Area	Agent planning and autonomous execution
Assurance Objective	Determine whether a multi-step agent remains within its approved objective, authority and safety constraints throughout execution.
Illustrative Assessment Activities	Review how goals become plans and tool calls; test approved tasks containing ambiguous or conflicting objectives; examine whether authority is re-evaluated at each step; verify approval when a plan changes materially; inspect intermediate data and state.
Expected Evidence	Plans, execution traces, tool calls, intermediate state, approval records, policy decisions and final outcomes.
Potential Indicators of Weakness	The agent expands the task beyond the user’s request; a permitted first step enables an unauthorised later action; plan changes bypass approval; intermediate output exposes sensitive information; security checks occur only at the start of the task.

A.9.8 AI-MCP-01 - MCP Server Trust and Capability Exposure

Field	Details
Assessment Area	MCP server and capability governance
Assurance Objective	Determine whether MCP clients connect only to approved servers and receive capabilities appropriate to the user and use case.

Field	Details
Illustrative Assessment Activities	Inventory configured servers and connection methods; verify server ownership, provenance and approval; review exposed tools, resources and prompts; compare discovered capabilities across roles; examine how capability changes are detected and approved.
Expected Evidence	MCP server register, client configuration, server identity and deployment records, discovery responses, capability definitions, approval records and change history.
Potential Indicators of Weakness	Users can connect unapproved servers; a server exposes unnecessary privileged tools; capabilities differ from the approved configuration without detection; development or personal MCP servers are trusted in production.

Tool discovery is normal MCP behaviour. A finding requires unauthorised discovery, excessive capability exposure or misplaced trust in the server.

A.9.9 AI-MCP-02 - MCP Authentication and Authorisation

Field	Details
Assessment Area	MCP connection and access control
Assurance Objective	Establish whether MCP connections authenticate the relevant parties and authorise each requested capability according to the transport and deployment model.
Illustrative Assessment Activities	Review authentication for remote MCP servers; validate token issuer, audience, scope and expiry where tokens are used; assess user-to-client and client-to-server identity propagation; examine credential storage and forwarding; review local process and operating-system boundaries for local servers.
Expected Evidence	Transport architecture, authentication configuration, token claims, authorisation policies, client and server logs, credential-storage controls and role-based results.
Potential Indicators of Weakness	A remote server accepts unauthenticated requests; tokens intended for another service are accepted; client credentials are forwarded as user authority without validation; local servers inherit excessive filesystem or process access; authorisation is not enforced per capability.

Authentication requirements differ between local and remote transports. A local *stdio* server may rely on operating-system and process boundaries, while a remotely reachable server requires controls appropriate to its network exposure and clients.

A.9.10 AI-MCP-03 - MCP Tool Contract and Message Validation

Field	Details
Assessment Area	MCP message and tool integrity
Assurance Objective	Determine whether MCP clients and servers validate tool arguments, responses and protocol messages before acting on them.
Illustrative Assessment Activities	Review tool input and output schemas; test approved missing, additional, malformed and boundary values; examine server-side validation and error handling; test whether tool results can introduce untrusted instructions or unsafe structured content; verify that protocol metadata cannot override authenticated context.
Expected Evidence	Tool definitions, schemas, server validation logic, JSON-RPC samples, error responses, invocation logs and controlled test results.
Potential Indicators of Weakness	The server trusts schema validation performed only by the client; additional arguments alter the operation; malformed messages trigger unintended behaviour; tool results are treated as trusted instructions; client-supplied metadata changes identity or scope.

Apply this module where MCP introduces a distinct protocol boundary. General tool authorisation and argument risks remain covered by the agent modules and should not be reported twice.

A.9.11 AI-MEM-01 - Agent Memory and Persistent Context

Field	Details
Assessment Area	Agent memory security
Assurance Objective	Determine whether persistent memory is authorised, isolated, accurate and managed throughout its lifecycle.
Illustrative Assessment Activities	Identify what information enters memory and how it is selected; test isolation between users, groups and tenants; examine whether untrusted content can create or alter durable memory; review access, correction, deletion, expiry and administrative visibility; test whether memory can influence later tools or decisions.
Expected Evidence	Memory architecture, storage and access controls, creation and retrieval logs, retention settings, user controls, controlled memory records and cross-user test results.
Potential Indicators of Weakness	One user receives another user’s memory; untrusted content creates persistent instructions; sensitive information is retained without a defined need; users cannot correct or remove inaccurate memory; stale memory drives an unauthorised action.

Memory is not limited to a dedicated product feature. Conversation summaries, profiles, checkpoints, cached plans and persisted agent state should be included where they influence later behaviour.

A.10 Model and Provider Security

This domain assesses whether models and hosting providers are selected, configured and changed in a way that remains consistent with the system's approved purpose, data requirements and security boundaries.

A.10.1 AI-MODEL-01 - Model Selection, Routing, and Allowlisting

Field	Details
Assessment Area	Model selection and routing
Assurance Objective	Determine whether the system uses only models and endpoints approved for the intended data, users, capabilities and deployment context.
Illustrative Assessment Activities	Review approved models and selection criteria; inspect model gateway and routing configuration; test whether users or applications can select unapproved models; examine fallbacks, aliases and regional endpoints; verify that capability differences are considered where models can use tools, process files or accept multimodal input.
Expected Evidence	Approved model register, risk assessments, gateway configuration, routing rules, model identifiers, endpoint settings, exception records and request logs.
Potential Indicators of Weakness	Sensitive requests are routed to an unapproved model or region; users can bypass the allowlist; a fallback model provides weaker controls; generic aliases obscure the model actually used; model capabilities are enabled without assessment.

A newer or more capable model is not automatically more secure. Approval should consider the deployed use case, available controls and information that the model will process.

A.10.2 AI-MODEL-02 - Provider Data Handling and Access

Field	Details
Assessment Area	Provider-managed data
Assurance Objective	Establish whether provider handling of prompts, responses, files, fine-tuning data and telemetry is compatible with organisational requirements and applicable obligations.

Field	Details
Illustrative Assessment Activities	Review retention and deletion settings; determine whether customer content may be used for model training or service improvement; examine processing and storage locations; identify subprocessors and support access; review encryption, tenant isolation, incident notification and deletion commitments; compare contractual terms with deployed settings.
Expected Evidence	Provider terms, data-processing agreement, privacy and security documentation, platform configuration, regional settings, subprocessor information, assurance reports and approved exceptions.
Potential Indicators of Weakness	Provider use of customer content conflicts with approved purpose; retention settings exceed organisational requirements; processing locations are not understood; provider personnel can access sensitive content without sufficient control; contractual commitments and technical configuration do not align.

Public product statements should not be assumed to apply equally to consumer services, enterprise subscriptions, APIs and preview features. The assessment should verify the terms and settings that apply to the deployed service.

A.10.3 AI-MODEL-03 - Model Change and Behavioural Regression

Field	Details
Assessment Area	Model version and change control
Assurance Objective	Determine whether model changes are identified, assessed and tested before they materially alter production behaviour or security controls.
Illustrative Assessment Activities	Identify pinned versions, aliases and provider-managed updates; review notification and approval processes; inspect regression tests for permission handling, retrieval, prompts, tools and structured output; examine fallback and rollback options; assess how deprecated models are replaced.
Expected Evidence	Model identifiers, version history, provider notices, change records, evaluation results, deployment configuration, monitoring records and rollback or migration plans.
Potential Indicators of Weakness	A mutable alias changes the production model without review; security regression tests are absent; provider changes alter tool or output behaviour without detection; deprecated models remain in use; no response plan exists where rollback is unavailable.

Not every provider exposes an exact model version or permits rollback. Where these controls are unavailable, the organisation should use behavioural baselines, monitoring and release gates appropriate to the resulting uncertainty.

A.10.4 AI-MODEL-04 - Fine-Tuning and Model Adaptation Security

Field	Details
Assessment Area	Model customisation and adaptation
Assurance Objective	Determine whether fine-tuning, adapters and other model adaptations use authorised data and produce controlled, traceable artefacts.
Illustrative Assessment Activities	Identify customised models and adaptation methods; review training-data provenance, approval and access; assess protection against poisoned or unauthorised data; inspect handling of personal information, confidential content and secrets; examine versioning, storage, deployment and security evaluation of adapted artefacts.
Expected Evidence	Dataset inventory, provenance and approval records, data-access controls, adaptation configuration, model or adapter versions, storage permissions, evaluation results and deployment records.
Potential Indicators of Weakness	Unapproved or poorly sourced data is used for adaptation; sensitive information is included without assessment; contributors can introduce training data without review; adapters or checkpoints are unprotected; the deployed artefact cannot be matched to the evaluated version; security behaviour is not retested after adaptation.

