

ODA3 INSTITUTE

SECTOR IMPLEMENTATION GUIDANCE

GAISSF™ Government Sector Implementation Guidance

Informative, evidence-driven implementation guidance for government security, governance, legal, procurement, audit, programme and oversight stakeholders

GAISSF-SEC-041 | Draft for Publication — Public — External Review Open

Document Control

Document ID	GAISST-SEC-041
Title	GAISST TM Government Sector Implementation Guidance
Doc Type	Sector implementation guidance — informative
Framework	GAISST TM v1.0 — Global AI Safety and Security Framework (59 controls across nine domains)
Normative source	GAISST-NOR-001 (Framework Standard); GAISST-NOR-004 (Control Catalogue)
Status	Draft for Publication (Public, External Review Open). Open gates: jurisdiction-specific legal and public-law review; accessibility, equality and affected-person rights review; current-policy verification; trademark-status verification.
Dates	Issued 30 June 2026 (research cut-off 30 June 2026); updated 05 October 2026
Companion	GAISST TM Government Sector Executive Brief
Review cycle	At least annually and on material legal, threat, technology or GAISST change
Contact	contact@oda3.org
Classification	Public (on release)

GAISSTTM Government Sector Implementation Guidance

Informative sector implementation guidance

Normative status: This guide is informative and subordinate to GAISST-NOR-001. It is not legal advice, does not determine whether a deployment is lawful, and does not guarantee security, safety, fairness, accuracy, availability or regulatory compliance. Profile membership is defined in GAISST-NOR-001. GAISST certification is not currently offered. © 2026 ODA3 Pvt Ltd. GAISST Ecosystem developed by ODA3 Institute, licensed under GEL v1.0.

1. Executive Summary

The central government-sector AI risk is not the mere presence of an AI model. It is the unsafe, insecure, opaque, or unlawful translation of AI output into public authority, public action, public records, or material consequences for people and institutions.

Government organisations should distinguish informational assistance, recommendation or prioritisation, constrained operational action, and material or determinative public decisions. Control intensity should rise with decision consequence, data sensitivity, autonomy, affected-person impact, and difficulty of correction.

Accountability remains with the authorised public body and responsible officials. Procurement, outsourcing, managed services, model hosting, or supplier certification do not transfer statutory or public accountability to a vendor.

Not every government AI system is inherently high risk. Risk becomes materially greater where systems affect rights, eligibility, liberty, legal status, public safety, taxation, licensing, immigration, enforcement, identity, access to essential services, or authoritative government records.

Five decisions for senior leadership

1. Define permitted, restricted, and prohibited AI uses and exception authority.
2. Assign accountable owners for authority, operation, security, data, records, suppliers, and affected-person safeguards.
3. Require a documented decision-authority level for every use case.
4. Set minimum evidence and independent-assurance requirements for material public-impact systems.
5. Fund continuity, appeal, correction, incident response, and decommissioning.

Notably Absent — executive view

This guide does not assert that all government AI is high risk; that human review is inherently effective; that model accuracy establishes legality or fairness; that vendor certification transfers accountability; or that framework adoption alone proves operational security.

2. Purpose, Scope and Audience

This guide translates the authoritative GAISST control baseline into government-sector implementation guidance, evidence expectations, assessment considerations, examples, and operating records. It does not alter normative GAISST requirements.

- central/federal, state/provincial, regional, and local government
- ministries, departments, agencies, regulators, authorities, and commissions
- shared services and government-controlled entities where relevant
- benefits, taxation, licensing, grants, procurement, regulatory, public-safety, emergency, border, justice-administration, public-health-administration, education-administration, identity, citizen-service, records, cybersecurity, and policy-support functions

Boundary conditions

Defence, intelligence, law-enforcement, healthcare, education, and critical-infrastructure activities may require specialist sector guides and additional controls.

Intended audience

Government CISOs, security architects, AI governance leads, chief data officers, digital-transformation leaders, procurement officials, legal and compliance personnel, internal auditors, inspectors general, programme owners, standards participants, and risk committees.

3. How to Use This Guide

1. Inventory the AI system, use case, supplier, data, integration, decision role, affected parties, and environment.
2. Assign a decision-authority level and public-impact classification before production use.
3. Map all 59 GAISST controls and document applicability; no control may be silently omitted.
4. Implement sector interpretations without changing authoritative control text.
5. Collect operating-effectiveness evidence, not only policy or design intent.
6. Record legal and policy dependencies, exceptions, residual risk, review dates, and approvals.
7. Use the workbook and disposition log to prepare for independent assessment.

4. Government Operating Context

Government AI operates within delegated authority, administrative procedure, records obligations, procurement constraints, rights protections, accessibility duties, public accountability, and continuity expectations. These vary by jurisdiction and function.

Operating characteristic	Implementation consequence
Delegated public authority	System roles must remain within lawful delegation; authority must not emerge implicitly from automation.
Affected-person consequences	Notice, reasons, review, correction, appeal, and redress may be required.
Public records and audit	Inputs, outputs, versions, approvals, and actions may require preservation and retrieval.
Procurement dependence	Contracts must provide evidence, change control, incident notification, audit rights, portability, and exit.
Continuity obligation	Manual fallback and degraded-mode service must be tested where interruption would cause harm.
Public trust	Disclosure, communications integrity, accessibility, and documented limitations are operational controls.

5. Government AI Use-Case Taxonomy

ID	Use case	Function	Role	Authority	Impact
UC-01	Citizen-facing conversational assistants	Citizen services	Advisory	2	Medium to High
UC-02	Benefits eligibility and prioritisation	Social protection	Recommendatory	4	High
UC-03	Fraud, waste, and abuse detection	Integrity and assurance	Recommendatory	3	High
UC-04	Tax risk scoring and audit selection	Revenue administration	Recommendatory	4	High
UC-05	Licensing and permit processing	Regulatory administration	Recommendatory	4	High
UC-06	Public procurement analysis	Procurement	Advisory	3	High
UC-07	Grant application screening	Public funding	Recommendatory	4	High
UC-08	Regulatory inspection prioritisation	Regulation	Recommendatory	4	High

ID	Use case	Function	Role	Authority	Impact
UC-09	Policy modelling and forecasting	Policy	Advisory	2	Medium
UC-10	Document classification and records management	Information management	Operational	2	Medium
UC-11	Public-records processing	Transparency	Advisory	3	High
UC-12	Translation and accessibility services	Citizen access	Advisory	2	Medium
UC-13	Identity verification and biometric matching	Digital identity	Recommendatory	4	Very High
UC-14	Public safety decision support	Public safety	Recommendatory	4	Very High
UC-15	Emergency response resource allocation	Emergency management	Operational	4	Very High
UC-16	Immigration, customs, and border processing	Border administration	Recommendatory	4	Very High
UC-17	Justice-sector administrative support	Justice administration	Advisory	4	Very High
UC-18	Government HR screening and workforce analytics	Workforce	Recommendatory	4	High
UC-19	Cybersecurity monitoring and incident analysis	Cybersecurity	Operational	3	High
UC-20	Intelligence summarisation using authorised data	Analysis	Advisory	3	Very High
UC-21	Code generation and development support	Digital delivery	Advisory	2	High
UC-22	Automated drafting of notices and decisions	Administration	Advisory	4	Very High
UC-23	Regulatory enforcement support	Enforcement	Recommendatory	4	Very High

ID	Use case	Function	Role	Authority	Impact
UC-24	Public consultation and sentiment analysis	Democratic participation	Advisory	3	High
UC-25	Synthetic media for government communications	Communications	Operational	3	High
UC-26	Inter-agency data matching	Shared services	Operational	4	Very High
UC-27	Vendor-hosted AI services	Cross-government	Operational	3	High
UC-28	General-purpose AI assistants for officials	Cross-government	Advisory	2	High

6. Government AI Risk Taxonomy

Risk ID	Risk	Category	Rating	Operational description
R-AUTH-01	Unlawful delegation of statutory authority	Authority and legality	High	AI output is treated as binding without a valid delegation or accountable official decision.
R-RIGHTS-01	Discriminatory or disparate effects	Rights and public impact	High	Model errors or data bias produce unequal access, burden, delay, scrutiny, or denial.
R-SEC-01	Prompt injection and malicious content ingestion	Security	High	Untrusted content alters instructions, causes data disclosure, or triggers unauthorised actions.
R-DATA-01	Inaccurate, incompatible, or unlawful data use	Data and records	High	Data lacks provenance, is stale, or is used outside authorised purpose.
R-PROC-01	Opaque supplier and weak contractual control	Procurement and vendor	High	Government cannot inspect, test, preserve evidence, or manage supplier changes.
R-OPS-01	Automation bias and ineffective human review	Operational	High	Reviewers defer to AI without sufficient competence, time, authority, or evidence.

Risk ID	Risk	Category	Rating	Operational description
R-INST-01	Erosion of public trust and institutional accountability	Democratic and institutional	High	AI use is opaque, politically sensitive, misleading, or not contestable.
R-CONT-01	Loss of public-service continuity	Operational	High	AI dependency prevents service delivery or manual fallback during outage or degradation.
R-PHYS-01	Unsafe physical action	Physical safety	Very High	AI commands affect people, vehicles, facilities, or equipment outside safe limits.

Risk treatment should identify the government function, lawful authority, data, users, affected persons, threat actors, failure modes, detectability, reversibility, service continuity, and evidence required to accept residual risk.

7. Decision-Authority Model

Level	Permitted role	Minimum expectations
Level 1 — Informational assistance	Drafting, search, summarisation, translation, or administration; no direct material decision.	Logging, data controls, training, output labelling, periodic review.
Level 2 — Recommendation or prioritisation	AI ranks, scores, flags, or recommends; an authorised official independently decides.	Decision records, reviewer competence, override, explanation, monitoring, escalation.
Level 3 — Constrained operational action	AI performs predefined reversible actions inside approved limits.	Least privilege, action logs, change control, kill switch, rollback, incident response.
Level 4 — Material or determinative public decision	AI materially influences or executes decisions affecting rights, status, enforcement, benefits, taxation, licensing, immigration, safety, or essential services.	Jurisdiction-specific legal validation, independent assurance, reasons, notice, contestability, authority, records, enhanced monitoring, executive risk acceptance.

