

ODA3 INSTITUTE

SECTOR IMPLEMENTATION GUIDE

GAISSF™ Defence Sector Implementation Guide

Global AI Safety and Security Framework

Mission assurance, security, safety, human authority and evidence

Document ID: GAISSF-SEC-040

Version: 1.0

Classification: Public · Non-normative sector implementation guidance

Status: Draft for Publication — Public — External Review Open

Issued: 29 June 2026

Updated: 5 October 2026

© 2026 ODA3 Pvt Ltd

Document Control

Field	Value
Status	Draft for Publication — Public — External Review Open
Normative baseline	GAISSF-NOR-001 v1.0; 59 controls across 9 domains. Applicable-Control Baseline = 52 D1-D8 controls + applicable D9 controls.
Classification	Non-normative sector implementation guidance
Issued	29 June 2026
Updated	5 October 2026
Publisher	ODA3 Institute
Legal entity	ODA3 Pvt Ltd
Document ID	GAISSF-SEC-040
Version	1.0

Copyright, Licensing and Reliance Notice

© 2026 ODA3 Pvt Ltd. Published by ODA3 Institute. Use is governed by the GAISSF Ecosystem Licence (GEL) v1.0. This guide is not legal or IHL advice, command advice, rules-of-engagement guidance, weapons-employment authority, certification, or a guarantee of security, safety, mission effectiveness or compliance.

How to Use This Guide

1. Classify the system and document mission context.
2. Trace mission requirements to threats, controls, tests, residual risk and approval.
3. Maintain owners and evidence in the workbook.
4. Escalate legal, IHL, safety, security and command questions to competent authorised functions.

Normative Boundary

Only GAISSF-NOR-001 control statements are normative. All defence interpretations, examples, profiles and thresholds are informative.

Methodology Note

Evidence sourcing follows ODA3's controlling T1-T4 evidence-sourcing taxonomy where that taxonomy is invoked; this sector guide does not redefine T1-T4. Control-evidence strength and implementation progression are recorded separately without reusing T1-T4 labels. Defence Implementation Stages A0-A5 describe implementation progression only; they are not GAISSF conformance tiers and must not be conflated with Foundational, Operational or Optimized.

Applicable-Control Baseline

GAISSF v1.0 contains 59 controls across nine domains. The canonical Foundational scope is the 52 controls in D1-D8. D9 contains seven additional/conditional controls and is mandatory where Physical AI or cyber-physical actuation is in scope. Operational uses the same applicable control scope with operating-effectiveness evidence. Optimized uses the same applicable control scope with continuous monitoring and enhanced assurance.

Physical AI and PAI-SF™

Where Physical AI or cyber-physical actuation is in scope, use GAISSF D9 alongside PAI-SF™ v1.0 as applicable. CSX-PAISF-004 provides mapping only; it does not establish equivalence, evidence sufficiency or automatic conformance.

Public / Classified Information Boundary

This public guide must not contain or be used to store classified operational data, mission parameters, target information or controlled technical data. Such records belong only in organisation-approved environments with the required classification, handling and access controls.

Executive Summary

Defence AI assurance integrates operational risk, programme accountability, safety, cybersecurity, legal acceptability, human authority and supplier control.

D1 - MODEL INTEGRITY & ADVERSARIAL ROBUSTNESS

D1-CTL-01 - DATASET PROVENANCE & POISONING PREVENTION

Defence interpretation: Establish mission-authorised lineage for every training, tuning and evaluation dataset, including source authority, classification, coalition restrictions, transformations and poisoning indicators.

Implementation actions: Implement hash verification + source allowlist + poisoning detection within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with hash, source, scan results, poisoning_score Chain of Custody: Signed by scanning service + independent assessor or reviewer review log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for dataset provenance & poisoning prevention using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track supplier/version changes, unsupported components, signature failures, vulnerability notices, licence or hosting changes and unapproved dependencies.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D1-CTL-01 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D1-CTL-02 - MODEL EXTRACTION RESISTANCE

Defence interpretation: Protect sovereign and mission-sensitive model capability from systematic query-based replication, insider copying and unauthorised weight access.

Implementation actions: Implement rate limiting + diversity detection + extraction monitoring within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `extraction_success_count`, `detection_alerts`, `rate_limit_logs` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for model extraction resistance using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to model extraction resistance.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D1-CTL-02 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D1-CTL-03 - BEHAVIORAL DRIFT DETECTION

Defence interpretation: Detect performance or behaviour change caused by new missions, sensor conditions, adversary tactics, operator practices or model/data updates before operational reliance becomes unsafe.