This module includes fine-tuning, prompt tuning, adapters and comparable methods that modify or extend model behaviour. It does not require access to provider-managed training internals that the organisation cannot inspect.

A.11 Application, API, and Customer-Facing Security

This domain assesses the application components that expose, orchestrate and integrate AI capabilities. These modules supplement rather than replace established web application and API security testing.

A.11.1 AI-APPSEC-01 - Authentication and Session Security

Field	Details
Assessment Area	Authentication and session management
Assurance Objective	Determine whether users, administrators and calling applications are reliably authenticated and bound to the correct session and AI context.
Illustrative Assessment Activities	Review authentication flows, session creation and termination; test conversation and file identifiers across accounts; examine token storage, expiry and revocation; assess session sharing and concurrent access; verify reauthentication for sensitive administrative or agent functions.

Field	Details
Expected Evidence	Authentication architecture, token and cookie configuration, session records, identity claims, revocation settings, controlled test results and audit logs.
Potential Indicators of Weakness	Conversation identifiers provide access across accounts; sessions remain active after required revocation; sensitive operations do not require sufficient authentication; identity context changes without creating a new session; tokens are exposed to unauthorised client-side components.

A.11.2 AI-APPSEC-02 - API and Object Authorisation

Field	Details
Assessment Area	API authorisation
Assurance Objective	Establish whether every API operation enforces authorisation for the authenticated identity, requested function and affected object.
Illustrative Assessment Activities	Test conversation, file, agent, tool, workspace and configuration identifiers across representative roles; invoke APIs directly rather than through the interface; vary object references and action types; review batch and asynchronous operations; verify authorisation at downstream services.
Expected Evidence	API specifications, identity claims, role mappings, authorisation logic, request and response samples, gateway logs and cross-role test results.
Potential Indicators of Weakness	A user accesses another user's conversation or file by changing an identifier; hidden administrative operations are directly callable; write operations rely on interface restrictions; asynchronous jobs lose the initiating user's authority; downstream APIs trust unverified role data.

A.11.3 AI-APPSEC-03 - File Upload and Content Processing

Field	Details
Assessment Area	Uploaded and generated file security
Assurance Objective	Determine whether files and documents are accepted, processed, stored and returned without compromising the application or connected systems.
Illustrative Assessment Activities	Review type, size and count limits; test approved mismatches between extension, content type and file content; examine archive, document, image, OCR and media processing; assess parser isolation, malware controls, storage permissions and filename handling; verify cleanup of temporary files.

Field	Details
Expected Evidence	Upload configuration, processing architecture, parser and sandbox controls, storage permissions, scanning records, error responses and controlled test files.
Potential Indicators of Weakness	Untrusted files reach vulnerable processing services; extension checks are treated as content validation; archives cause uncontrolled resource consumption; uploaded content is publicly accessible; filenames affect paths or downstream commands; temporary files retain sensitive data.

Instructions concealed within uploaded content are assessed under the retrieval, prompt and multimodal modules. This module addresses the security of the file-processing path itself.

A.11.4 AI-APPSEC-04 - Integration Endpoint and Outbound Request Security

Field	Details
Assessment Area	Webhooks, callbacks and external requests
Assurance Objective	Determine whether integration endpoints authenticate messages and restrict where the AI application can send requests or data.
Illustrative Assessment Activities	Review webhook authentication, signatures, timestamps and replay controls; test approved callback and connector endpoints; examine URL-fetching features and outbound network restrictions; assess redirects, DNS changes and access to internal or cloud metadata services; inspect error handling and retry behaviour.
Expected Evidence	Integration architecture, webhook configuration, signature validation, outbound network policy, destination allowlists, request logs and controlled test results.
Potential Indicators of Weakness	Unsigned or replayed webhooks are accepted; user-controlled URLs cause server-side requests to internal services; redirects bypass destination restrictions; integration secrets appear in logs or callbacks; retries repeat a consequential operation.

A.11.5 AI-APPSEC-05 - Supporting Application and Business Logic Security

Field	Details
Assessment Area	Application security
Assurance Objective	Determine whether conventional application weaknesses can undermine AI-specific controls, data boundaries or workflows.

Field	Details
Illustrative Assessment Activities	Select established web and API security tests according to the architecture; assess input validation, injection, request integrity, workflow state, concurrency, error handling and security headers; review whether client-side restrictions can be bypassed; examine trust placed in model-generated values.
Expected Evidence	Application architecture, source or configuration evidence where available, request and response samples, workflow records, error behaviour and controlled test results.
Potential Indicators of Weakness	Injection reaches a data source or tool; workflow steps can be skipped; concurrent requests bypass an action limit; errors disclose prompts, credentials or internal configuration; client-side controls are treated as authoritative; model-generated values enter trusted operations without validation.

The selected procedures should align with recognised application and API testing guidance. EAISA does not reproduce a complete conventional penetration-testing standard.

A.11.6 AI-CUSTOMER-01 - External Exposure and Abuse Controls

Field	Details
Assessment Area	Customer-facing and publicly accessible AI
Assurance Objective	Establish whether externally accessible AI capabilities withstand untrusted input and foreseeable abuse without exposing data or causing unauthorised actions.
Illustrative Assessment Activities	Review anonymous and authenticated capabilities; test account and tenant isolation; examine rate, usage and file limits; assess automated and repeated interaction; review escalation to human operators; test whether generated communications or actions clearly retain the responsible identity and approval context.
Expected Evidence	Exposure architecture, user journeys, authentication settings, usage controls, abuse-monitoring rules, escalation procedures, controlled test results and audit logs.
Potential Indicators of Weakness	Anonymous users reach protected data or tools; one customer can influence another customer’s context; inexpensive automation causes material cost or service exhaustion; generated communications appear to come from an authorised employee without review; abuse cannot be attributed or contained.

Broader content safety, fairness and responsible-AI evaluation should be included only where it forms part of the agreed scope. This module focuses on security and operational abuse arising from external exposure.

A.12 Cloud and Integration Security

This domain assesses the cloud resources, identities, networks, storage and connectors supporting the AI system. It focuses on whether infrastructure controls preserve the boundaries established at the application and AI layers.

A.12.1 AI-CLOUD-01 - Workload Identity and Cloud Privilege

Field	Details
Assessment Area	Cloud identity and access management
Assurance Objective	Determine whether AI workloads use identifiable, managed and least-privileged identities for access to cloud resources and enterprise services.
Illustrative Assessment Activities	Inventory workload and service identities; review assigned roles and resource scopes; identify shared or long-lived credentials; examine identity federation and managed identities where supported; test representative access; review creation, rotation, disablement and monitoring.
Expected Evidence	Identity inventory, role assignments, trust policies, token claims, cloud IAM configuration, access reviews, authentication logs and controlled access results.
Potential Indicators of Weakness	One identity is shared across unrelated workloads or environments; broad subscription, project or account roles are assigned unnecessarily; inactive identities remain enabled; static credentials are used where a managed alternative is available; identity activity cannot be attributed to a workload.

This module evaluates the cloud identity itself. User-to-agent delegation and authorisation remain covered by the identity and agent modules.

A.12.2 AI-CLOUD-02 - Cloud Storage and Data-Service Access

Field	Details
Assessment Area	Cloud data-plane security
Assurance Objective	Establish whether storage services, databases, queues and AI data stores are accessible only to approved identities and network paths.
Illustrative Assessment Activities	Review access policies, public exposure and resource sharing; examine data-plane and control-plane permissions; assess temporary access links and delegated access tokens; test separation between environments; review encryption and audit configuration; identify unauthorised secondary data paths.

Field	Details
Expected Evidence	Storage and database configuration, access policies, resource-sharing settings, encryption configuration, network controls, access logs and controlled test results.
Potential Indicators of Weakness	AI data is publicly accessible; non-production workloads can read production stores; temporary access links are overprivileged or excessively long-lived; control-plane administrators have unmonitored access to sensitive content; a secondary export or backup bypasses primary access controls.

Data classification, retention, and deletion are assessed in the Data Security and Lifecycle domain. This module focuses on infrastructure-level access and exposure.