Level 4 is not presumed lawful or appropriate.

8. Public-Law, Rights and Accountability Considerations

Document the legal or administrative basis for the function and the official who retains decision authority.

Separate advisory AI output from official reasons and decisions.

Assess notice, explanation, accessibility, procedural fairness, equality, contestability, correction, appeal, and remedy requirements.

Provide oversight bodies sufficient evidence to inspect the system, supplier dependencies, decisions, exceptions, incidents, and remediation.

9. Data, Records, Privacy and Information Management

Maintain inventories of datasets, prompts, retrieval sources, models, versions, tools, integrations, logs, and records classifications.

Preserve evidence needed to reconstruct material decisions.

Do not enter controlled or protected information into unapproved services.

Retention follows the longest applicable statutory, archival, litigation-hold, security, contractual, or GAISST requirement.

10. Procurement and Third-Party Governance

Contracts should cover authorised use, prohibited use, data ownership, training restrictions, confidentiality, security, vulnerabilities, incidents, subcontractors, changes, audits, evidence, logging, retention, continuity, portability, exit, deletion, and suspension.

Supplier certifications are due-diligence inputs, not substitutes for verification.

Government should retain suspension, evidence-preservation, migration, and continuity rights.

11. Security Architecture Considerations

Use environment separation, least privilege, identity-bound access, approved data paths, prompt isolation, secure retrieval, egress control, secrets management, tool allowlists, monitoring, and recovery.

Public-facing systems require abuse controls, filtering, rate limiting, prompt-injection testing, leakage prevention, and human escalation.

Agentic systems require action boundaries, transaction limits, approvals, rollback, and emergency stop.

12. Human Oversight and Administrative Accountability

Human involvement is effective only when reviewers are competent, able to challenge the output, given sufficient time, and provided evidence.

Measure override rates, review quality, reversals, appeal outcomes, workload, and automation-bias indicators.

A nominal approval click is not meaningful oversight.

13. Transparency, Notice, Explanation, Contestability and Appeal

Identify AI interaction where relevant and distinguish automated content from approved government decisions.

Material decisions should provide appropriate reasons and routes for review or correction.

Explanations must be useful; technical feature importance alone may be insufficient.

14. Incident Management, Escalation and Continuity

Define incidents for security compromise, harmful output, unlawful action, data exposure, supplier change, disparity, records failure, outage, and physical safety.

Preserve evidence, constrain unsafe operation, notify accountable officials, correct records and decisions, and communicate where required.

Test manual fallback, degraded operation, supplier failure, model withdrawal, data restoration, and emergency shutdown.

15. GAISSE Government-Sector Control Interpretations

Every authoritative control ID and title is preserved. Sector treatment is informative; authoritative control detail remains in GAISSE-NOR-001.

D1-CTL-01 — Dataset Provenance & Poisoning Prevention

Field	Government-sector treatment
Domain	D1: MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS
Authoritative title	DATASET PROVENANCE & POISONING PREVENTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to dataset lineage, authorised sources, integrity and poisoning controls. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Maintain a dataset register with source, licence, collection purpose, transformations, hashes, approvals and poisoning-test results; quarantine unverified data before training or retrieval use.
Minimum evidence	dataset register; source approvals; hashes; poisoning-test report; quarantine record
Enhanced evidence	dataset register; source approvals; hashes; poisoning-test report; quarantine record; independent test evidence; affected-person or service-impact analysis where material
Assessment	Sample source records, reproduce integrity checks, and confirm unverified data cannot enter production pipelines.
Common failure modes	Untraceable datasets, unverifiable licences, missing hashes, or poisoning scans performed only after deployment.
Related risks	Data and records; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D1-CTL-02 — Model Extraction Resistance

Field	Government-sector treatment
Domain	D1: MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS
Authoritative title	MODEL EXTRACTION RESISTANCE
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to resistance to model extraction through rate, query and output controls. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define extraction threat scenarios; apply authentication, rate limits, query-pattern detection, output minimisation and response actions for suspicious harvesting.
Minimum evidence	extraction threat model; rate-limit policy; anomalous-query alerts; response logs
Enhanced evidence	extraction threat model; rate-limit policy; anomalous-query alerts; response logs; independent test evidence; affected-person or service-impact analysis where material
Assessment	Run controlled extraction attempts and verify detection, throttling and escalation without materially blocking legitimate public access.
Common failure modes	Unlimited high-volume querying, excessive confidence/logit exposure, or no investigation path for extraction alerts.
Related risks	Data and records; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D1-CTL-03 — Behavioral Drift Detection

Field	Government-sector treatment
Domain	D1: MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS
Authoritative title	BEHAVIORAL DRIFT DETECTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to behavioural drift detection across population, language, season and policy changes. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Establish approved baselines and monitor performance, safety, fairness and refusal behaviour by relevant cohort; define drift thresholds, review cadence and rollback criteria.
Minimum evidence	baseline report; drift metrics; cohort analysis; alert history; rollback decision record
Enhanced evidence	baseline report; drift metrics; cohort analysis; alert history; rollback decision record; independent test evidence; affected-person or service-impact analysis where material

Field	Government-sector treatment
Assessment	Introduce representative distribution shifts and verify alerts, investigation and corrective action occur within approved thresholds.
Common failure modes	Only aggregate accuracy is monitored; drift thresholds are undocumented; affected cohorts are invisible.
Related risks	Data and records; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D1-CTL-04 — Federated Learning Poisoning Prevention

Field	Government-sector treatment
Domain	D1: MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS
Authoritative title	FEDERATED LEARNING POISONING PREVENTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to poisoning resistance in federated or distributed training. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Authenticate participants, validate updates, use robust aggregation, detect anomalous gradients and retain participant/update provenance.
Minimum evidence	participant register; update signatures; anomaly reports; aggregation configuration; exclusion decisions
Enhanced evidence	participant register; update signatures; anomaly reports; aggregation configuration; exclusion decisions; independent test evidence; affected-person or service-impact analysis where material
Assessment	Submit malicious or anomalous updates in a controlled test and confirm they are detected, contained and excluded.
Common failure modes	Unauthenticated participants, blind aggregation, or inability to reconstruct which update changed the model.
Related risks	Data and records; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D1-CTL-05 — Embedding Space Robustness

Field	Government-sector treatment
Domain	D1: MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS
Authoritative title	EMBEDDING SPACE ROBUSTNESS
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.

Field	Government-sector treatment
Government interpretation	Government entities should apply this control to robustness of embeddings and vector retrieval against manipulation and collision. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Validate embedding models and indexes; test adversarial nearest-neighbour manipulation, tenant separation, poisoned documents and retrieval relevance for protected workflows.
Minimum evidence	embedding evaluation; index access controls; poisoned-document tests; retrieval logs
Enhanced evidence	embedding evaluation; index access controls; poisoned-document tests; retrieval logs; independent test evidence; affected-person or service-impact analysis where material
Assessment	Seed adversarial and cross-tenant documents and verify isolation, retrieval integrity and alerting.
Common failure modes	Shared indexes without tenant controls, no retrieval attack testing, or silent index replacement.
Related risks	Data and records; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D1-CTL-06 — Post-Quantum Model Signing & Crypto Hardening

Field	Government-sector treatment
Domain	D1: MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS
Authoritative title	POST-QUANTUM MODEL SIGNING & CRYPTO HARDENING
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to cryptographic authenticity and migration readiness for model artifacts. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Sign model artifacts and manifests, protect signing keys, verify signatures before load, inventory algorithms and maintain a risk-based cryptographic migration plan.
Minimum evidence	signed manifest; key-management record; verification logs; algorithm inventory; migration plan
Enhanced evidence	signed manifest; key-management record; verification logs; algorithm inventory; migration plan; independent test evidence; affected-person or service-impact analysis where material
Assessment	Attempt to load altered and unsigned artifacts and confirm rejection; review migration triggers and key-rotation evidence.
Common failure modes	Signatures not checked at runtime, shared signing keys, or unsupported post-quantum claims presented as current assurance.
Related risks	Data and records; Security; Operational

Field	Government-sector treatment
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D1-CTL-07 — LoRA/Adapter Integrity Verification

Field	Government-sector treatment
Domain	D1: MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS
Authoritative title	LORA/ADAPTER INTEGRITY VERIFICATION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to integrity and compatibility of LoRA, adapters and other parameter-efficient modifications. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Treat adapters as executable model changes: approve sources, hash and scan artifacts, test compatibility and safety, and bind each adapter to an authorised base model/version.
Minimum evidence	adapter inventory; hashes; compatibility tests; approval record; deployment binding
Enhanced evidence	adapter inventory; hashes; compatibility tests; approval record; deployment binding; independent test evidence; affected-person or service-impact analysis where material
Assessment	Load an altered or mismatched adapter and verify rejection or safe quarantine.
Common failure modes	Adapters deployed outside change control or accepted solely because the base model was approved.
Related risks	Data and records; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D1-CTL-08 — Model Merge Attack Detection

Field	Government-sector treatment
Domain	D1: MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS
Authoritative title	MODEL MERGE ATTACK DETECTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to detection of unsafe or malicious behaviour introduced through model merging. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.