Implementation actions: Implement baseline profiling + kl divergence monitoring + accuracy tracking within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `daily_kl_divergence_values`, `accuracy_trend`, `alert_log` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for behavioral drift detection using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track mission-segment performance, class-specific error, calibration, data distribution, environment and operator override trends; trigger revalidation on meaningful change.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D1-CTL-03 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D1-CTL-04 - FEDERATED LEARNING POISONING PREVENTION

Defence interpretation: Prevent compromised coalition, laboratory or edge clients from manipulating shared model updates or revealing protected contributions.

Implementation actions: Implement gradient anomaly detection + robust aggregation within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with detection_rate, aggregation_log, client_anomaly_scores Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for federated learning poisoning prevention using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to federated learning poisoning prevention.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D1-CTL-04 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D1-CTL-05 - EMBEDDING SPACE ROBUSTNESS

Defence interpretation: Test whether adversarially crafted text, imagery, signals or retrieval content can distort similarity search, clustering or nearest-neighbour decisions used in defence workflows.

Implementation actions: Implement adversarial training + certified robustness measurement within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with classification_change_rate, certified_radius, attack_log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for embedding space robustness using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to embedding space robustness.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D1-CTL-05 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D1-CTL-06 - POST-QUANTUM MODEL SIGNING & CRYPTO HARDENING

Defence interpretation: Maintain cryptographic authenticity and long-term integrity of model weights, adapters and deployment packages for systems with extended defence lifecycles or store-now-decrypt-later exposure.

Implementation actions: Implement pqc signing (ml-dsa/slh-dsa) + pqc key exchange (ml-kem) within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with signature_verification_log, tls_cipher_suite_audit, pqc_migration_plan Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for post-quantum model signing & crypto hardening using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to post-quantum model signing & crypto hardening.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D1-CTL-06 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D1-CTL-07 - LORA/ADAPTER INTEGRITY VERIFICATION

Defence interpretation: Verify that every adapter or low-rank update is authorised, signed, provenance-traceable and screened for behavioural changes before integration into a defence model.

Implementation actions: Implement adapter scanning + provenance verification + registry allowlist within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with scan_results, hash, source_verification, detection_flag Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for lora/adapter integrity verification using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to lora/adapter integrity verification.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D1-CTL-07 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D1-CTL-08 - MODEL MERGE ATTACK DETECTION

Defence interpretation: Detect malicious or unintended capability changes introduced when combining checkpoints, experts, adapters or fine-tuned variants.

Implementation actions: Implement pre-registration behavioural evaluation + regression testing within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with behavioral_test_results, regression_delta, registration_audit Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for model merge attack detection using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to model merge attack detection.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D1-CTL-08 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D1-CTL-09 - QUANTIZATION BACKDOOR SCREENING

Defence interpretation: Screen compressed or quantised models for accuracy collapse, hidden backdoors and changed safety behaviour before fielding on constrained edge hardware.

Implementation actions: Implement cross-precision behavioural comparison + delta threshold monitoring within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with pre_quant_results, post_quant_results, behavioral_delta, audit_log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for quantization backdoor screening using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to quantization backdoor screening.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D1-CTL-09 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D2 - RUNTIME SECURITY & ADVERSARIAL DEFENSE

D2-CTL-01 - DIRECT PROMPT INJECTION PREVENTION

Defence interpretation: Prevent operators or external users from overriding system instructions, disclosure controls or authorised task boundaries through direct input.

Implementation actions: Implement input validation + adversarial pattern matching + system prompt isolation + guardrail sidecar within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `bypass_count`, `refusal_rate`, `payload_hashes` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for direct prompt injection prevention using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track blocked and successful attack patterns, instruction-source conflicts, policy bypass, tool invocation anomalies and affected data contexts.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D2-CTL-01 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D2-CTL-02 - INDIRECT PROMPT INJECTION PREVENTION

Defence interpretation: Prevent operators or external users from overriding system instructions, disclosure controls or authorised task boundaries through direct input.

Implementation actions: Implement contextual separation + source allowlisting + output validation + rag sanitization pipeline within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `sanitization_log`, `execution_rate`, `source_audit` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for indirect prompt injection prevention using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track blocked and successful attack patterns, instruction-source conflicts, policy bypass, tool invocation anomalies and affected data contexts.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D2-CTL-02 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D2-CTL-03 - JAILBREAK RESISTANCE TESTING

Defence interpretation: Measure and reduce the ability of users to bypass prohibited-use, disclosure or action constraints through iterative adversarial prompting.