A.12.3 AI-CLOUD-03 - Network Exposure and Connectivity

Field	Details
Assessment Area	Cloud network security
Assurance Objective	Determine whether inbound, outbound and service-to-service connectivity is limited to the communication required by the approved architecture.
Illustrative Assessment Activities	Map public and private endpoints; review firewall, gateway and security-group rules; assess private connectivity and service endpoints where used; examine outbound access and destination controls; test administrative interfaces; verify DNS, proxy and certificate validation.
Expected Evidence	Network and data-flow diagrams, firewall and gateway configuration, endpoint inventory, DNS configuration, egress policy, flow or access logs and controlled connectivity results.
Potential Indicators of Weakness	Administrative or model endpoints are unnecessarily internet-accessible; broad network rules expose internal services; workloads can send data to arbitrary destinations; private connectivity is bypassed through a public endpoint; certificate or hostname validation is disabled.

A public endpoint is not automatically a vulnerability. Its risk depends on authentication, authorisation, exposed functionality, network controls, and monitoring.

A.12.4 AI-CLOUD-04 - Secrets and Credential Management

Field	Details
Assessment Area	Secrets protection
Assurance Objective	Determine whether model keys, API credentials, connector tokens and other secrets are generated, stored, used and revoked securely.

Field	Details
Illustrative Assessment Activities	Identify credentials used by AI components; review storage in code, configuration, prompts, logs and deployment systems; examine secret-store access; assess rotation and revocation; inspect client-side applications and generated diagnostic output; verify response procedures for exposed credentials.
Expected Evidence	Secret inventory, vault configuration, workload access policies, repository and pipeline controls, rotation records, redacted configuration and relevant audit logs.
Potential Indicators of Weakness	Secrets appear in prompts, source code, logs or client-side content; several services share the same credential; credentials are not rotated after exposure; secret-store access is broader than required; unused credentials remain valid.

Secret values should never be reproduced in assessment evidence. Evidence should show the location, exposure and affected identity using redacted or fingerprinted values.

A.12.5 AI-INT-01 - Connector Authentication and Integration Trust

Field	Details
Assessment Area	SaaS and enterprise connectors
Assurance Objective	Establish whether connectors are approved, authenticated and limited to the users, data and operations required by the integration.
Illustrative Assessment Activities	Inventory enabled connectors; verify publisher, ownership and approval; review delegated and application permissions; examine OAuth consent and token storage; assess user mapping and tenant restrictions; test disablement and consent revocation; review how connector failures affect access decisions.
Expected Evidence	Connector inventory, application registrations, consent grants, scopes, publisher information, identity mappings, approval records, token controls and connector logs.
Potential Indicators of Weakness	An unapproved connector accesses enterprise data; application-wide permissions are used where delegated access is expected; tokens remain valid after connector removal; users can connect personal or external tenants; connector failures cause access controls to fail open.

Where a connector also exposes agent tools or MCP capabilities, apply the relevant agent or MCP modules to its actions. Do not report the same excessive permission under multiple modules.

A.13 Privacy and Regulatory Considerations

This domain assesses whether personal information is handled consistently with the system's approved purpose and applicable privacy requirements. It supports technical assurance and does not replace advice from qualified privacy or legal specialists.

Applicable requirements depend on the organisation, deployment, affected individuals and jurisdictions. Relevant sources may include the EU GDPR, the UK GDPR and Data Protection Act 2018, Australia's Privacy Act 1988 and Australian Privacy Principles, New Zealand's Privacy Act 2020 and Information Privacy Principles, applicable United States privacy requirements, sector-specific privacy rules and contractual obligations.

A.13.1 AI-PRIV-01 - Privacy Impact and Lawful Use Review

Field	Details
Assessment Area	Privacy governance and purpose
Assurance Objective	Determine whether the collection, creation, use, and disclosure of personal information through the AI system are understood, necessary and permitted for the intended purpose.
Illustrative Assessment Activities	Review the privacy impact assessment and data flows; identify personal information collected directly, retrieved from other systems or inferred by the AI; compare processing with notices, policies and approved purposes; examine data minimisation and secondary uses; verify that identified privacy controls were implemented.
Expected Evidence	Privacy impact assessment, processing and data-flow records, privacy notices, approved purposes, data classifications, control implementation records and relevant approvals.
Potential Indicators of Weakness	Personal information is processed for an undocumented purpose; AI-generated inferences are not included in the privacy assessment; notices do not reflect actual handling; the system collects more information than required; identified privacy mitigations were not implemented.

Consent is not the only consideration and should not be presented as a universal basis for processing. The applicable requirements depend on the jurisdiction, entity, information type, purpose and relevant exceptions.

A.13.2 AI-PRIV-02 - Sensitive and High-Impact Personal Information

Field	Details
Assessment Area	Higher-risk personal information

Field	Details
Assurance Objective	Establish whether information requiring greater protection is identified and subject to controls proportionate to its sensitivity and potential effect on individuals.
Illustrative Assessment Activities	Identify health, biometric, financial, employment, identity, legal, government and other high-impact information; review collection and use restrictions; examine access, masking, approval and disclosure controls; assess whether the AI can infer or aggregate sensitive attributes; review applicable sector requirements.
Expected Evidence	Data classification, processing inventory, applicable privacy or sector requirements, access controls, masking configuration, approved use cases, privacy assessments and controlled test results.
Potential Indicators of Weakness	Sensitive information is processed without appropriate assessment; broad prompts or retrieval expose unnecessary attributes; the system infers sensitive characteristics outside its approved purpose; controls do not reflect the consequence of misuse; sector-specific requirements have not been considered.

Legal definitions of sensitive information differ between jurisdictions. The assessment should use the definitions and obligations applicable to the organisation rather than treating the examples above as a universal legal category.

A.13.3 AI-PRIV-03 - Overseas Processing and Provider Disclosure

Field	Details
Assessment Area	Cross-border and third-party handling
Assurance Objective	Determine whether overseas processing and disclosure to providers or subprocessors are identified, approved and managed according to applicable requirements.
Illustrative Assessment Activities	Map where prompts, responses, files, logs, telemetry and support data may be processed or accessed; identify providers and subprocessors; review regional settings and contractual safeguards; examine support and diagnostic access; verify how provider changes are monitored.
Expected Evidence	Cross-border data-flow map, provider agreements, subprocessor list, processing locations, regional configuration, risk assessments, contractual controls and change notifications.
Potential Indicators of Weakness	Overseas processing is not documented; regional settings do not provide the assumed restriction; provider personnel access content from unassessed locations; subprocessors are unknown; contractual and technical controls do not align with applicable cross-border requirements.

For covered Australian entities, APP 8 may apply to cross-border disclosure. For New Zealand agencies, Information Privacy Principle 12 addresses disclosure outside New

Zealand. Other privacy regimes, including the EU GDPR and UK GDPR, impose their own requirements on international transfers. Whether a particular provider arrangement constitutes a disclosure or restricted transfer requires assessment of the actual service and legal context.

A.13.4 AI-PRIV-04 - Transparency, Access, and Correction

Field	Details
Assessment Area	Individual transparency and privacy rights
Assurance Objective	Determine whether the organisation can explain its handling of personal information and respond to applicable access and correction requests.
Illustrative Assessment Activities	Review what individuals are told about AI-related collection and use; identify where personal information and AI-generated inferences are stored; assess search and export capability across conversations, files and derived stores; examine correction workflows and propagation to indexes, memory and downstream systems; test approved representative requests.
Expected Evidence	Privacy notices, information-location records, request procedures, search and export results, correction records, propagation logs and exception-handling procedures.
Potential Indicators of Weakness	The organisation cannot locate personal information held in AI stores; notices omit material AI processing; corrections do not propagate to retrieval or memory; inferred information is treated as outside privacy handling without assessment; request responses disclose another person’s information.

This module should not imply a universal right to deletion. Access, correction, retention, and deletion obligations differ by jurisdiction and circumstance.

A.14 Monitoring, Auditability, and Incident Response

This domain assesses whether the organisation can reconstruct AI activity, detect material abuse and contain incidents without relying on complete prompt or response capture. Logging should provide sufficient evidence while limiting unnecessary collection of sensitive information.

A.14.1 AI-MON-01 - Security Logging and Event Correlation

Field	Details
Assessment Area	AI security telemetry
Assurance Objective	Determine whether security-relevant activity can be traced across the user, application, model, retrieval and integration layers.

Field	Details
Illustrative Assessment Activities	Map available logs; trace representative requests across components; review user, session, model, retrieval and administrative events; examine timestamps and correlation identifiers; assess log access, integrity, retention and forwarding; verify redaction and minimisation.
Expected Evidence	Logging architecture, event schemas, sample records, correlation identifiers, retention settings, access controls, SIEM configuration and trace results.
Potential Indicators of Weakness	Requests cannot be traced across components; user or session identity is absent; administrative changes are not recorded; timestamps prevent reliable correlation; logs contain unnecessary sensitive content; security teams cannot access relevant provider telemetry.