Field	Government-sector treatment
Implementation guidance	Record component models and merge method, scan components, compare pre/post-merge behaviour and require approval before release.
Minimum evidence	merge manifest; component provenance; differential evaluation; approval; rollback package
Enhanced evidence	merge manifest; component provenance; differential evaluation; approval; rollback package; independent test evidence; affected-person or service-impact analysis where material
Assessment	Perform differential tests for backdoors, capability shifts and safety regressions against component baselines.
Common failure modes	Merged models lack component lineage or are approved on benchmark averages that conceal targeted regressions.
Related risks	Data and records; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D1-CTL-09 — Quantization Backdoor Screening

Field	Government-sector treatment
Domain	D1: MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS
Authoritative title	QUANTIZATION BACKDOOR SCREENING
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to backdoor and safety regression screening after quantisation or compression. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Treat quantisation as a material model change; compare full-precision and quantised behaviour, test trigger sets and preserve conversion parameters and toolchain provenance.
Minimum evidence	conversion manifest; tool versions; differential tests; trigger-set results; approval
Enhanced evidence	conversion manifest; tool versions; differential tests; trigger-set results; approval; independent test evidence; affected-person or service-impact analysis where material
Assessment	Execute targeted trigger and regression tests across precision variants and verify release thresholds.
Common failure modes	Quantised artifacts inherit approval automatically from the source model without re-testing.
Related risks	Data and records; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D2-CTL-01 — Direct Prompt Injection Prevention

Field	Government-sector treatment
Domain	D2: RUNTIME SECURITY & ADVERSARIAL DEFENSE
Authoritative title	DIRECT PROMPT INJECTION PREVENTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to direct prompt-injection resistance for user-controlled inputs. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Separate instructions from user data, constrain tool access, apply input/output controls, test known attack patterns and log attempted instruction override.
Minimum evidence	prompt architecture; attack tests; tool policy; alert logs; remediation records
Enhanced evidence	prompt architecture; attack tests; tool policy; alert logs; remediation records; independent test evidence; affected-person or service-impact analysis where material
Assessment	Run representative direct-injection attacks and verify privileged instructions, data and tools remain protected.
Common failure modes	Reliance on a single system prompt or keyword blocklist with no tool-boundary testing.
Related risks	Security; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D2-CTL-02 — Indirect Prompt Injection Prevention

Field	Government-sector treatment
Domain	D2: RUNTIME SECURITY & ADVERSARIAL DEFENSE
Authoritative title	INDIRECT PROMPT INJECTION PREVENTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to indirect prompt-injection resistance in retrieved documents, webpages, email and files. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Label external content as untrusted, sanitise and isolate retrieval, restrict tool actions, and require confirmation for consequential operations.
Minimum evidence	content trust policy; retrieval filters; sandbox configuration; indirect-injection tests
Enhanced evidence	content trust policy; retrieval filters; sandbox configuration; indirect-injection tests; independent test evidence; affected-person or service-impact analysis where material
Assessment	Embed malicious instructions in authorised source formats and verify they cannot override policy or trigger unauthorised actions.

Field	Government-sector treatment
Common failure modes	Retrieved content is treated as trusted instruction or can silently influence tool calls.
Related risks	Security; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D2-CTL-03 — Jailbreak Resistance Testing

Field	Government-sector treatment
Domain	D2: RUNTIME SECURITY & ADVERSARIAL DEFENSE
Authoritative title	JAILBREAK RESISTANCE TESTING
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to systematic jailbreak and policy-bypass testing. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Maintain a risk-based adversarial test suite covering multi-turn, encoding, role-play and multilingual attacks; track attack success rate and remediation.
Minimum evidence	jailbreak suite; test results; defect tickets; retest evidence; accepted-risk record
Enhanced evidence	jailbreak suite; test results; defect tickets; retest evidence; accepted-risk record; independent test evidence; affected-person or service-impact analysis where material
Assessment	Re-run fixed and novel attacks against each release and verify residual bypasses are documented and bounded.
Common failure modes	One-time testing, undocumented prompts, or pass criteria based only on vendor statements.
Related risks	Security; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D2-CTL-04 — Multi-Modal Injection Defense

Field	Government-sector treatment
Domain	D2: RUNTIME SECURITY & ADVERSARIAL DEFENSE
Authoritative title	MULTI-MODAL INJECTION DEFENSE
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.

Field	Government-sector treatment
Government interpretation	Government entities should apply this control to injection defence across text, image, audio, video and document channels. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Threat-model each modality and cross-modal interaction; validate parsers, metadata and OCR/transcription outputs before they reach privileged prompts or tools.
Minimum evidence	modality threat model; parser tests; sample corpus; cross-modal attack results
Enhanced evidence	modality threat model; parser tests; sample corpus; cross-modal attack results; independent test evidence; affected-person or service-impact analysis where material
Assessment	Insert hidden or conflicting instructions in supported media and verify isolation and safe handling.
Common failure modes	Security testing covers text only while images, documents or audio can carry instructions.
Related risks	Security; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D2-CTL-05 — Function Call/Tool Call Injection Prevention

Field	Government-sector treatment
Domain	D2: RUNTIME SECURITY & ADVERSARIAL DEFENSE
Authoritative title	FUNCTION CALL/TOOL CALL INJECTION PREVENTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to prevention of malicious or unauthorised function and tool calls. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Use explicit allowlists, typed schemas, least privilege, parameter validation, transaction limits and human approval for consequential actions.
Minimum evidence	tool catalogue; permission matrix; schema validation; transaction logs; approval evidence
Enhanced evidence	tool catalogue; permission matrix; schema validation; transaction logs; approval evidence; independent test evidence; affected-person or service-impact analysis where material
Assessment	Attempt unauthorised tools, malformed parameters and privilege escalation; confirm block, alert and audit trail.
Common failure modes	Tools inherit broad service-account permissions or model-generated parameters are executed without validation.
Related risks	Security; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D2-CTL-06 — Cross-Context Hijacking Mitigation

Field	Government-sector treatment
Domain	D2: RUNTIME SECURITY & ADVERSARIAL DEFENSE
Authoritative title	CROSS-CONTEXT HIJACKING MITIGATION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to prevention of session, tenant or task context hijacking. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Isolate sessions and tenants, bind context to authenticated principals, minimise persistent state and detect cross-context references or memory leakage.
Minimum evidence	session design; tenant tests; memory policy; access logs; isolation test results
Enhanced evidence	session design; tenant tests; memory policy; access logs; isolation test results; independent test evidence; affected-person or service-impact analysis where material
Assessment	Attempt cross-user context retrieval and session fixation; verify isolation and incident detection.
Common failure modes	Shared conversation state, weak tenant keys, or cached context reused across unrelated cases.
Related risks	Security; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D3-CTL-01 — Least Agency Enforcement

Field	Government-sector treatment
Domain	D3: AGENTIC RISK & AUTONOMOUS SYSTEM SECURITY
Authoritative title	LEAST AGENCY ENFORCEMENT
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to least-agency enforcement for autonomous planning and action. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define allowed goals, actions, resources, time and spend; require explicit approval for authority expansion and enforce stop conditions.
Minimum evidence	agent authority statement; action allowlist; budget limits; intervention logs
Enhanced evidence	agent authority statement; action allowlist; budget limits; intervention logs; independent test evidence; affected-person or service-impact analysis where material
Assessment	Test out-of-scope goals and actions and confirm the agent stops or seeks authorised approval.

Field	Government-sector treatment
Common failure modes	Agent capability is assumed to equal authorised agency.
Related risks	Authority and legality; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D3-CTL-02 — Inter-Agent Communication Security

Field	Government-sector treatment
Domain	D3: AGENTIC RISK & AUTONOMOUS SYSTEM SECURITY
Authoritative title	INTER-AGENT COMMUNICATION SECURITY
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to secure and authenticated inter-agent communication. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Authenticate agents and messages, validate schemas, apply trust boundaries, prevent confused-deputy behaviour and log delegation chains.
Minimum evidence	agent registry; message signatures; schema rules; delegation logs; trust-boundary tests
Enhanced evidence	agent registry; message signatures; schema rules; delegation logs; trust-boundary tests; independent test evidence; affected-person or service-impact analysis where material
Assessment	Inject spoofed, replayed and malformed agent messages and verify rejection and traceability.
Common failure modes	Agents trust names or natural-language assertions without cryptographic or policy verification.
Related risks	Authority and legality; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D3-CTL-03 — Agentic Prompt Chaining Detection

Field	Government-sector treatment
Domain	D3: AGENTIC RISK & AUTONOMOUS SYSTEM SECURITY
Authoritative title	AGENTIC PROMPT CHAINING DETECTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.

Field	Government-sector treatment
Government interpretation	Government entities should apply this control to detection of unsafe prompt chains and emergent multi-step action paths. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Inspect plans and intermediate steps, constrain recursion and tool sequences, and detect policy-violating chains before execution.
Minimum evidence	plan logs; chain policy; recursion limits; prohibited-sequence tests
Enhanced evidence	plan logs; chain policy; recursion limits; prohibited-sequence tests; independent test evidence; affected-person or service-impact analysis where material
Assessment	Construct benign-looking steps that combine into a prohibited outcome and verify detection before execution.
Common failure modes	Controls inspect individual calls but not the cumulative effect of a chain.
Related risks	Authority and legality; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D3-CTL-04 — Embodied AI Safety Controls

Field	Government-sector treatment
Domain	D3: AGENTIC RISK & AUTONOMOUS SYSTEM SECURITY
Authoritative title	EMBODIED AI SAFETY CONTROLS
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to safety controls for agents that influence physical or operational environments. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define operational envelopes, simulation and staged testing, human override, safe state and continuity controls before real-world action.
Minimum evidence	operational envelope; simulation results; override tests; safe-state evidence
Enhanced evidence	operational envelope; simulation results; override tests; safe-state evidence; independent test evidence; affected-person or service-impact analysis where material
Assessment	Test boundary violations, sensor faults and communication loss under controlled conditions.
Common failure modes	Digital approval is treated as sufficient for physical deployment without hazard analysis.
Related risks	Authority and legality; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D3-CTL-05 — Multi-Agent Trust Chain Attestation

Field	Government-sector treatment
Domain	D3: AGENTIC RISK & AUTONOMOUS SYSTEM SECURITY
Authoritative title	MULTI-AGENT TRUST CHAIN ATTESTATION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to attestation of identity, capability and trust across multi-agent chains. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Require verifiable agent identity, approved capability claims, provenance of delegated tasks and expiry/revocation of trust.
Minimum evidence	attestation records; capability registry; delegation chain; revocation tests
Enhanced evidence	attestation records; capability registry; delegation chain; revocation tests; independent test evidence; affected-person or service-impact analysis where material
Assessment	Introduce an untrusted or expired agent into a chain and confirm refusal and alerting.
Common failure modes	Trust propagates transitively without verification or revocation.
Related risks	Authority and legality; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D3-CTL-06 — Persistent Memory Exfiltration Prevention

Field	Government-sector treatment
Domain	D3: AGENTIC RISK & AUTONOMOUS SYSTEM SECURITY
Authoritative title	PERSISTENT MEMORY EXFILTRATION PREVENTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to prevention of sensitive-data exfiltration through persistent agent memory. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Classify memory content, minimise retention, isolate tenants, apply DLP and access controls, and test extraction through prompts and tools.
Minimum evidence	memory inventory; retention rules; DLP logs; access tests; deletion evidence
Enhanced evidence	memory inventory; retention rules; DLP logs; access tests; deletion evidence; independent test evidence; affected-person or service-impact analysis where material
Assessment	Attempt to retrieve another case's or user's memory and verify prevention and detection.