Implementation actions: Implement quarterly red-team prompt library + adversarial training + automated refusal monitoring within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with success_rate, technique_breakdown, refusal_log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for jailbreak resistance testing using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track blocked and successful attack patterns, instruction-source conflicts, policy bypass, tool invocation anomalies and affected data contexts.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D2-CTL-03 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D2-CTL-04 - MULTI-MODAL INJECTION DEFENSE

Defence interpretation: Treat text, imagery, audio, video, metadata and sensor payloads as potential instruction channels and isolate untrusted content from control instructions.

Implementation actions: Implement multi-modal content scanning + steganography detection + modality-specific guardrails within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with modality_scan_results, detection_rate, payload_metadata Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for multi-modal injection defense using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track blocked and successful attack patterns, instruction-source conflicts, policy bypass, tool invocation anomalies and affected data contexts.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D2-CTL-04 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D2-CTL-05 - FUNCTION CALL/TOOL CALL INJECTION PREVENTION

Defence interpretation: Constrain model-triggered tool and function execution to approved actions, validated arguments, least privilege and explicit high-impact approval.

Implementation actions: Implement parameter schema validation + allowlist enforcement + sandboxed execution within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with validation_log, allowlist_hits, rejection_reasons Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for function call/tool call injection prevention using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track blocked and successful attack patterns, instruction-source conflicts, policy bypass, tool invocation anomalies and affected data contexts.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D2-CTL-05 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D2-CTL-06 - CROSS-CONTEXT HIJACKING MITIGATION

Defence interpretation: Prevent instructions, data or authority from one mission, user, tenant or classification context from influencing another.

Implementation actions: Implement context window segmentation + prompt anchoring + attention boundary enforcement within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with position_test_results, adherence_score, override_log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for cross-context hijacking mitigation using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track blocked and successful attack patterns, instruction-source conflicts, policy bypass, tool invocation anomalies and affected data contexts.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D2-CTL-06 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D3 - AGENTIC RISK & AUTONOMOUS SYSTEM SECURITY

D3-CTL-01 - LEAST AGENCY ENFORCEMENT

Defence interpretation: Give agents only the minimum tools, permissions, persistence and action horizon needed for the approved defence task.

Implementation actions: Implement role-based tool scoping + policy-as-code + dynamic permission revocation within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with permission_audit_log, denial_rate, revocation_latency Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for least agency enforcement using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to least agency enforcement.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D3-CTL-01 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D3-CTL-02 - INTER-AGENT COMMUNICATION SECURITY

Defence interpretation: Authenticate and authorise messages among agents, platforms and orchestration services so that one compromised component cannot impersonate trusted authority.

Implementation actions: Implement mtls for agent mesh + message signing + payload validation within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with tls_audit, message_validation_log, rejection_count Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for inter-agent communication security using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to inter-agent communication security.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D3-CTL-02 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D3-CTL-03 - AGENTIC PROMPT CHAINING DETECTION

Defence interpretation: Detect multi-step instruction chains that gradually move an agent beyond its approved objective, data boundary or action envelope.

Implementation actions: Implement cross-session behavioural correlation + chain pattern detection + anomaly scoring within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with chain_detection_log, correlation_scores, latency_metrics Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for agentic prompt chaining detection using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to agentic prompt chaining detection.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D3-CTL-03 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D3-CTL-04 - EMBODIED AI SAFETY CONTROLS

Defence interpretation: Constrain embodied or cyber-physical AI through verified perception, action limits, safe states, override and recovery behaviour.

Implementation actions: Implement sensor integrity verification + safety interlocks + fail-safe state enforcement within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with sensor_integrity_log, interlock_activation_log, fail_safe_metrics Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for embodied ai safety controls using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track sensor disagreement, command rejection, boundary approach, override use, safe-state entry, near misses and recovery success.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D3-CTL-04 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D3-CTL-05 - MULTI-AGENT TRUST CHAIN ATTESTATION

Defence interpretation: Require attestable identities, configurations, permissions and trust decisions across multi-agent mission teams.

Implementation actions: Implement spiffe/spire workload identity + short-lived certificates + continuous attestation within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `attestation_log`, `certificate_rotation_audit`, `rejection_rate` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for multi-agent trust chain attestation using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to multi-agent trust chain attestation.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D3-CTL-05 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D3-CTL-06 - PERSISTENT MEMORY EXFILTRATION PREVENTION

Defence interpretation: Prevent long-lived agent memory from retaining or exposing classified, personal, operational or cross-mission information.