Prompt and response content should be logged only where necessary and permitted. Metadata, identifiers, policy decisions and controlled content capture may provide sufficient evidence with less privacy risk.

A.14.2 AI-MON-02 - Agent and Tool Action Auditability

Field	Details
Assessment Area	Delegated-action logging
Assurance Objective	Establish whether actions performed through agents and tools are attributable, complete and suitable for investigation.
Illustrative Assessment Activities	Execute approved representative actions; verify capture of the initiating user, execution identity, agent, tool, target, material arguments, approval, outcome and time; examine multi-step correlation and partial failures; compare AI logs with target-system audit records.
Expected Evidence	Agent traces, tool invocation logs, approval records, target-system audit logs, identity claims, transaction identifiers and before-and-after state.
Potential Indicators of Weakness	Actions are attributed only to a shared service account; approval cannot be linked to execution; material parameters are absent; failed and retried actions are indistinguishable; agent records conflict with target-system logs.

Sensitive arguments and returned data may be redacted, hashed or referenced rather than recorded in full, provided the remaining evidence supports attribution and investigation.

A.14.3 AI-MON-03 - Detection and Alerting for AI Abuse

Field	Details
Assessment Area	Abuse detection

Field	Details
Assurance Objective	Determine whether material misuse, control failure and anomalous AI activity can generate timely and actionable detection.
Illustrative Assessment Activities	Identify credible detection scenarios from the threat model; generate approved representative events; test alerts for repeated access denials, cross-boundary retrieval, unusual tool activity, excessive execution, new privileged agents or connectors and security-control changes; review alert routing, context and response ownership.
Expected Evidence	Detection rules, alert records, test-event logs, investigation procedures, escalation routes, tuning history and response results.
Potential Indicators of Weakness	Material agent actions generate no alert; security-control changes are not monitored; alerts omit the responsible identity or affected resource; events are collected but not routed to an owner; detection depends only on matching known malicious prompt text.

Prompt-pattern detection may provide useful signals but is not a reliable security boundary. Detection should also use identity, access, retrieval, tool and behavioural telemetry.

A.14.4 AI-MON-04 - Usage, Resource and Cost Monitoring

Field	Details
Assessment Area	Consumption and cost visibility
Assurance Objective	Determine whether abnormal or uncontrolled AI consumption can be identified before it creates material service or financial impact.
Illustrative Assessment Activities	Review usage metrics by user, workload, model, agent and tool; examine token, request, file-processing and execution limits; test approved threshold events; inspect budget alerts and escalation; verify visibility into retries, loops and provider charges.
Expected Evidence	Usage dashboards, quota settings, billing configuration, alert rules, agent execution records, budget thresholds and controlled alert results.
Potential Indicators of Weakness	Usage cannot be attributed to a user or workload; agent loops are hidden within aggregate billing; alerts occur only after material cost is incurred; compromised credentials can consume unrestricted resources; provider and internal measurements cannot be reconciled.

Cost alerts provide visibility but do not prevent abuse. High-exposure or autonomous workloads may also require enforceable quotas, transaction limits or automatic containment.

A.14.5 AI-IR-01 - Incident Preparedness and Triage

Field	Details
Assessment Area	AI incident readiness
Assurance Objective	Establish whether the organisation can recognise, classify and investigate incidents involving AI-specific components and behaviour.
Illustrative Assessment Activities	Review incident scenarios, roles and escalation paths; conduct an approved walkthrough or tabletop exercise; assess procedures for data exposure, poisoned retrieval, compromised agents, unsafe tool actions and provider incidents; verify access to technical, privacy, legal and vendor support.
Expected Evidence	Incident response plan, AI-specific playbooks, contact lists, severity criteria, tabletop records, vendor escalation procedures and notification decision processes.
Potential Indicators of Weakness	AI incidents have no defined owner; responders cannot determine which model, data or tools were involved; provider escalation is unavailable or untested; privacy and legal review is not integrated; incident severity ignores delegated actions or operational impact.

The assessment should confirm that notification obligations can be evaluated by authorised privacy and legal personnel rather than making the legal determination itself.

A.14.6 AI-IR-02 – Containment, Evidence Preservation, Recovery

Field	Details
Assessment Area	AI incident containment and recovery
Assurance Objective	Determine whether affected AI capabilities can be contained and restored while preserving sufficient evidence for investigation.
Illustrative Assessment Activities	Review or simulate approved containment of models, agents, tools, connectors, accounts and retrieval sources; examine session and credential revocation; test quarantine of poisoned content; verify configuration and model identification; review evidence preservation, recovery and re-enablement criteria.
Expected Evidence	Containment procedures, emergency controls, revocation records, evidence-handling requirements, configuration history, recovery plans, backup records and exercise results.
Potential Indicators of Weakness	A compromised agent or connector cannot be disabled independently; poisoned content remains indexed after quarantine; sessions or service credentials cannot be revoked promptly; model and prompt versions cannot be established; recovery restores the same unsafe configuration.

Containment controls should be tested in a manner consistent with the approved safety class. A documented emergency switch provides little assurance if its dependencies, authority and operational consequences have not been examined.

A.15 Resilience and Abuse Resistance

This domain assesses whether the AI system can withstand excessive, malformed or repeated use without uncontrolled cost, service degradation or unsafe failure. Active load testing must remain within the approved safety class and rules of engagement.

A.15.1 AI-RES-01 - Request, Execution, and Consumption Controls

Field	Details
Assessment Area	Resource and abuse controls
Assurance Objective	Determine whether enforceable limits prevent a user, workload or agent from consuming disproportionate AI and integration resources.
Illustrative Assessment Activities	Review limits for requests, tokens, concurrency, model calls, tool invocations and queued work; examine enforcement by user, tenant, workload and credential; test approved threshold behaviour; assess backpressure, cancellation and recovery; verify whether limits apply across distributed entry points.
Expected Evidence	Rate and quota configuration, gateway policies, concurrency controls, queue settings, controlled threshold results, usage records and exception approvals.
Potential Indicators of Weakness	A single identity can exhaust shared capacity; limits apply only to the user interface; several credentials bypass a tenant-wide quota; queued work grows without a bound; rejected requests continue consuming downstream resources.

Usage monitoring is covered by *AI-MON-04*. This module assesses controls that actively limit or contain consumption.

A.15.2 AI-RES-02 - Input Size and Processing Boundaries

Field	Details
Assessment Area	Resource-intensive input processing
Assurance Objective	Establish whether text, files, images and other inputs are constrained before they cause excessive processing, storage or downstream requests.

Field	Details
Illustrative Assessment Activities	Review limits for prompt length, context size, file count, file size, extracted content, archive expansion, media resolution and batch operations; test approved boundary values; examine parser timeouts and cancellation; assess repeated summarisation, embedding and retrieval expansion.
Expected Evidence	Input limits, parser and extraction configuration, timeout settings, context-management logic, queue controls, resource telemetry and controlled boundary results.
Potential Indicators of Weakness	Compressed or nested content expands beyond processing limits; large inputs create unbounded model or embedding calls; client-side limits can be bypassed; timed-out requests continue processing; retrieved context grows without a defined maximum.

File parsing vulnerabilities remain within *AI-APPSEC-03*. This module focuses on computational and operational exhaustion caused by accepted inputs.

A.15.3 AI-RES-03 - Dependency Failure and Graceful Degradation

Field	Details
Assessment Area	Provider and integration resilience
Assurance Objective	Determine whether failure of a model, provider, data source or tool is contained without bypassing security controls or leaving business processes in an unsafe state.
Illustrative Assessment Activities	Identify critical external dependencies; review timeouts, retries, circuit breakers and fallback behaviour; simulate approved unavailable or degraded responses; assess security parity of fallback models and regions; examine queued actions, partial transactions and manual operating procedures.
Expected Evidence	Dependency map, availability design, timeout and retry configuration, fallback rules, recovery procedures, service objectives, controlled failure results and business continuity plans.
Potential Indicators of Weakness	Failure causes unrestricted fallback to an unapproved model; retries duplicate actions; unavailable authorisation or retrieval services cause controls to fail open; partial transactions cannot be reconciled; users receive confident output when required sources are unavailable.

A fallback is not automatically safer than an outage. It should preserve the required data, security and regional controls and make any reduction in capability clear to users and downstream systems.

A.16 AI Development Lifecycle and Supply Chain

This domain assesses whether AI systems are designed, tested, built and released through controlled processes, and whether the components on which they depend are understood and trustworthy. Modules should be applied proportionately to custom development, configured SaaS capabilities and externally supplied AI services.