Field	Government-sector treatment
Common failure modes	Persistent memory stores secrets or citizen data without classification, expiry or tenant isolation.
Related risks	Authority and legality; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D3-CTL-07 — Secure Memory Lifecycle Management

Field	Government-sector treatment
Domain	D3: AGENTIC RISK & AUTONOMOUS SYSTEM SECURITY
Authoritative title	SECURE MEMORY LIFECYCLE MANAGEMENT
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to secure creation, use, review, correction and deletion of agent memory. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Establish memory lifecycle states, provenance, correction rights, retention schedules, deletion verification and backup handling.
Minimum evidence	memory lifecycle procedure; provenance logs; correction records; deletion tests
Enhanced evidence	memory lifecycle procedure; provenance logs; correction records; deletion tests; independent test evidence; affected-person or service-impact analysis where material
Assessment	Create, amend and delete test memories and verify changes propagate to indexes, caches and backups as required.
Common failure modes	Deleted or corrected memory remains available to the agent through secondary stores.
Related risks	Authority and legality; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D4-CTL-01 — AI Bill Of Materials (AI BOM) Maintenance

Field	Government-sector treatment
Domain	D4: SUPPLY CHAIN & THIRD-PARTY AI SECURITY
Authoritative title	AI BILL OF MATERIALS (AI BOM) MAINTENANCE
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.

Field	Government-sector treatment
Government interpretation	Government entities should apply this control to complete AI bill of materials for models, data, code, services and tools. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Maintain versioned inventories of models, datasets, adapters, libraries, APIs, tools, licences, owners and deployment locations.
Minimum evidence	AI BOM; dependency versions; owners; licence records; change history
Enhanced evidence	AI BOM; dependency versions; owners; licence records; change history; independent test evidence; affected-person or service-impact analysis where material
Assessment	Sample deployed systems and reconcile observed dependencies to the AI BOM.
Common failure modes	Inventory stops at the primary model and omits retrieval stores, adapters, plugins or hosted dependencies.
Related risks	Procurement and vendor; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D4-CTL-02 — Model File & Artifact Scanning

Field	Government-sector treatment
Domain	D4: SUPPLY CHAIN & THIRD-PARTY AI SECURITY
Authoritative title	MODEL FILE & ARTIFACT SCANNING
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to malware, secrets, unsafe deserialisation and integrity scanning of model artifacts. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Scan files in quarantine, verify hashes/signatures, block unsafe formats and retain scan results before promotion.
Minimum evidence	quarantine logs; scanner outputs; hashes; format policy; release approval
Enhanced evidence	quarantine logs; scanner outputs; hashes; format policy; release approval; independent test evidence; affected-person or service-impact analysis where material
Assessment	Submit altered, malicious and unsafe-serialisation artifacts and verify quarantine and block.
Common failure modes	Artifacts are downloaded directly into production or scans exclude large model files.
Related risks	Procurement and vendor; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D4-CTL-03 — Model Hub & Registry Vetting

Field	Government-sector treatment
Domain	D4: SUPPLY CHAIN & THIRD-PARTY AI SECURITY
Authoritative title	MODEL HUB & REGISTRY VETTING
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to governance of model hubs, registries and provenance sources. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Approve registries and publishers, verify identity and licence, pin versions, monitor takedowns and preserve acquisition evidence.
Minimum evidence	approved registry list; publisher verification; acquisition record; licence review; takedown monitoring
Enhanced evidence	approved registry list; publisher verification; acquisition record; licence review; takedown monitoring; independent test evidence; affected-person or service-impact analysis where material
Assessment	Attempt acquisition from an unapproved or impersonated source and verify prevention.
Common failure modes	Popularity or download count is treated as assurance.
Related risks	Procurement and vendor; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D4-CTL-04 — MCP Server Behavioral Monitoring

Field	Government-sector treatment
Domain	D4: SUPPLY CHAIN & THIRD-PARTY AI SECURITY
Authoritative title	MCP SERVER BEHAVIORAL MONITORING
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to behavioural monitoring of Model Context Protocol and comparable tool servers. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Inventory servers and exposed tools, authenticate connections, constrain permissions, monitor changes and anomalous calls, and support rapid disablement.
Minimum evidence	server inventory; tool schemas; permissions; monitoring alerts; disable test
Enhanced evidence	server inventory; tool schemas; permissions; monitoring alerts; disable test; independent test evidence; affected-person or service-impact analysis where material

Field	Government-sector treatment
Assessment	Change a server tool or return malicious content and verify detection, containment and revocation.
Common failure modes	MCP servers are treated as passive data sources rather than privileged execution dependencies.
Related risks	Procurement and vendor; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D4-CTL-05 — Third-Party AI Api Security Assessment

Field	Government-sector treatment
Domain	D4: SUPPLY CHAIN & THIRD-PARTY AI SECURITY
Authoritative title	THIRD-PARTY AI API SECURITY ASSESSMENT
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to security assessment of third-party AI APIs. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Assess data handling, authentication, rate limits, logging, residency, model changes, incident notification and exit arrangements before use.
Minimum evidence	supplier assessment; data-flow map; contract clauses; API test results; exit plan
Enhanced evidence	supplier assessment; data-flow map; contract clauses; API test results; exit plan; independent test evidence; affected-person or service-impact analysis where material
Assessment	Verify supplier controls and simulate service/model change and incident notification workflows.
Common failure modes	API use begins under standard SaaS terms with no AI-specific evidence or change rights.
Related risks	Procurement and vendor; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D4-CTL-06 — Shadow AI Discovery & Governance

Field	Government-sector treatment
Domain	D4: SUPPLY CHAIN & THIRD-PARTY AI SECURITY
Authoritative title	SHADOW AI DISCOVERY & GOVERNANCE
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.

Field	Government-sector treatment
Government interpretation	Government entities should apply this control to discovery and governance of unapproved or shadow AI use. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Use network, identity, expense, browser and endpoint signals consistent with law and policy; provide approved alternatives and remediation paths.
Minimum evidence	discovery method; approved-service register; findings; remediation and exception records
Enhanced evidence	discovery method; approved-service register; findings; remediation and exception records; independent test evidence; affected-person or service-impact analysis where material
Assessment	Seed a controlled unapproved service and confirm detection, triage and proportionate response.
Common failure modes	Discovery relies only on staff self-reporting or suppresses use without providing safe alternatives.
Related risks	Procurement and vendor; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D4-CTL-07 — AI Software Composition Analysis (SCA)

Field	Government-sector treatment
Domain	D4: SUPPLY CHAIN & THIRD-PARTY AI SECURITY
Authoritative title	AI SOFTWARE COMPOSITION ANALYSIS (SCA)
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to software composition analysis for AI applications and supporting stacks. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Scan code, containers and dependencies; correlate vulnerabilities with reachability and AI-specific exposure; track remediation and exceptions.
Minimum evidence	SBOM/SCA results; reachability analysis; remediation tickets; exception approvals
Enhanced evidence	SBOM/SCA results; reachability analysis; remediation tickets; exception approvals; independent test evidence; affected-person or service-impact analysis where material
Assessment	Introduce a vulnerable reachable dependency and verify detection, prioritisation and closure.
Common failure modes	SCA excludes notebooks, model-serving images, GPU libraries or generated code.
Related risks	Procurement and vendor; Security; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D5-CTL-01 — Harmful Content Blocking

Field	Government-sector treatment
Domain	D5: CONTENT SAFETY & OUTPUT INTEGRITY
Authoritative title	HARMFUL CONTENT BLOCKING
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to blocking or safe handling of harmful content within lawful government purpose. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define prohibited and restricted content by use case, test over- and under-blocking, support escalation and preserve lawful access and records duties.
Minimum evidence	content policy; evaluation set; false-positive/negative analysis; escalation logs
Enhanced evidence	content policy; evaluation set; false-positive/negative analysis; escalation logs; independent test evidence; affected-person or service-impact analysis where material
Assessment	Test harmful and legitimate edge cases, including protected speech and statutory-service contexts.
Common failure modes	A generic vendor safety filter is used without government-purpose calibration or appeal.
Related risks	Rights and public impact; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D5-CTL-02 — PII Leakage Prevention

Field	Government-sector treatment
Domain	D5: CONTENT SAFETY & OUTPUT INTEGRITY
Authoritative title	PII LEAKAGE PREVENTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to prevention of personal and protected information leakage. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Classify data, minimise prompts and context, apply DLP and output checks, isolate tenants and test memorisation or retrieval leakage.
Minimum evidence	data-flow map; DLP rules; leakage tests; access logs; incident records
Enhanced evidence	data-flow map; DLP rules; leakage tests; access logs; incident records; independent test evidence; affected-person or service-impact analysis where material
Assessment	Use canary identifiers and cross-tenant tests to verify prevention and alerting.