Implementation actions: Implement user-scoped memory isolation + encryption at rest + query-level access controls within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `isolation_audit_log`, `leakage_rate`, `access_control_hits` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for persistent memory exfiltration prevention using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to persistent memory exfiltration prevention.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D3-CTL-06 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D3-CTL-07 - SECURE MEMORY LIFECYCLE MANAGEMENT

Defence interpretation: Control creation, classification, access, retention, correction and deletion of agent memory throughout deployment and retirement.

Implementation actions: Implement cryptographic deletion + lifecycle policy enforcement + retention auditing within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with lifecycle_audit, wipe_verification_log, retention_compliance Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for secure memory lifecycle management using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to secure memory lifecycle management.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D3-CTL-07 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D4 - SUPPLY CHAIN & THIRD-PARTY AI SECURITY

D4-CTL-01 - AI BILL OF MATERIALS (AI BOM) MAINTENANCE

Defence interpretation: Maintain an authoritative AI bill of materials covering model files, adapters, prompts, datasets, retrieval stores, libraries, tools, APIs, hardware dependencies and accountable suppliers.

Implementation actions: Implement automated bom generation + version tracking + registry synchronization within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON/SBOM-compatible with component_list, version_hashes, coverage_metric Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for ai bill of materials (ai bom) maintenance using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track supplier/version changes, unsupported components, signature failures, vulnerability notices, licence or hosting changes and unapproved dependencies.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D4-CTL-01 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D4-CTL-02 - MODEL FILE & ARTIFACT SCANNING

Defence interpretation: Scan model files, weights, archives and associated artefacts for malware, unsafe serialization, embedded executables, anomalous metadata and known malicious signatures before registry admission or deployment.

Implementation actions: Implement static analysis + deserialization sandboxing + signature verification within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with scan_results, quarantine_log, detection_metrics Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for model file & artifact scanning using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to model file & artifact scanning.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D4-CTL-02 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D4-CTL-03 - MODEL HUB & REGISTRY VETTING

Defence interpretation: Allow models only from approved registries or hubs that meet provenance, licence, security, custody and supportability criteria; quarantine unreviewed artefacts.

Implementation actions: Implement provenance verification + license compliance + security scorecard within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with provenance_log, license_audit, security_scorecard Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for model hub & registry vetting using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to model hub & registry vetting.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D4-CTL-03 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D4-CTL-04 - MCP SERVER BEHAVIORAL MONITORING

Defence interpretation: Treat Model Context Protocol servers and equivalent tool gateways as privileged supply-chain components requiring identity, scope, code and configuration assurance.

Implementation actions: Implement tool-call logging + anomaly detection + access control enforcement within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with tool_call_log, anomaly_scores, block_metrics Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for mcp server behavioral monitoring using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to mcp server behavioral monitoring.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D4-CTL-04 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D4-CTL-05 - THIRD-PARTY AI API SECURITY ASSESSMENT

Defence interpretation: Assess external AI APIs for authentication, data use, retention, jurisdiction, availability, model change, incident reporting, logging and exit before mission integration.

Implementation actions: Implement contractual security requirements + penetration testing + data flow mapping within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with vendor_name, assessment_date, security_gaps, remediation_plan Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for third-party ai api security assessment using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track supplier/version changes, unsupported components, signature failures, vulnerability notices, licence or hosting changes and unapproved dependencies.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D4-CTL-05 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D4-CTL-06 - SHADOW AI DISCOVERY & GOVERNANCE

Defence interpretation: Discover and govern unapproved AI services, local models, browser extensions, coding assistants and agent tools that process defence information outside authorised controls.

Implementation actions: Implement network traffic analysis + saas discovery + policy enforcement within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `discovery_log`, `unauthorized_count`, `governance_actions` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for shadow ai discovery & governance using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to shadow ai discovery & governance.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D4-CTL-06 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D4-CTL-07 - AI SOFTWARE COMPOSITION ANALYSIS (SCA)

Defence interpretation: Analyse AI software and model-serving stacks for vulnerable, malicious, unsupported or licence-incompatible components and transitive dependencies.

Implementation actions: Implement dependency scanning + cve matching + automated patching within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `dependency_list`, `cve_matches`, `remediation_status` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for ai software composition analysis (sca) using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to ai software composition analysis (sca).