A.16.1 AI-SDLC-01 - Secure AI Design Review

Field	Details
Assessment Area	Secure design
Assurance Objective	Determine whether security, privacy and operational risks were considered before the system architecture and authority model were established.
Illustrative Assessment Activities	Review security requirements, architecture and threat modelling; identify trust boundaries, sensitive data flows and delegated actions; examine alternatives that reduce access or autonomy; verify that failure, misuse and human oversight were considered during design.
Expected Evidence	Security requirements, architecture decisions, threat model, data-flow diagrams, abuse cases, design reviews and accepted risks.
Potential Indicators of Weakness	Security review occurred only after deployment; trust boundaries are undocumented; the system receives unnecessary data or authority; high-impact actions lack an explicit control design; accepted risks have no accountable owner.

A.16.2 AI-SDLC-02 - Security Testing and Release Validation

Field	Details
Assessment Area	Pre-release assurance
Assurance Objective	Establish whether material AI security controls are tested before release and after relevant changes.
Illustrative Assessment Activities	Review test coverage for permissions, retrieval, prompts, tools, APIs and logging; examine representative release evidence; assess how probabilistic behaviour is tested; review release gates, defect handling and security regression testing; verify independence appropriate to the system’s risk.
Expected Evidence	Test plans, test data, role matrices, results, defect records, release approvals, regression suites and exception records.

Field	Details
Potential Indicators of Weakness	Releases proceed without permission-boundary testing; tests cover only expected prompts; one successful result is treated as proof of consistent behaviour; material defects are accepted without authority; security tests are not repeated after relevant changes.

Test design should account for variable model behaviour while retaining deterministic pass conditions for controls such as authorisation and tool policy.

A.16.3 AI-SDLC-03 - Environment and Deployment Separation

Field	Details
Assessment Area	Build and deployment security
Assurance Objective	Determine whether development, testing and production environments remain appropriately separated throughout build and release.
Illustrative Assessment Activities	Review separation of identities, data, indexes, vector stores, model keys and tools; examine deployment pipelines and approvals; inspect configuration promotion and secret injection; assess whether production data is used outside production; verify rollback and emergency release controls.
Expected Evidence	Environment architecture, IAM configuration, pipeline definitions, deployment records, data-use approvals, secret controls, release history and rollback procedures.
Potential Indicators of Weakness	Non-production users can access production data; environments share privileged credentials or indexes; developers deploy directly to production without review; test agents can invoke production tools; rollback restores insecure configuration.

A.16.4 AI-SUPPLY-01 - Component Provenance and Integrity

Field	Details
Assessment Area	AI supply chain provenance
Assurance Objective	Establish whether the origin, version and integrity of material AI components can be determined and verified.
Illustrative Assessment Activities	Identify models, adapters, datasets, libraries, containers, plugins, tools and MCP servers; review acquisition sources and version controls; examine hashes, signatures or attestations where available; assess unverified downloads and transitive dependencies; inspect integrity controls in build and deployment.

Field	Details
Expected Evidence	Component inventory, source locations, versions, lock files, hashes, signatures, attestations, build records and repository history.
Potential Indicators of Weakness	A model or plugin has unknown origin; mutable or unversioned artefacts are deployed; integrity is not checked before release; dependencies are obtained from untrusted sources; the deployed component cannot be matched to a reviewed version.

Not every component ecosystem supports signed artefacts or formal attestations. Where these are unavailable, the assessment should identify the resulting uncertainty and evaluate alternative provenance controls.

A.16.5 AI-SUPPLY-02 - Third-Party Component and Service Risk

Field	Details
Assessment Area	External component assurance
Assurance Objective	Determine whether third-party models, plugins, connectors, tools and services are assessed and monitored according to the access and trust granted to them.
Illustrative Assessment Activities	Review publisher and service ownership; assess maintenance and vulnerability practices; examine permissions, update mechanisms and transitive services; review incident notification, support and end-of-life arrangements; determine how compromise or withdrawal would be contained.
Expected Evidence	Supplier assessment, contractual terms, security documentation, permission records, update configuration, vulnerability information, subprocessor details and exit plans.
Potential Indicators of Weakness	An abandoned component retains privileged access; automatic updates introduce unreviewed capability; publisher identity is not verified; no process exists for security advisories; the organisation cannot disable or replace a compromised dependency.

A.16.6 AI-AIBOM-01 - AI Bill of Materials Completeness

Field	Details
Assessment Area	AI component inventory
Assurance Objective	Determine whether the AI Bill of Materials accurately represents the components and dependencies of the deployed system.

Field	Details
Illustrative Assessment Activities	Compare the AI Bill of Materials with architecture, cloud resources, application configuration and runtime observations; sample model, prompt, agent, tool, data and provider records; verify ownership, version and environment fields; assess update and review processes.
Expected Evidence	AI Bill of Materials, architecture diagrams, deployment configuration, runtime inventory, model and agent registers, data-source inventory and change records.
Potential Indicators of Weakness	Deployed models or tools are absent; versions and providers are unknown; data sources and service identities are omitted; the inventory does not distinguish environments; updates do not trigger corresponding inventory changes.

The AI Bill of Materials may reference sensitive prompts, credentials or configurations by identifier and version. It should not reproduce secret values or other content that would create an additional security risk.

A.16.7 AI-EVAL-01 - Evaluation Data and Security Test Coverage

Field	Details
Assessment Area	Evaluation integrity
Assurance Objective	Establish whether evaluation data and procedures provide credible coverage of the system’s security requirements and intended operating conditions.
Illustrative Assessment Activities	Review datasets and scenarios used for security evaluation; examine coverage of roles, permissions, data types, languages, retrieval states and tool actions; assess separation between development and evaluation data; review dataset provenance, access and contamination risks; inspect pass criteria and repeated-trial methodology.
Expected Evidence	Evaluation plan, dataset inventory, provenance records, access controls, scenario coverage, pass criteria, results and regression history.
Potential Indicators of Weakness	Evaluation covers only normal user behaviour; role or tenant boundaries are absent; test data has been used to tune the same behaviour being evaluated; results lack defined pass criteria; sensitive evaluation data is inadequately protected.

Fairness and broad model-quality evaluation should be included only where agreed. This module focuses on evidence supporting security and operational assurance.

A.16.8 AI-CODE-01 - AI Coding Assistant and Repository Security

Field	Details
Assessment Area	AI-assisted software development
Assurance Objective	Determine whether coding assistants preserve repository boundaries, protect source material and operate within established development controls.
Illustrative Assessment Activities	Review repository and workspace access; examine what code, files and terminal context are sent to the provider; assess retention and training-use settings; test cross-repository isolation; review generated code validation; inspect terminal, tool and deployment permissions; verify pull-request, review and scanning requirements.
Expected Evidence	Repository permissions, assistant configuration, provider data settings, extension or CLI permissions, tool logs, generated changes, pipeline controls and review records.
Potential Indicators of Weakness	The assistant accesses unrelated repositories; proprietary code or secrets are sent under unsuitable provider terms; generated changes bypass review or scanning; terminal tools run with excessive privilege; development context crosses users or projects.

AI-generated code should be treated as untrusted contribution. Existing peer review, testing, secret scanning and deployment controls should remain enforceable regardless of how the code was produced.

A.17 Advanced and Context-Specific Assurance

A.17.1 AI-REDTEAM-01 - Scenario-Led Adversarial Assessment

Field	Details
Assessment Area	Cross-domain adversarial testing
Assurance Objective	Determine whether realistic attack paths can combine weaknesses across identity, retrieval, prompts, agents, tools and business workflows.
Illustrative Assessment Activities	Develop scenarios from the threat model and *Appendix C* catalogue; define objectives, identities, starting access and safety constraints; conduct controlled attack-path testing; assess prevention, detection and response; record where the scenario was stopped or succeeded.
Expected Evidence	Threat model, scenario plans, rules of engagement, execution records, system and security logs, response actions and attack-path analysis.

Field	Details
Potential Indicators of Weakness	Individually minor weaknesses combine into material impact; security controls operate in isolation but fail across components; defenders cannot observe the attack path; the system permits actions beyond the approved scenario or user authority.

This module normally supports Level 4 assurance. It should not be used to duplicate findings already recorded under a more specific module.

A.17.2 AI-WORKFLOW-01 - Autonomous Workflow Control

Field	Details
Assessment Area	End-to-end workflow security
Assurance Objective	Establish whether AI-enabled workflows remain authorised, bounded and recoverable across multiple systems and decision points.
Illustrative Assessment Activities	Map the workflow from initiation to completion; identify human and automated decisions; test approved changes to order, state, recipient and target; examine partial failure, retries and concurrent execution; verify approval, reconciliation and rollback across connected systems.
Expected Evidence	Workflow diagrams, state models, identity flows, decision rules, transaction logs, approvals, exception records and controlled end-to-end results.
Potential Indicators of Weakness	A workflow skips a required approval; authority is checked only at initiation; one agent's output becomes another agent's trusted instruction; retries duplicate a consequential action; partial completion cannot be detected or reconciled.