Field	Government-sector treatment
Common failure modes	Sensitive data is placed in prompts because the supplier claims not to train on it.
Related risks	Rights and public impact; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D5-CTL-03 — Copyright Detection

Field	Government-sector treatment
Domain	D5: CONTENT SAFETY & OUTPUT INTEGRITY
Authoritative title	COPYRIGHT DETECTION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to copyright and rights-risk detection for inputs and outputs. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Record source and licence constraints, screen outputs where appropriate, provide review for publication and retain attribution or permission evidence.
Minimum evidence	rights register; licence records; output review; attribution/permission evidence
Enhanced evidence	rights register; licence records; output review; attribution/permission evidence; independent test evidence; affected-person or service-impact analysis where material
Assessment	Test representative copyrighted inputs and publication workflows; verify escalation for uncertain rights.
Common failure modes	Automated similarity detection is treated as a legal conclusion.
Related risks	Rights and public impact; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D5-CTL-04 — AI Watermarking Robustness

Field	Government-sector treatment
Domain	D5: CONTENT SAFETY & OUTPUT INTEGRITY
Authoritative title	AI WATERMARKING ROBUSTNESS
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to robust provenance and disclosure for AI-generated government content. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.

Field	Government-sector treatment
Implementation guidance	Apply durable provenance metadata or disclosure appropriate to channel, test transformations and define exceptions for security or accessibility.
Minimum evidence	content provenance records; disclosure policy; transformation tests; exception approvals
Enhanced evidence	content provenance records; disclosure policy; transformation tests; exception approvals; independent test evidence; affected-person or service-impact analysis where material
Assessment	Resize, transcode and repost marked content and verify provenance/disclosure remains usable.
Common failure modes	Watermarking is claimed to prove authenticity without key, custody and verification controls.
Related risks	Rights and public impact; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D5-CTL-05 — Privacy-By-Design Verification

Field	Government-sector treatment
Domain	D5: CONTENT SAFETY & OUTPUT INTEGRITY
Authoritative title	PRIVACY-BY-DESIGN VERIFICATION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to privacy-by-design throughout purpose, data, architecture and lifecycle decisions. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Document purpose and lawful basis, minimise data, assess privacy impacts, implement rights and retention controls, and review changes.
Minimum evidence	privacy impact assessment; minimisation record; retention schedule; rights workflow
Enhanced evidence	privacy impact assessment; minimisation record; retention schedule; rights workflow; independent test evidence; affected-person or service-impact analysis where material
Assessment	Trace a data subject request and a purpose change through the system and verify controls.
Common failure modes	Privacy is reviewed only at procurement or after deployment.
Related risks	Rights and public impact; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D5-CTL-06 — Privacy-Preserving ML Validation

Field	Government-sector treatment
Domain	D5: CONTENT SAFETY & OUTPUT INTEGRITY
Authoritative title	PRIVACY-PRESERVING ML VALIDATION
Applicability	Applicable with Government Interpretation
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to validation of privacy-preserving machine-learning techniques and their limits. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define the privacy threat model, parameters and utility trade-offs; test attacks relevant to the chosen technique and monitor configuration drift.
Minimum evidence	threat model; parameter record; attack tests; utility analysis; approval
Enhanced evidence	threat model; parameter record; attack tests; utility analysis; approval; independent test evidence; affected-person or service-impact analysis where material
Assessment	Run membership, reconstruction or linkage tests as applicable and verify claimed protection.
Common failure modes	Use of differential privacy, federated learning or synthetic data is asserted without measured guarantees.
Related risks	Rights and public impact; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D6-CTL-01 — Human-In-The-Loop For High-Risk Actions

Field	Government-sector treatment
Domain	D6: GOVERNANCE, ACCOUNTABILITY & HUMAN OVERSIGHT
Authoritative title	HUMAN-IN-THE-LOOP FOR HIGH-RISK ACTIONS
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to effective human control over high-impact or irreversible actions. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Identify decisions requiring human authority, ensure reviewers have competence, time, evidence and power to change outcomes, and test override effectiveness.
Minimum evidence	decision-rights matrix; reviewer training; review logs; override tests; quality metrics
Enhanced evidence	decision-rights matrix; reviewer training; review logs; override tests; quality metrics; independent test evidence; affected-person or service-impact analysis where material
Assessment	Sample decisions and confirm meaningful independent review, reasons and correction capability.

Field	Government-sector treatment
Common failure modes	A click-through approval is counted as human oversight.
Related risks	Authority and legality; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D6-CTL-02 — Audit Trail Completeness

Field	Government-sector treatment
Domain	D6: GOVERNANCE, ACCOUNTABILITY & HUMAN OVERSIGHT
Authoritative title	AUDIT TRAIL COMPLETENESS
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to complete and reconstructable audit trails. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Log inputs, outputs, model/version, retrieval context, tools, approvals, changes and outcomes with integrity, access and retention controls.
Minimum evidence	logging standard; sample logs; integrity controls; retention evidence; reconstruction test
Enhanced evidence	logging standard; sample logs; integrity controls; retention evidence; reconstruction test; independent test evidence; affected-person or service-impact analysis where material
Assessment	Reconstruct a material decision end-to-end from retained records.
Common failure modes	Logs omit prompts, retrieval sources, tool actions or human edits needed to explain the outcome.
Related risks	Authority and legality; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D6-CTL-03 — AI Model Card Completeness

Field	Government-sector treatment
Domain	D6: GOVERNANCE, ACCOUNTABILITY & HUMAN OVERSIGHT
Authoritative title	AI MODEL CARD COMPLETENESS
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to complete, current and decision-useful model and system cards. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.

Field	Government-sector treatment
Implementation guidance	Document purpose, limitations, data, performance, risks, authority level, dependencies, monitoring and change history; update after material change.
Minimum evidence	approved system card; review history; linked evaluations; change records
Enhanced evidence	approved system card; review history; linked evaluations; change records; independent test evidence; affected-person or service-impact analysis where material
Assessment	Compare the card with the deployed system and verify accuracy and currency.
Common failure modes	Model cards repeat supplier marketing and omit local use, limitations or modifications.
Related risks	Authority and legality; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D6-CTL-04 — AI Incident Response Readiness

Field	Government-sector treatment
Domain	D6: GOVERNANCE, ACCOUNTABILITY & HUMAN OVERSIGHT
Authoritative title	AI INCIDENT RESPONSE READINESS
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to AI-specific incident readiness integrated with government escalation. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define triggers, roles, evidence preservation, citizen-impact response, supplier coordination, notification and lessons learned.
Minimum evidence	incident plan; exercises; contact list; evidence procedure; corrective-action log
Enhanced evidence	incident plan; exercises; contact list; evidence procedure; corrective-action log; independent test evidence; affected-person or service-impact analysis where material
Assessment	Run a tabletop covering harmful output, data exposure and unauthorised action.
Common failure modes	AI incidents are forced into generic cyber playbooks with no affected-person or model-change response.
Related risks	Authority and legality; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D6-CTL-05 — Model Deprecation & Decommissioning

Field	Government-sector treatment
Domain	D6: GOVERNANCE, ACCOUNTABILITY & HUMAN OVERSIGHT

Field	Government-sector treatment
Authoritative title	MODEL DEPRECATION & DECOMMISSIONING
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to controlled model deprecation and system decommissioning. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Plan replacement, notice, records preservation, data and credential removal, dependency closure, citizen continuity and post-retirement monitoring.
Minimum evidence	decommission plan; migration record; deletion evidence; archive record; dependency closure
Enhanced evidence	decommission plan; migration record; deletion evidence; archive record; dependency closure; independent test evidence; affected-person or service-impact analysis where material
Assessment	Retire a test component and verify no residual access, calls or unsupported dependencies remain.
Common failure modes	A model is disabled but data, keys, endpoints and downstream references remain active.
Related risks	Authority and legality; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D6-CTL-06 — Third-Party AI Vendor Governance

Field	Government-sector treatment
Domain	D6: GOVERNANCE, ACCOUNTABILITY & HUMAN OVERSIGHT
Authoritative title	THIRD-PARTY AI VENDOR GOVERNANCE
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to ongoing governance of third-party AI vendors. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Assign accountable owners, monitor evidence and changes, enforce audit/incident rights, manage subcontractors and maintain exit capability.
Minimum evidence	vendor register; review cadence; contract evidence; change notices; exit test
Enhanced evidence	vendor register; review cadence; contract evidence; change notices; exit test; independent test evidence; affected-person or service-impact analysis where material
Assessment	Review a material supplier change and verify risk reassessment and approval before use.
Common failure modes	Initial due diligence is treated as permanent assurance.
Related risks	Authority and legality; Democratic and institutional; Operational

Field	Government-sector treatment
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D6-CTL-07 — AI Resilience & Business Continuity

Field	Government-sector treatment
Domain	D6: GOVERNANCE, ACCOUNTABILITY & HUMAN OVERSIGHT
Authoritative title	AI RESILIENCE & BUSINESS CONTINUITY
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to resilience and continuity of AI-enabled public services. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define service tolerances, manual alternatives, degraded modes, backup dependencies, recovery objectives and exercise results.
Minimum evidence	continuity plan; manual procedure; dependency map; recovery test; lessons learned
Enhanced evidence	continuity plan; manual procedure; dependency map; recovery test; lessons learned; independent test evidence; affected-person or service-impact analysis where material
Assessment	Simulate model/API outage and data-quality degradation and verify essential service continuity.
Common failure modes	Continuity assumes the same unavailable AI supplier or omits staff capacity for manual fallback.
Related risks	Authority and legality; Democratic and institutional; Operational
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D7-CTL-H01 — Ai-Generated Phishing Simulation

Field	Government-sector treatment
Domain	D7: HUMAN & SOCIETAL HARMS
Authoritative title	AI-GENERATED PHISHING SIMULATION
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to safe simulation of AI-generated phishing for workforce preparedness. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Authorise scope, protect participants, avoid real credential capture, measure learning and prevent simulation content from escaping.