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D4-CTL-07 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D5 - CONTENT SAFETY & OUTPUT INTEGRITY

D5-CTL-01 - HARMFUL CONTENT BLOCKING

Defence interpretation: Define mission-appropriate output constraints and escalation for content that could cause operational, human, legal or security harm, without suppressing legitimate intelligence or red-team analysis.

Implementation actions: Implement content safety classifier + refusal engine within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with refusal_rate, harmful_content_categories, refusal_log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for harmful content blocking using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to harmful content blocking.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D5-CTL-01 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D5-CTL-02 - PII LEAKAGE PREVENTION

Defence interpretation: Prevent disclosure of personal, classified, controlled unclassified, mission and coalition-restricted information through outputs, logs or model context.

Implementation actions: Implement pii detection + masking + access controls within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with pii_detection_recall, false_positive_rate, redaction_log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for pii leakage prevention using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to pii leakage prevention.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D5-CTL-02 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D5-CTL-03 - COPYRIGHT DETECTION

Defence interpretation: Identify copyright, licence and provenance concerns in generated or retrieved content before publication, distribution, training reuse or incorporation into defence deliverables.

Implementation actions: Implement n-gram overlap detection + refusal within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with ngram_overlap_scores, reproduction_rate, refusal_log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for copyright detection using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to copyright detection.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D5-CTL-03 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D5-CTL-04 - AI WATERMARKING ROBUSTNESS

Defence interpretation: Evaluate whether provenance marks or watermarks survive compression, transformation, adversarial removal and operational media pipelines; do not rely on watermarking as sole authenticity evidence.

Implementation actions: Implement c2pa-compliant watermarking + tamper resistance testing within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with detection_rate_by_attack, false_positives, c2pa_compliance_certificate Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for ai watermarking robustness using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to ai watermarking robustness.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D5-CTL-04 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D5-CTL-05 - PRIVACY-BY-DESIGN VERIFICATION

Defence interpretation: Verify that data minimisation, purpose limitation, access, retention, user rights and privacy risk controls are engineered into the system rather than added after deployment.

Implementation actions: Implement data minimization + purpose limitation + machine unlearning within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with data_inventory, purpose_audit_log, erasure_request_log, unlearning_attestation Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for privacy-by-design verification using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-

approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to privacy-by-design verification.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D5-CTL-05 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D5-CTL-06 - PRIVACY-PRESERVING ML VALIDATION

Defence interpretation: Validate that privacy-preserving techniques such as differential privacy, secure aggregation or confidential computing achieve stated protection without unacceptable mission-performance loss.

Implementation actions: Implement differential privacy + membership inference testing within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `dp_epsilon_value`, `membership_inference_results`, `dp_training_certificate` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for privacy-preserving ml validation using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to privacy-preserving ml validation.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D5-CTL-06 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D6 - GOVERNANCE, ACCOUNTABILITY & HUMAN OVERSIGHT

D6-CTL-01 - HUMAN-IN-THE-LOOP FOR HIGH-RISK ACTIONS

Defence interpretation: Require explicit, competent and timely human approval for high-consequence actions, with sufficient information, independent corroboration and authority to refuse or abort.

Implementation actions: Implement approval workflow + policy enforcement + audit log within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `approval_request_log`, `approver_id`, `timestamp`, `justification` Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for human-in-the-loop for high-risk actions using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track operator error, escalation, override quality, simulation outcomes, identity-verification failures and recurring human-factor conditions.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D6-CTL-01 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D6-CTL-02 - AUDIT TRAIL COMPLETENESS

Defence interpretation: Capture complete, time-synchronised and tamper-evident records of model version, inputs, outputs, tool calls, human decisions, overrides, configuration and data access.

Implementation actions: Implement structured logging + siem integration + retention enforcement within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON/SIEM log with decision_id, timestamp, input_hash, output_hash, system_id, approver_id Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for audit trail completeness using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to audit trail completeness.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D6-CTL-02 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D6-CTL-03 - AI MODEL CARD COMPLETENESS

Defence interpretation: Maintain a controlled model card describing purpose, prohibited use, training/evaluation context, limitations, dependencies, safety/security tests, monitoring and responsible owners.

Implementation actions: Implement standardized template + version control + public accessibility within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: Markdown/JSON with model_id, owner, purpose, data_sources, limitations, risk_tier, review_date Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for ai model card completeness using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-

approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to ai model card completeness.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D6-CTL-03 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D6-CTL-04 - AI INCIDENT RESPONSE READINESS

Defence interpretation: Integrate AI-specific failure, misuse and compromise scenarios into defence incident command, evidence preservation and recovery.