AI-AGENT-07 assesses control of an individual agent's plan. This module assesses the wider business workflow in which one or more agents participate.

A.17.3 AI-HUMAN-01 - Human Oversight and Decision Accountability

Field	Details
Assessment Area	Human review of consequential output and action
Assurance Objective	Determine whether human oversight is meaningful, informed and assigned to a person with the authority and information needed to intervene.
Illustrative Assessment Activities	Identify decisions and actions requiring human review; examine what evidence and context reviewers receive; test rejection, modification and escalation paths; assess workload and time pressure; verify responsibility for the final decision and resulting action.

Field	Details
Expected Evidence	Decision and approval design, reviewer roles, interface examples, escalation procedures, action logs, override records and representative review results.
Potential Indicators of Weakness	Reviewers receive insufficient context; approval is presented as a routine confirmation; the reviewer lacks authority to stop the action; no person is accountable for the final decision; automated execution occurs before review is complete.

Human involvement is not an effective control merely because a person appears in the workflow. The review must be timely, informed and capable of changing the outcome.

A.17.4 AI-MULTI-01 - Multimodal Input and Hidden Instruction Security

Field	Details
Assessment Area	Image, audio, video and mixed-content security
Assurance Objective	Determine whether untrusted instructions or sensitive information carried through non-text inputs can alter system behaviour or cross a security boundary.
Illustrative Assessment Activities	Identify accepted modalities and preprocessing stages; use approved images, audio, video or mixed documents containing visible and concealed test instructions; examine OCR, transcription, metadata and extracted context; test whether multimodal content influences retrieval, tools or downstream actions; verify file and modality limits.
Expected Evidence	Modality architecture, preprocessing configuration, controlled test content, extracted text or metadata, model responses, tool records and security logs.
Potential Indicators of Weakness	Instructions in an image or audio file override trusted policy; hidden metadata enters model context without control; sensitive content is extracted and disclosed outside the user's authority; a non-text input causes an unintended tool action.

The existence of hidden or machine-readable content is not itself a vulnerability. A finding requires an unintended security-relevant effect in the deployed workflow.

A.17.5 AI-CONT-01 - Continuous Assurance and Regression Monitoring

Field	Details
Assessment Area	Ongoing AI assurance
Assurance Objective	Establish whether material changes and security regressions are identified after initial assessment and production release.

Field	Details
Illustrative Assessment Activities	Identify changes that trigger reassessment; review automated and manual regression coverage; examine baselines for permissions, retrieval, prompts, models and tools; sample recent changes; assess alerting for configuration drift and failed evaluations; verify ownership of resulting remediation.
Expected Evidence	Continuous assurance plan, change triggers, evaluation suites, baseline results, drift records, alerts, reassessment decisions and remediation records.
Potential Indicators of Weakness	Model or prompt changes bypass regression testing; permission or connector drift is not detected; failed evaluations do not block or escalate a release; tests no longer reflect production architecture; no owner reviews assurance results.

This module supports Level 5 assurance. It should repeat the relevant Level 2–4 activities rather than create a separate set of weaker checks.

A.17.6 AI-IND-01 - Industry and Use-Case Risk Review

Field	Details
Assessment Area	Sector and business-context risk
Assurance Objective	Determine whether assessment scope and controls account for risks specific to the organisation’s sector, operating environment and use case.
Illustrative Assessment Activities	Identify applicable safety, security, privacy, recordkeeping and operational requirements; review sector-specific threat scenarios; examine consequences of incorrect output or action; consult relevant business, legal, privacy, safety and operational owners; tailor modules and evidence requirements accordingly.
Expected Evidence	Applicable requirements, business impact analysis, sector threat information, specialist input, tailored assessment plan, control mappings and risk decisions.
Potential Indicators of Weakness	Generic controls overlook a sector-specific consequence; safety or operational owners were not consulted; assessment scenarios do not reflect the deployed use; applicable requirements are absent from scope; residual risk is accepted without appropriate authority.

This module does not require the security assessor to provide legal, clinical, engineering or safety certification. It requires the assessment to incorporate qualified specialist input where those consequences are relevant.

A.18 Assessment Tailoring Guidance

The complete matrix is not intended to be applied to every AI system. Module selection should follow the system's capabilities, data, exposure, authority and credible attack paths rather than its product name alone.

The system archetypes in *Appendix B* provide an initial classification. The assessor should then use the architecture and threat model to identify relevant modules, select the required assurance level and determine how testing can be performed within the approved safety class.

An assurance level changes the depth of evidence required. A safety class constrains how that evidence may be obtained. Neither should be used to exclude a relevant security objective without documenting an alternative assessment approach.

System characteristic	Modules normally considered
Processes personal, confidential or regulated information	Data Security and Lifecycle; Privacy and Regulatory Considerations; Model and Provider Security; Cloud Storage Security
Retrieves enterprise knowledge	Identity and Boundary Isolation; Retrieval and Knowledge Security; Output Handling; Connector Security
Performs actions through agents or tools	Delegated Identity; Agent and Tool Security; Output-to-Workflow Injection; Action Auditability; Resilience
Uses MCP servers	MCP Server Trust, Authentication and Message Validation; Agent and Tool controls; Connector and Network Security
Accepts external or untrusted input	Application and API Security; Prompt and Orchestration Security; External Abuse Controls; Monitoring and Resilience
Supports consequential decisions or workflows	Decision and Operational Integrity; Human Oversight; Autonomous Workflow Control; Incident Response
Accepts files, images, audio or video	File Processing Security; Multimodal Input Security; Indirect Instruction Injection; Resource Boundaries
Is custom-built or extensively configured	Secure Design; Application and API Security; Cloud Security; Supply Chain; AI Bill of Materials
Changes models, prompts, tools, or data sources frequently	Model Change; Security Regression Testing; Change Management; Continuous Assurance

System characteristic	Modules normally considered
Depends on external models, plugins, connectors or providers	Provider Data Handling; Third-Party Component Risk; Component Provenance; Dependency Resilience
Is shared across tenants, workspaces or environments	Tenant and Workspace Isolation; Data Store Separation; Workload Identity; Environment Separation
Is publicly accessible or exposed to a large user population	Authentication and Session Security; External Exposure and Abuse Controls; Rate and Consumption Controls; Detection and Alerting

Table 6: Indicative module-selection triggers

The final assessment plan should record:

- Included modules and the reason for selection
- Excluded modules and the basis for exclusion
- Modules adapted to the system architecture
- Assurance level and safety class
- Evidence expected for each selected objective
- Dependencies, limitations and untested assumptions

Module exclusion does not establish that the corresponding risk is absent. It records that the area was not applicable, was addressed through another procedure or remained outside the agreed scope.

Appendix B - AI System Archetypes

AI system archetypes provide a practical starting point for assessment planning. They describe common combinations of capabilities, integrations and operating contexts rather than fixed product categories.

A system may match several archetypes. For example, an enterprise copilot may use retrieval, invoke tools and operate within a SaaS platform. Assessors should select all relevant archetypes and then use the system architecture and threat model to determine the applicable Appendix A modules.

An archetype does not determine the system's risk rating or required assurance level. Those decisions depend on its data, users, exposure, authority, business impact and applicable obligations.

Archetype	Typical characteristics	Primary assessment focus
Enterprise Copilot or Assistant	Provides chat, search, summarisation or assistance across enterprise applications and user-accessible content.	Permission boundaries, user and group isolation, connector access, aggregation, output handling and auditability.
Retrieval-Augmented Generation System	Retrieves content from repositories, indexes, vector stores or knowledge services to construct responses.	Source permission enforcement, retrievable-unit access, metadata exposure, ingestion integrity, indirect instruction injection, grounding and citations.
AI Agent or Tool-Calling System	Selects or invokes tools, APIs, plugins or workflows to retrieve information or perform actions. May include single-agent or multi-agent operation.	Delegated identity, least privilege, action authorisation, tool arguments, approval, confused-deputy risk, execution bounds and recovery.
AI Gateway, Proxy, or MCP Broker	Routes requests between users, applications, models and tools while applying policy, authentication or monitoring.	Authentication, authorisation, routing integrity, policy enforcement, capability mediation, tenant isolation, logging and abuse controls.
AI-Enabled SaaS Platform	Provides AI capabilities within a third-party business platform such as productivity, CRM, ITSM, HR or finance.	Tenant configuration, role enforcement, feature enablement, connector governance, provider data handling, administrative access and change visibility.
AI Coding or SDLC Assistant	Accesses source repositories or development environments and may generate code, invoke terminal tools or participate in build and deployment workflows.	Repository boundaries, source-code handling, secrets, generated-code validation, tool permissions, supply-chain risk and enforcement of review and deployment controls.