Field	Government-sector treatment
Minimum evidence	exercise approval; scenario design; participant safeguards; metrics; deletion evidence
Enhanced evidence	exercise approval; scenario design; participant safeguards; metrics; deletion evidence; independent test evidence; affected-person or service-impact analysis where material
Assessment	Confirm simulations cannot collect live secrets or send outside approved recipients.
Common failure modes	Realistic simulation is prioritised over proportionality, consent/policy and data protection.
Related risks	Rights and public impact; Democratic and institutional
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D7-CTL-H02 — Deepfake Detection Training

Field	Government-sector treatment
Domain	D7: HUMAN & SOCIETAL HARMS
Authoritative title	DEEPPFAKE DETECTION TRAINING
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to workforce capability to detect and escalate deepfakes. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Train staff using role-relevant media, verification channels and escalation procedures; test performance and refresh for emerging techniques.
Minimum evidence	training materials; attendance; exercise results; escalation logs
Enhanced evidence	training materials; attendance; exercise results; escalation logs; independent test evidence; affected-person or service-impact analysis where material
Assessment	Run controlled deepfake recognition and verification exercises for high-risk roles.
Common failure modes	Training focuses on visual artefacts only and omits process verification.
Related risks	Rights and public impact; Democratic and institutional
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D7-CTL-H03 — Out-Of-Band Authentication

Field	Government-sector treatment
Domain	D7: HUMAN & SOCIETAL HARMS
Authoritative title	OUT-OF-BAND AUTHENTICATION
Applicability	Applicable with Enhanced Evidence

Field	Government-sector treatment
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to independent authentication for high-risk requests and communications. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Require out-of-band verification through a separately controlled channel for payments, credentials, sensitive disclosure and authority changes.
Minimum evidence	authentication procedure; protected contact registry; verification logs; exception records
Enhanced evidence	authentication procedure; protected contact registry; verification logs; exception records; independent test evidence; affected-person or service-impact analysis where material
Assessment	Simulate spoofed executive or citizen requests and verify independent confirmation.
Common failure modes	The second channel relies on contact details supplied in the suspicious message.
Related risks	Rights and public impact; Democratic and institutional
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D7-CTL-H04 — AI Social Engineering IR

Field	Government-sector treatment
Domain	D7: HUMAN & SOCIETAL HARMS
Authoritative title	AI SOCIAL ENGINEERING IR
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to incident response for AI-enabled social engineering. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Include deepfake, synthetic identity, automated targeting and impersonation in triage, containment, evidence and communications plans.
Minimum evidence	playbook; exercises; evidence checklist; coordination records
Enhanced evidence	playbook; exercises; evidence checklist; coordination records; independent test evidence; affected-person or service-impact analysis where material
Assessment	Run a scenario combining synthetic media and credential abuse and verify coordinated response.
Common failure modes	Cases are classified only as user error and synthetic evidence is not preserved.
Related risks	Rights and public impact; Democratic and institutional
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D7-CTL-H05 — Ai-Enhanced External Attack Defense

Field	Government-sector treatment
Domain	D7: HUMAN & SOCIETAL HARMS
Authoritative title	AI-ENHANCED EXTERNAL ATTACK DEFENSE
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to defence against AI-enhanced external attacks. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Use threat intelligence to adapt detection, rate limits, identity controls and response for automated reconnaissance, exploitation and evasion.
Minimum evidence	threat model; detection updates; exercise results; incident metrics
Enhanced evidence	threat model; detection updates; exercise results; incident metrics; independent test evidence; affected-person or service-impact analysis where material
Assessment	Exercise high-volume adaptive attacks and verify detection and service protection.
Common failure modes	Claims of 'AI-powered attack' are accepted without evidence, or controls focus on attribution rather than observable behaviour.
Related risks	Rights and public impact; Democratic and institutional
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D8-CTL-01 — EU AI Act Risk Tier Mapping

Field	Government-sector treatment
Domain	D8: REGULATORY ALIGNMENT & COMPLIANCE
Authoritative title	EU AI ACT RISK TIER MAPPING
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to documented mapping to the EU AI Act where the organisation and use are legally in scope. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Identify actor role, system classification, applicable dates and obligations using current official text; obtain qualified legal review and avoid globalising EU requirements.
Minimum evidence	jurisdiction map; legal applicability record; system classification; review date
Enhanced evidence	jurisdiction map; legal applicability record; system classification; review date; independent test evidence; affected-person or service-impact analysis where material
Assessment	Sample systems and verify role, risk classification and effective-date assumptions against official sources.

Field	Government-sector treatment
Common failure modes	The EU AI Act is treated as universally applicable or a framework mapping is presented as legal compliance.
Related risks	Authority and legality; Legal and policy dependency
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D8-CTL-02 — ISO 42001 Gap Analysis

Field	Government-sector treatment
Domain	D8: REGULATORY ALIGNMENT & COMPLIANCE
Authoritative title	ISO 42001 GAP ANALYSIS
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to gap analysis against ISO/IEC 42001 without equating mapping to certification. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define scope, compare requirements to implemented practices, record objective evidence, gaps and treatment, and use licensed standard text appropriately.
Minimum evidence	scope statement; gap matrix; evidence references; remediation plan
Enhanced evidence	scope statement; gap matrix; evidence references; remediation plan; independent test evidence; affected-person or service-impact analysis where material
Assessment	Verify sampled 'met' ratings against objective evidence and current standard edition.
Common failure modes	Self-assessment is labelled certification or controls are inferred from secondary summaries.
Related risks	Authority and legality; Legal and policy dependency
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D8-CTL-03 — GPAI Technical Documentation Verification

Field	Government-sector treatment
Domain	D8: REGULATORY ALIGNMENT & COMPLIANCE
Authoritative title	GPAI TECHNICAL DOCUMENTATION VERIFICATION
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.

Field	Government-sector treatment
Government interpretation	Government entities should apply this control to verification of technical documentation obligations for general-purpose AI where applicable. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Determine provider/deployer role, preserve model and system documentation, limitations, evaluation and downstream information, and obtain legal review for applicability.
Minimum evidence	role assessment; technical documentation; evaluation summary; downstream notices; update history
Enhanced evidence	role assessment; technical documentation; evaluation summary; downstream notices; update history; independent test evidence; affected-person or service-impact analysis where material
Assessment	Trace required documentation through a material model update and downstream deployment.
Common failure modes	Generic model cards are assumed to satisfy all GPAI legal obligations.
Related risks	Authority and legality; Legal and policy dependency
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D8-CTL-04 — DORA ICT Incident Reporting (Financial Sector)

Field	Government-sector treatment
Domain	D8: REGULATORY ALIGNMENT & COMPLIANCE
Authoritative title	DORA ICT INCIDENT REPORTING (FINANCIAL SECTOR)
Applicability	Context Dependent
Rationale	DORA applies to EU financial entities and their ICT providers. Apply only where the public body is, or supports, a DORA-scoped entity.
Government interpretation	Government entities should apply this control to incident-reporting readiness for DORA only where financial-sector scope applies, while applying the underlying reporting discipline elsewhere. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Document whether DORA applies; if not, map incident classification, timelines, evidence and notification to the governing public-sector regime.
Minimum evidence	scope decision; reporting matrix; incident classification; notification records
Enhanced evidence	scope decision; reporting matrix; incident classification; notification records; independent test evidence; affected-person or service-impact analysis where material
Assessment	Test classification and timed escalation under the applicable regime; verify DORA references are not used outside scope.
Common failure modes	The authoritative title is read as imposing DORA on every government system.
Related risks	Authority and legality; Legal and policy dependency
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D8-CTL-05 — NIST SP 800-218A Compliance Check

Field	Government-sector treatment
Domain	D8: REGULATORY ALIGNMENT & COMPLIANCE
Authoritative title	NIST SP 800-218A COMPLIANCE CHECK
Applicability	Applicable with Enhanced Evidence
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to secure AI software development alignment with NIST SP 800-218A or another authorised regime. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Map AI development practices to the selected secure-development framework, retain evidence and document deviations; verify current edition and applicability.
Minimum evidence	framework mapping; secure-development records; test evidence; deviation approvals
Enhanced evidence	framework mapping; secure-development records; test evidence; deviation approvals; independent test evidence; affected-person or service-impact analysis where material
Assessment	Sample releases and trace requirements through design, build, test, deployment and response.
Common failure modes	A checklist is completed without evidence or is applied to acquired services as though the agency developed them.
Related risks	Authority and legality; Legal and policy dependency
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D9-CTL-01 — Physical Harm Boundary Enforcement

Field	Government-sector treatment
Domain	D9: PHYSICAL AI SAFETY
Authoritative title	PHYSICAL HARM BOUNDARY ENFORCEMENT
Applicability	Context Dependent
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to enforcement of physical safety and harm boundaries. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define hazards, operating envelope, prohibited states and independent safety constraints; validate before live operation.
Minimum evidence	hazard analysis; boundary specification; independent interlocks; test results
Enhanced evidence	hazard analysis; boundary specification; independent interlocks; test results; independent test evidence; affected-person or service-impact analysis where material
Assessment	Attempt boundary violations and sensor anomalies and verify safe prevention.

Field	Government-sector treatment
Common failure modes	Safety limits exist only in the AI policy layer that can fail with the model.
Related risks	Physical safety; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D9-CTL-02 — Safe State And Graceful Degradation

Field	Government-sector treatment
Domain	D9: PHYSICAL AI SAFETY
Authoritative title	SAFE STATE AND GRACEFUL DEGRADATION
Applicability	Context Dependent
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to transition to a safe state and graceful degradation. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Define safe states for loss of model, data, communications or power; preserve essential service where possible and test degraded modes.
Minimum evidence	safe-state design; degraded-mode procedure; failure tests; recovery records
Enhanced evidence	safe-state design; degraded-mode procedure; failure tests; recovery records; independent test evidence; affected-person or service-impact analysis where material
Assessment	Induce dependency failures and verify timely safe transition and controlled recovery.
Common failure modes	Fail-safe behaviour creates a different public-safety or service-continuity hazard.
Related risks	Physical safety; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D9-CTL-03 — Human Override And Emergency Stop

Field	Government-sector treatment
Domain	D9: PHYSICAL AI SAFETY
Authoritative title	HUMAN OVERRIDE AND EMERGENCY STOP
Applicability	Context Dependent
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to effective human override and emergency stop. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.