Archetype	Typical characteristics	Primary assessment focus
AI Decision Support System	Produces recommendations, classifications, scores or summaries used in consequential business or operational decisions.	Data provenance, output integrity, uncertainty, human oversight, traceability, privacy and applicable sector requirements.
Customer-Facing AI System	Accepts input from customers, partners, public users or other untrusted parties through chat, voice or embedded interfaces.	External abuse, authentication, customer isolation, sensitive-data handling, rate controls, unsafe output, escalation and monitoring.
Multimodal AI System	Processes or generates combinations of text, images, audio, video, documents or other media.	File processing, hidden instructions, preprocessing, cross-modal data exposure, output handling, resource limits and modality-specific logging.
Model Development or Adaptation Platform	Trains, fine-tunes, adapts, evaluates or hosts custom models and related artefacts.	Training-data provenance, model and adapter integrity, evaluation coverage, artefact protection, environment separation, supply chain and change control.
Custom AI Application	Uses one or more models within a purpose-built application, API, analytics service, classifier, assistant or workflow.	Application and API security, identity, orchestration, data handling, model routing, output validation, cloud configuration and monitoring.

Table 7: EAISA AI system archetypes

Archetype selection should be recorded during system discovery and revisited if the system's capabilities change. Enabling retrieval, adding a connector, introducing persistent memory or allowing an assistant to perform actions may add a new archetype and require additional assessment modules.

Appendix C - Enterprise AI Threat Scenario Catalogue

This catalogue contains representative abuse paths for enterprise AI threat modelling and scenario-led assessment. It is not exhaustive, and not every scenario applies to every system.

Assessors should select and adapt scenarios according to the system archetype, architecture, data, identities, integrations and business impact. Execution must remain within the approved rules of engagement and testing safety class. Any resulting findings should be mapped to the relevant Appendix A modules.

Id	Threat scenario	Representative abuse path	Potential consequence
EAISA-TS-01	Cross-Boundary Retrieval	A user obtains content through AI search, chat or summarisation that they cannot access in the source system.	Unauthorised disclosure of documents, records or protected facts.
EAISA-TS-02	Cross-Tenant or Workspace Exposure	Shared indexes, service identities or weak scope enforcement allow content to cross tenant, workspace, project or environment boundaries.	Broad data exposure and loss of tenant isolation.
EAISA-TS-03	Metadata and Reference Disclosure	Filenames, titles, labels, paths, result counts, snippets or citations reveal the existence or context of restricted material.	Disclosure of sensitive projects, customers, investigations or records without exposing the full content.
EAISA-TS-04	Stale Access Exposure	Revoked, deleted or reclassified information remains available through an index, cache, session, memory or derived store beyond the approved propagation period.	Continued access after the source permission or lifecycle control has changed.
EAISA-TS-05	Aggregation and Inference Exposure	The system combines authorised information from several sources to reveal a protected fact or materially more sensitive conclusion.	Defeat of an intended data boundary or disclosure that was impractical through the source systems alone.
EAISA-TS-06	Sensitive AI Telemetry Exposure	Prompts, responses, retrieved context, files or tool outputs are stored in logs or monitoring platforms with broader access than the source information.	Secondary exposure through administrators, support personnel, exports or security tooling.

Id	Threat scenario	Representative abuse path	Potential consequence
EAISA-TS-07	Direct Prompt Control Bypass	User instructions override trusted instructions and cause a security-relevant policy, data or capability boundary to fail.	Unauthorised disclosure, prohibited capability use or manipulation of downstream behaviour.
EAISA-TS-08	Indirect Instruction Injection	Instructions embedded in documents, websites, messages or tool results are treated as trusted instructions by the AI system.	Altered responses, data disclosure, redirected retrieval or unintended tool execution.
EAISA-TS-09	Retrieval Poisoning	An attacker introduces or modifies content in a trusted knowledge source to influence later retrieval and generation.	Persistent misinformation, manipulated decisions or indirect instruction injection.
EAISA-TS-10	Persistent Memory Manipulation	Untrusted content creates or changes durable agent memory that influences later users, sessions or actions.	Persistent control manipulation, cross-user exposure or repeated unsafe behaviour.
EAISA-TS-11	Confused Deputy Execution	A user causes an agent or privileged service identity to perform an action the user could not perform directly.	Privilege misuse, cross-user modification or unauthorised access to enterprise systems.
EAISA-TS-12	Tool Selection or Argument Manipulation	User or retrieved content changes the selected tool, object identifier, recipient, path, query or execution parameter.	Action against an unintended target or execution outside the approved task.
EAISA-TS-13	Approval or Confirmation Bypass	An agent performs a consequential action without valid approval, changes material parameters after approval or reuses approval for another action.	Unauthorised transactions, communications, record changes or configuration changes.
EAISA-TS-14	Multi-Step Goal Expansion	An agent expands an approved task into additional steps that exceed the original objective or authority.	Unauthorised downstream actions that were not apparent when the task began.
EAISA-TS-15	Repeated or Non-Idempotent Execution	Retries, loops or replay cause an agent to repeat messages, transactions, updates or tool calls.	Duplicate actions, inconsistent records, operational disruption or uncontrolled cost.
EAISA-TS-16	Unsafe Agent Publication	An unreviewed or compromised agent is published to a broad user	Scaled exposure of unsafe capability, excessive permissions or manipulated instructions.

Id	Threat scenario	Representative abuse path	Potential consequence
		population with sensitive data or tool access.	
EAISA-TS-17	Malicious or Excessive MCP Capability	An unapproved or compromised MCP server exposes misleading, unnecessary or privileged tools and resources to an AI client.	Data theft, unintended execution or extension of agent authority through a trusted integration.
EAISA-TS-18	Unsafe Output Rendering	Generated markup, links, files or other content is rendered without controls appropriate to the destination.	Script execution, deceptive navigation, unsafe resource loading or compromise of a downstream user interface.
EAISA-TS-19	Output-to-Workflow Injection	Generated text or structured output is passed into a command, API, ticket, email, query or automation without sufficient validation.	Manipulation of workflow state, recipients, commands or downstream business logic.
EAISA-TS-20	Multimodal Hidden Instruction	Instructions or sensitive content concealed in images, audio, video, metadata or mixed documents influence model or tool behaviour.	Security-control bypass, unintended action or cross-modal data exposure.
EAISA-TS-21	Decision-Support Manipulation	Poisoned, incomplete or misleading data causes the system to produce an unsupported recommendation that is trusted by a consequential workflow.	Harmful financial, employment, legal, health, safety or operational decisions.
EAISA-TS-22	Resource and Cost Exhaustion	Repeated requests, large inputs, recursive agents or excessive tool use consume model, processing or provider resources.	Service degradation, denial of service or material financial loss.
EAISA-TS-23	Provider or Region Misrouting	Routing, fallback or configuration sends data to an unapproved model, provider, tenant or processing region.	Breach of data-handling requirements, contractual commitments or regional restrictions.
EAISA-TS-24	Component or Model Supply-Chain Compromise	A model, adapter, plugin, connector, package or MCP server is replaced, modified or updated through an untrusted source.	Compromised output, concealed data access, malicious tool behaviour or persistent system control.
EAISA-TS-25	Security Regression after Change	A model, prompt, permission, connector, tool or index	Reintroduction of data exposure, unsafe actions

Id	Threat scenario	Representative abuse path	Potential consequence
		change invalidates previously tested controls.	or weakened monitoring after release.
EAISA-TS-26	Unattributable AI Action	An agent or shared service performs an action that cannot be linked to the initiating user, approval, prompt or tool call.	Delayed investigation, disputed accountability and inability to determine the scope of misuse.

Table 8: Enterprise AI threat scenario catalogue

Threat scenarios should be converted into system-specific test cases with defined preconditions, identities, targets, expected controls and stop conditions. A scenario may combine several Appendix A modules, but any confirmed weakness should be reported once under the module that best represents its root cause.

Appendix D - Secure Architecture Patterns and Common Anti-Patterns

These patterns support secure enterprise AI design but do not guarantee that a system is secure. Their suitability depends on the system architecture, business purpose, data, authority and operational constraints. Each pattern should be validated in the deployed implementation.