Field	Government-sector treatment
Implementation guidance	Provide accessible, independent and tested override mechanisms with clear authority, training and post-event logging.
Minimum evidence	override design; authority matrix; training; test logs; maintenance records
Enhanced evidence	override design; authority matrix; training; test logs; maintenance records; independent test evidence; affected-person or service-impact analysis where material
Assessment	Test override under realistic load, communications loss and partial system failure.
Common failure modes	Emergency stop is inaccessible, shares the failed control path or is not maintained.
Related risks	Physical safety; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D9-CTL-04 — Cyber-Physical Attack Detection

Field	Government-sector treatment
Domain	D9: PHYSICAL AI SAFETY
Authoritative title	CYBER-PHYSICAL ATTACK DETECTION
Applicability	Context Dependent
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to detection and response to cyber-physical attacks. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Correlate cyber, sensor, actuator and process signals; define containment that protects people and services; preserve evidence.
Minimum evidence	threat model; correlated monitoring; response playbook; exercise evidence
Enhanced evidence	threat model; correlated monitoring; response playbook; exercise evidence; independent test evidence; affected-person or service-impact analysis where material
Assessment	Simulate spoofed sensors and malicious commands and verify detection and safe containment.
Common failure modes	Cyber and operational monitoring remain siloed and neither sees the complete attack.
Related risks	Physical safety; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D9-CTL-05 — Physical Environment Integrity Monitoring

Field	Government-sector treatment
Domain	D9: PHYSICAL AI SAFETY

Field	Government-sector treatment
Authoritative title	PHYSICAL ENVIRONMENT INTEGRITY MONITORING
Applicability	Context Dependent
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to integrity monitoring of the physical environment and sensing context. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Validate sensor provenance, calibration, redundancy and environmental assumptions; detect tampering and implausible conditions.
Minimum evidence	sensor inventory; calibration records; redundancy tests; tamper alerts
Enhanced evidence	sensor inventory; calibration records; redundancy tests; tamper alerts; independent test evidence; affected-person or service-impact analysis where material
Assessment	Manipulate or obstruct sensors and verify plausibility checks and escalation.
Common failure modes	The AI trusts sensor readings without independent checks or maintenance evidence.
Related risks	Physical safety; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D9-CTL-06 — Actuator Command Verification

Field	Government-sector treatment
Domain	D9: PHYSICAL AI SAFETY
Authoritative title	ACTUATOR COMMAND VERIFICATION
Applicability	Context Dependent
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to verification of actuator commands before execution. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Authenticate commands, validate ranges and sequences, enforce rate and safety limits, and use independent interlocks for hazardous actions.
Minimum evidence	command schema; authentication logs; safety rules; interlock tests
Enhanced evidence	command schema; authentication logs; safety rules; interlock tests; independent test evidence; affected-person or service-impact analysis where material
Assessment	Submit replayed, malformed, out-of-sequence and unsafe commands and verify rejection.
Common failure modes	Natural-language or model-generated commands reach actuators without deterministic validation.
Related risks	Physical safety; Operational; Public-service continuity

Field	Government-sector treatment
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

D9-CTL-07 — Physical Incident Evidence Preservation

Field	Government-sector treatment
Domain	D9: PHYSICAL AI SAFETY
Authoritative title	PHYSICAL INCIDENT EVIDENCE PRESERVATION
Applicability	Context Dependent
Rationale	The control remains in the 59-control baseline. Government implementation depends on documented system scope, decision consequence, jurisdiction, data sensitivity, and operating context.
Government interpretation	Government entities should apply this control to preservation of evidence after physical-AI incidents. The accountable public body retains decision and assurance responsibility even when a supplier operates the technology.
Implementation guidance	Synchronise time, retain cyber and physical logs, protect chain of custody, capture configuration and support safety/legal investigations.
Minimum evidence	evidence plan; time-sync evidence; sealed logs; custody records; configuration snapshot
Enhanced evidence	evidence plan; time-sync evidence; sealed logs; custody records; configuration snapshot; independent test evidence; affected-person or service-impact analysis where material
Assessment	Run an incident exercise and verify complete, tamper-evident evidence can be reconstructed.
Common failure modes	Volatile model, sensor or actuator evidence is overwritten before investigators can preserve it.
Related risks	Physical safety; Operational; Public-service continuity
Legal/policy dependency	Jurisdiction-specific public law, records, privacy, procurement, accessibility, cybersecurity, sector, and agency requirements.

16. Assessment Model

Assessment dimension	Question
Design adequacy	Can the control design address the documented risk and authority model?
Implementation completeness	Are required people, process, technology, supplier, and records elements present?
Operating effectiveness	Did the control work consistently during the assessment period and scenarios?
Evidence sufficiency	Is evidence authentic, complete, attributable, timely, and reproducible?
Residual risk	What material risk remains, who accepted it, and when is review due?
Legal/policy dependency	Which conclusions require qualified jurisdiction-specific review?

17. Implementation Maturity Model

Level	Observable characteristics
1 — Initial	Inventories incomplete; controls reactive; ownership unclear; evidence inconsistent.
2 — Defined	Policies, roles, classification, procedures, and minimum evidence documented.
3 — Implemented	Controls operate across in-scope systems; suppliers and exceptions governed.
4 — Measured	Effectiveness, disparities, incidents, appeals, overrides, and metrics reviewed.
5 — Adaptive	Threats, complaints, audits, incidents, legal changes, and performance drive improvement.

Documentation volume is not maturity.

18. Phased Implementation Roadmap

Timeframe	Phase	Required outcomes
0–30 days	Establish authority and scope	Inventory systems and suppliers; assign owners; classify use cases; constrain unacceptable use.
31–60 days	Establish minimum controls	Implement access, logging, data restrictions, review rules, supplier clauses, incident reporting, and evidence retention.
61–90 days	Validate operation	Test security, oversight, notice, contestability, fallback, supplier evidence, and reconstruction.
3–6 months	Institutionalise assurance	Periodic assessment, model-change governance, independent review, reporting, remediation.
6–12 months	Improve and adapt	Use incidents, appeals, complaints, audits, threat intelligence, and performance data.

19. Worked Examples

19.1 Citizen-service chatbot

Element	Treatment
Decision-authority level	Level 2
Context and role	Answers service questions and triages cases; it cannot determine eligibility or alter an official record.
Material risks	Direct or indirect prompt injection, inaccurate advice, sensitive-data disclosure, inaccessible interaction, and false attribution of authority.

Element	Treatment
Required safeguards	Approved knowledge base; untrusted-content isolation; disclosure that the service is automated; privacy controls; accessible human channel; query and output monitoring.
Evidence	Knowledge-source approvals; injection tests; privacy assessment; accessibility test; escalation logs; sampled-answer review.
Failure scenario	A webpage retrieved by the assistant contains hidden instructions that cause the chatbot to disclose case details or provide unauthorised procedural advice.
Detection and response	Indirect-injection monitoring or a complaint identifies the issue; disable affected retrieval sources, preserve prompts and logs, notify the service owner, correct published advice, assess affected users, and retest before restoration.
Notably Absent	No claim is made that every citizen chatbot is high risk or that disclosure alone makes inaccurate advice safe.

19.2 Benefits eligibility support

Element	Treatment
Decision-authority level	Level 4
Context and role	Ranks or recommends benefit eligibility cases; an authorised official must determine entitlement and provide reasons.
Material risks	Disparate denial or delay, unlawful delegation, stale data, automation bias, inadequate notice, and ineffective appeal.
Required safeguards	Documented legal authority; cohort testing; independent review; reasons and notice; appeal route; correction workflow; manual fallback.
Evidence	Authority assessment; data provenance; impact analysis; cohort metrics; reviewer records; decision notices; appeal outcomes.
Failure scenario	A historical proxy variable suppresses eligibility recommendations for a protected community, and reviewers routinely accept the score without examining evidence.
Detection and response	Disparity monitoring or appeal data triggers investigation; suspend the affected rule/model, identify and correct decisions, notify accountable officials, validate revised features, and provide redress where required.
Notably Absent	Nominal human approval is not treated as evidence of meaningful oversight, and no universal fairness threshold is asserted.

19.3 Tax audit prioritisation

Element	Treatment
Decision-authority level	Level 4
Context and role	Prioritises taxpayers for audit; the model does not establish liability and selection must remain legally authorised and reviewable.
Material risks	Biased selection, opaque risk factors, strategic gaming, data incompatibility, excessive scrutiny, and inability to explain selection.
Required safeguards	Feature and purpose review; protected-attribute/proxy analysis; calibrated thresholds; reason codes; sampling of low-score cases; independent audit.
Evidence	Selection policy; model card; feature approvals; disparity and calibration tests; reviewer decisions; challenge outcomes.

Element	Treatment
Failure scenario	A geographic proxy concentrates audits in a community without a defensible risk relationship and without review of disparate impact.
Detection and response	Selection-distribution monitoring identifies concentration; freeze the model-assisted queue, conduct legal and statistical review, re-run affected selections, document corrections, and update governance thresholds.
Notably Absent	The example does not assert that risk scoring is unlawful per se or that equal selection rates are always the correct benchmark.

19.4 Fraud detection

Element	Treatment
Decision-authority level	Level 3
Context and role	Flags potential fraud for investigation but cannot block payment, impose sanctions, or create an adverse finding automatically.
Material risks	False positives, adversarial evasion, sensitive-data leakage, confirmation bias, and excessive investigative burden.
Required safeguards	Case-evidence review; calibrated alert thresholds; feedback controls; adversarial testing; separation between alert and adverse action.
Evidence	Alert logic; precision/recall by segment; investigation outcomes; override reasons; adversarial tests; incident records.
Failure scenario	Fraud actors learn a stable threshold and alter claims just below it, while the system continues to report apparently strong historical performance.
Detection and response	Outcome drift and threat intelligence reveal evasion; rotate detection features under change control, preserve affected cases, conduct retrospective sampling, and validate that new controls do not increase unjustified false positives.
Notably Absent	No claim is made that every flagged case is fraudulent or that high aggregate accuracy establishes lawful investigation.