D.1 Secure Architecture Patterns

Pattern	Security Benefit and implementation intent
End-to-end identity and security-context propagation	Preserves the initiating user, tenant, role and approved task across application, retrieval, agent and tool boundaries. Downstream systems should independently validate the supplied context.
Permission-aware retrieval	Enforces effective source-system permissions before content reaches the model or user. Enforcement may use replicated ACLs, query-time checks or another equivalent server-side control.
Isolation aligned to permission domains	Reduces cross-boundary exposure by separating or strongly partitioning indexes, vector stores, caches and service identities according to tenant, workspace, environment or data sensitivity.
Security policy enforcement outside the model	Places access control, approval, transaction limits and other critical decisions in deterministic application, policy or target-system controls.
Explicit instruction provenance	Distinguishes system and developer instructions from user input, retrieved content and tool output so that untrusted material is not presented as authoritative policy.
Least-privileged delegated execution	Uses delegated user authority where practical or a constrained service identity with independent per-action authorisation and user attribution.
Capability mediation and tool allowlisting	Exposes only the tools, MCP capabilities and operations required by the approved agent and user context.
Prepare, approve and execute for consequential actions	Shows the user the intended operation, target and material parameters before execution and binds the approval to that specific action.
Server-side argument and business-rule validation	Validates tool and API inputs against the authenticated identity, approved task and current system state rather than relying only on generated schemas.

Pattern	Security Benefit and implementation intent
Controlled data ingestion and provenance	Restricts who and what can contribute to retrieval sources, records origin and version, and provides a way to quarantine manipulated content.
Data-minimised security telemetry	Records identities, policy decisions, retrieval events, tool actions, approvals and outcomes while avoiding unnecessary storage of sensitive prompts and responses.
Correlated and attributable audit records	Links user activity, sessions, retrieval, model calls, agent plans, tool invocations, approvals and target-system outcomes using reliable identifiers.
Restricted outbound connectivity	Limits AI workloads and tools to approved providers, APIs, destinations and network paths, reducing data-exfiltration and server-side request risks.
Environment and credential separation	Separates production and non-production data, indexes, credentials, agents, tools and deployment authority.
Versioned change and regression control	Tracks changes to models, prompts, permissions, tools, connectors and retrieval sources and retests the controls affected by those changes.
Tested containment and recovery controls	Allows affected models, agents, tools and connectors to be isolated or disabled while preserving evidence and maintaining safe business operation.

Table 9: Secure enterprise AI architecture patterns

D.2 Common Architecture Anti-Patterns

Anti-Pattern	Resulting risk
The model decides whether access is permitted	Prompt manipulation or inconsistent model behaviour can become an authorisation bypass.
Global index with weak or client-controlled filtering	Content may cross user, tenant, workspace or classification boundaries.
Shared over-privileged execution identity	Low-privileged users may exercise permissions they do not hold directly, while actions become difficult to attribute.
Retrieved content treated as trusted instruction	Documents, websites or tool results can redirect model behaviour or influence downstream actions.
All tools exposed to every agent or MCP client	Compromise or manipulation of one agent provides unnecessary access to sensitive capabilities.

Anti-Pattern	Resulting risk
Approval detached from the executed action	Parameters, targets or action types may change after approval, or the same approval may be replayed.
Generated output passed directly to trusted workflows	Model-controlled text or structured data can alter commands, API calls, records, routing or business logic.
Generated markup rendered without destination-specific controls	Output may enable unsafe scripts, links, resource loading or deceptive content.
Full prompts and responses retained by default	Sensitive information may accumulate in logs, transcripts, monitoring platforms and support systems.
Production and non-production share trust boundaries	Test users, workloads or compromised development systems may reach production data or capabilities.
Unrestricted model or provider fallback	Data may be routed to an unapproved model, provider or processing region when the primary service fails.
Mutable models, prompts or connectors without regression testing	Previously validated controls may fail after an update without detection.
Agent actions recorded only under a service account	The initiating user, approval and business purpose cannot be reliably reconstructed.
High-impact capabilities enabled broadly by default	Users gain access to sensitive retrieval, agents or tools before ownership, permissions and monitoring are established.
Emergency controls exist only on paper	Responders may be unable to contain an unsafe agent, poisoned source or compromised connector during an incident.

Table 10: Common enterprise AI architecture anti-patterns

A pattern should be judged by the security outcome it provides, not by the technology used to implement it. Alternative designs may provide equivalent assurance where identity, authorisation, isolation, attribution and recovery remain effective.

Appendix E - Relationship to OWASP AI Testing Guide and MITRE ATLAS

EAISA was developed with reference to established public AI security resources, including the OWASP AI Testing Guide and MITRE ATLAS. These resources serve different but complementary purposes.

The OWASP AI Testing Guide provides a technology-agnostic methodology for testing AI trustworthiness across application, model, infrastructure and data layers. MITRE ATLAS is a living knowledge base of adversary tactics and techniques affecting AI-enabled systems.

EAISA uses these resources to inform threat modelling and test design. Its specific role is to structure enterprise security assessments around deployed identities, permissions, data sources, retrieval systems, agents, tools, integrations and business workflows.

This appendix describes high-level relationships only. It does not claim formal conformance, complete coverage or one-to-one equivalence with either resource.

Area	OWASP AI Testing Guide	MITRE ATLAS	EAISA Methodology Contribution
Prompt and instruction manipulation	Provides test approaches for adversarial input and trustworthiness failures.	Describes relevant adversary behaviours and techniques.	Connects manipulation to enterprise data access, policy enforcement, retrieval and tool execution.
Sensitive information exposure	Addresses privacy, confidentiality and information leakage across AI components.	Supports threat modelling for discovery, collection and exfiltration behaviours.	Tests effective user, role, tenant and repository boundaries and defines evidence requirements for confirmed exposure.
Model-level attacks	Includes model robustness, manipulation and trustworthiness testing.	Provides tactics and techniques for attacks against models and AI supply chains.	Applies model-level testing where it is within scope, particularly for custom, adapted or self-hosted models.
Retrieval and knowledge security	Covers AI data, retrieval and trustworthiness concerns.	Provides techniques relevant to poisoning, manipulation and information access.	Examines source permission enforcement, retrievable-unit isolation, metadata, ingestion integrity, indirect instructions and traceability.
Agent and tool security	Addresses autonomous behaviour, unsafe	Supports modelling of adversary actions involving AI-	Follows delegated authority through agents, tools, service identities, approvals, target systems and recovery processes.

Area	OWASP AI Testing Guide	MITRE ATLAS	EAISA Methodology Contribution
	agency and AI application testing.	enabled systems and connected capabilities.	
MCP and capability brokers	Relevant test principles apply through application, agent and integration testing.	Relevant techniques may be used to model compromise or abuse of connected capabilities.	Defines specific modules for MCP server trust, capability exposure, authentication, authorisation and message validation.
Enterprise permission boundaries	General access-control and data-security principles apply.	Relevant adversary techniques inform attempts to cross boundaries.	Provides role-based, cross-user, cross-tenant and stale-permission validation for enterprise repositories and SaaS platforms.
Testing safety	Provides test methodology and trustworthiness-testing considerations.	Not intended to define rules of engagement.	Defines safety classes, excluded activities, operational limits and stop conditions for enterprise assessment.
Evidence confidence	Supports repeatable testing and evidence-based conclusions.	Provides structured threat terminology rather than an evidence-rating system.	Defines confidence levels for observed, reproduced, corroborated, systemic, and impact-demonstrated behaviour.
Risk rating	Addresses AI risk and trustworthiness outcomes.	Describes adversary behaviour but is not a finding-severity methodology.	Supplements conventional scoring with deployment-specific likelihood, impact, and validated-control factors.
Production readiness	Provides testing that can support deployment decisions.	Provides threat information that can inform readiness and monitoring.	Defines minimum security evidence and risk decisions for enterprise go-live approval.
Continuous assurance	Supports lifecycle testing and reassessment.	As a living knowledge base, supports threat-informed monitoring and scenario updates.	Applies regression testing and reassessment to changes in models, prompts, data, permissions, agents and tools.

Table 11: High-level relationship between EAISA, the OWASP AI Testing Guide and MITRE ATLAS

The table identifies complementary areas of use. It should not be interpreted as a measurement of which resource provides greater coverage.

EAISA should be used with the public resources relevant to the system and assessment objective. Threat techniques from MITRE ATLAS may inform scenarios, OWASP guidance may inform test design, and governance or control frameworks may inform expected outcomes. EAISA provides the enterprise assessment structure through which those sources can be applied and evidenced.