19.5 Permit processing

Element	Treatment
Decision-authority level	Level 4
Context and role	Checks permit applications and recommends approval, refusal, or further evidence; an authorised officer owns the decision.
Material risks	Incorrect refusal, missing statutory discretion, inaccessible submissions, inconsistent local rules, and weak reasons.
Required safeguards	Current rule base; jurisdiction and date controls; exception routing; reason generation tied to evidence; human sign-off; appeal and correction.
Evidence	Rule provenance; policy-change log; test cases; officer review; notices; appeals; correction records.
Failure scenario	A planning-rule update is not propagated, causing compliant applications to be recommended for refusal under superseded criteria.
Detection and response	Change reconciliation or applicant challenge identifies the defect; stop automated recommendations, identify affected applications, issue corrected decisions or notices, update and validate the rule base, and record the incident.

Element	Treatment
Notably Absent	The example does not assume that codified rules eliminate statutory discretion or local variation.

19.6 Procurement assistant

Element	Treatment
Decision-authority level	Level 3
Context and role	Summarises bids and highlights risks; it cannot score, exclude, negotiate, or award without authorised procurement officials.
Material risks	Confidential bid leakage, hallucinated criteria, supplier bias, prompt injection in bid documents, and inadequate auditability.
Required safeguards	Isolated document processing; approved criteria; no cross-bid leakage; source-linked summaries; conflict checks; procurement review.
Evidence	Data-flow map; bid isolation tests; criteria configuration; source citations; reviewer changes; access logs.
Failure scenario	A bidder embeds instructions in a proposal that cause the assistant to suppress competitor risks and promote its own submission.
Detection and response	Indirect-injection testing or anomalous summary review detects manipulation; quarantine the document, regenerate summaries in a clean environment, inform procurement integrity personnel, preserve evidence, and assess whether the process must be repeated.
Notably Absent	No supplier ranking or award is presumed valid because it was reviewed after generation.

19.7 Public-records classification

Element	Treatment
Decision-authority level	Level 3
Context and role	Classifies and routes records for retention, disclosure review, or archival processing; legal disposition remains governed by records authorities.
Material risks	Misclassification, premature deletion, hidden personal data, disclosure error, and loss of decision reconstruction.
Required safeguards	Authoritative schedule mapping; confidence thresholds; human review for destructive or disclosure actions; immutable logs; sampling.
Evidence	Records schedule mapping; labelled test corpus; error analysis; disposition approvals; audit logs; restoration test.
Failure scenario	Low-confidence records are automatically assigned a short retention category and deleted before a public-records request is processed.
Detection and response	Sampling or request reconciliation detects missing records; halt automated disposition, restore recoverable records, notify records and legal owners, investigate scope, correct classifications, and strengthen destructive-action approvals.
Notably Absent	Automation does not replace statutory records authority, and classifier confidence does not establish legal disposition.

19.8 Emergency resource allocation

Element	Treatment
Decision-authority level	Level 4
Context and role	Recommends emergency resource allocation under time pressure; incident command retains authority and must be able to override.
Material risks	Unequal service, stale situational data, communications loss, unsafe optimisation, and inability to operate manually.
Required safeguards	Predefined objectives and constraints; real-time data quality checks; command approval; manual fallback; degraded-mode exercises; post-event review.
Evidence	Authority plan; data feeds; allocation logs; override records; continuity exercise; after-action report.
Failure scenario	A communications outage makes one district appear to have low demand, diverting resources away from a severely affected population.
Detection and response	Data-quality alarms and field reports trigger degraded mode; incident command overrides the recommendation, switches to verified manual reports, documents reallocations, preserves telemetry, and reviews model assumptions after stabilisation.
Notably Absent	No claim is made that optimisation alone can resolve competing emergency duties or that speed eliminates accountability.

19.9 Regulatory inspection prioritisation

Element	Treatment
Decision-authority level	Level 4
Context and role	Prioritises entities for inspection; inspectors and enforcement officials retain authority over inspection, findings, and sanctions.
Material risks	Selective enforcement, stale compliance data, vendor opacity, strategic evasion, and weak explanation of prioritisation.
Required safeguards	Legally relevant factors; random and risk-based sampling; bias and drift monitoring; reason codes; independent review; challenge process.
Evidence	Inspection policy; feature rationale; selection distributions; random-sample results; inspector feedback; complaints and reviews.
Failure scenario	The model learns that entities using a particular reporting format are higher risk, producing a persistent but spurious enforcement concentration.
Detection and response	Feature review and distribution monitoring detect the proxy; suspend the feature, re-evaluate queued inspections, assess past decisions, document legal review, and validate a revised model with random controls.
Notably Absent	The example does not assert that equal inspection rates are required or that model opacity automatically proves discrimination.

19.10 Drafting official notices

Element	Treatment
Decision-authority level	Level 4
Context and role	Drafts notices or decisions from case evidence; an authorised official must verify facts, law, reasons, remedy and final text.

Element	Treatment
Material risks	Fabricated facts or citations, omitted evidence, unlawful reasons, privacy disclosure, and false appearance of official approval.
Required safeguards	Source-grounded drafting; citation verification; controlled templates; redaction; mandatory substantive review; signed approval; version history.
Evidence	Source links; draft/final comparison; reviewer checklist; approval signature; notice delivery; correction and appeal records.
Failure scenario	The system invents a statutory reference and omits contrary evidence, and the reviewer approves the notice because the prose appears authoritative.
Detection and response	Legal quality review or appeal identifies the defect; withdraw or correct the notice, preserve draft and approval evidence, assess similarly generated notices, retrain reviewers, and update citation and contrary-evidence checks.
Notably Absent	Professional wording, source citations, or a signature do not by themselves establish factual or legal correctness.

20. Notably Absent

- No verified basis is presented for claiming that all public-sector AI deployments are high risk.
- No framework adoption, policy, model card, attestation, or certification alone proves operating effectiveness.
- No assumption is made that human approval is effective without competence, time, authority, independence, and evidence.
- No claim is made that model accuracy alone establishes legality, fairness, security, accessibility, procedural validity, or fitness.
- No claim is made that procurement or supplier certification transfers public accountability.
- No verified evidence is asserted here of autonomous AI lawmaking or wholesale replacement of constitutional or statutory authority.
- No absence of reported incidents is treated as proof of safety.
- No universal threshold, maturity score, or profile establishes a government acceptance criterion.

21. Limitations

- This guide is implementation guidance, not legal advice.
- Jurisdiction-specific legal authority and public-law duties require qualified review.
- Conformance evidence does not by itself prove operating effectiveness.
- Absence of a reported incident does not prove safety.
- GAISST implementation does not eliminate residual risk.
- Government functions, legal systems, records regimes, procurement rules, and oversight structures differ materially.
- Threat conditions, models, suppliers, integrations, and system behaviour change over time.
- External sources are contextual unless legally applicable; verify status, edition, date, and jurisdiction.

- This public candidate contains no universal financial exposure range, cost estimate, ROI claim, or dependency on an unverified ODA3 test repository. Test specifications are illustrative and must be adapted to the implementing entity's architecture and assurance method.

22. Source and Evidence Register

ID	Source	Issuer	Date	Class/ status	URL or location	Limitations	Last verified
SRC-001	GAISST-NOR-001 Framework Standard v1.0	ODA3 Institute	2026-06-29	Authoritative / Normative	Internal controlled source	Canonical 59-control baseline	30 Jun 2026
SRC-002	Artificial Intelligence Risk Management Framework (AI RMF 1.0)	NIST	2023-01-26	Class 1 / Advisory	https://doi.org/10.6028/NIST.AI.100-1	Voluntary risk-management framework	30 Jun 2026
SRC-003	Regulation (EU) 2024/1689 (Artificial Intelligence Act)	European Union	2024-07-12	Class 1 / Binding where applicable	https://eur-lex.europa.eu/eli/reg/2024/1689/oj	Jurisdiction-specific; verify applicability and phased dates	30 Jun 2026
SRC-004	G7 Toolkit for Artificial Intelligence in the Public Sector	OECD/G7	2024	Class 2 / Advisory	https://www.oecd.org/en/publications/g7-toolkit-for-artificial-intelligence-in-the-public-sector_421c1244-en.html	Public-sector implementation context	30 Jun 2026
SRC-005	OMB Memorandum M-25-21, Accelerating Federal Use of AI through Innovation, Governance, and Public Trust	US OMB	2025-04-03	Class 1 / Binding for covered US federal agencies	https://www.whitehouse.gov/omb/information-resources/guidance/memoranda/	Jurisdiction-specific and subject to supersession	30 Jun 2026

ID	Source	Issuer	Date	Class/ status	URL or location	Limitations	Last verified
SRC-006	NIST AI 600-1 Generative AI Profile	NIST	2024-07-26	Class 1 / Advisory	https://doi.org/10.6028/NIST.AI.600-1	Generative AI risk profile	30 Jun 2026

22.1 Source-authority classes

Tier	Meaning
Class 1 — Authoritative	Legislation, regulation, binding policy, official standards/directives, formal decisions, audit findings, and primary government sources.
Class 2 — Strong supporting	Recognised standards-body guidance, cybersecurity guidance, peer-reviewed research, and formal assurance frameworks.
Class 3 — Contextual	Credible incident reporting, technical analysis, public case studies, and documented operational experience.
Class 4 — Illustrative	Hypothetical or composite examples, practitioner observations, and unverified public claims.

Source verification note: time-sensitive legal, regulatory, directive and policy sources must be reverified immediately before publication. NIST AI RMF 1.0 is voluntary guidance; Regulation (EU) 2024/1689 is binding only where applicable; WCAG 2.2 is an advisory web accessibility standard unless adopted by the governing regime.

GAISST Ecosystem developed by ODA3 Institute, licensed under GEL v1.0. | © 2026 ODA3 Pvt Ltd | oda3.org