Implementation actions: Implement ai-ir runbook + tabletop exercises + containment automation within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with exercise_date, participants, mttc_achieved, lessons_learned, improvement_tracker Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for ai incident response readiness using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to ai incident response readiness.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D6-CTL-04 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D6-CTL-05 - MODEL DEPRECATION & DECOMMISSIONING

Defence interpretation: Withdraw obsolete or unsafe models, remove dependencies, preserve records and prevent unauthorised continued use.

Implementation actions: Implement access revocation + decommission audit + scheduled lifecycle within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with model_id, decommission_date, access_revocation_log, traffic_verification Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for model deprecation & decommissioning using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier

working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to model deprecation & decommissioning.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D6-CTL-05 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D6-CTL-06 - THIRD-PARTY AI VENDOR GOVERNANCE

Defence interpretation: Assign acquisition, security and programme authorities to oversee vendor evidence, updates, incidents, subcontractors and exit.

Implementation actions: Implement contractual security requirements + annual assessment + audit rights within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with vendor_name, assessment_date, soc2_status, contract_review_summary Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for third-party ai vendor governance using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track supplier/version changes, unsupported components, signature failures, vulnerability notices, licence or hosting changes and unapproved dependencies.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D6-CTL-06 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D6-CTL-07 - AI RESILIENCE & BUSINESS CONTINUITY

Defence interpretation: Maintain mission continuity through degraded modes, alternate processes, local capability, rollback and recovery testing.

Implementation actions: Implement failover systems + degraded mode + rto/rpo definition within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with bcp_test_date, rto_achieved, rpo_achieved, degraded_mode_capabilities, improvement_tracker Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for ai resilience & business continuity using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and

inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to ai resilience & business continuity.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D6-CTL-07 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D7 - HUMAN & SOCIETAL HARMS

D7 boundary: D7 is the canonical GAISSF domain Human & Societal Harms. In GAISSF v1.0, its five controls address AI-enabled phishing, deepfakes, out-of-band authentication, social-engineering incident response and AI-enhanced external attack defence. Bias, fairness and generic human-factors concerns may be relevant elsewhere in the assurance case, but they do not redefine D7 or replace its canonical control set.

D7-CTL-H01 - AI-GENERATED PHISHING SIMULATION

Defence interpretation: Prepare personnel to recognise AI-generated phishing and spear-phishing without normalising unsafe simulation practices.

Implementation actions: Implement simulation campaigns + click tracking + remedial training within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with simulation_date, click_rate, coverage_percentage, training_completion Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for ai-generated phishing simulation using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track AI-enabled phishing, impersonation/deepfake, identity-verification, social-engineering incident, external-attack, operator escalation and response outcomes relevant to this D7 control. Human-factors observations may be supporting evidence but do not redefine D7 as a generic human-factors or bias domain.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D7-CTL-H01 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D7-CTL-H02 - DEEPPAKE DETECTION TRAINING

Defence interpretation: Detect and label synthetic or manipulated media while preserving escalation routes for uncertain or contested results.

Implementation actions: Implement training modules + quiz + simulated attacks within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with training_date, completion_rate_by_role, quiz_pass_rate Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for deepfake detection training using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track AI-enabled phishing, impersonation/deepfake, identity-verification, social-engineering incident, external-attack, operator escalation and response outcomes relevant to this D7 control. Human-factors observations may be supporting evidence but do not redefine D7 as a generic human-factors or bias domain.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D7-CTL-H02 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D7-CTL-H03 - OUT-OF-BAND AUTHENTICATION

Defence interpretation: Require an independent communication channel before acting on high-impact requests that may be AI-generated or impersonated.

Implementation actions: Implement independent channel verification + policy enforcement within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with oob_enforcement_log, policy_compliance_report, exception_log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for out-of-band authentication using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track AI-enabled phishing, impersonation/deepfake, identity-verification, social-engineering incident, external-attack, operator escalation and response outcomes relevant to this D7 control. Human-factors observations may be supporting evidence but do not redefine D7 as a generic human-factors or bias domain.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D7-CTL-H03 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D7-CTL-H04 - AI SOCIAL ENGINEERING IR

Defence interpretation: Define containment, identity verification, evidence handling and communications actions for AI-enabled social-engineering incidents.

Implementation actions: Implement tabletop exercises + ir plan + verification triggers within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with exercise_date, participants, verification_triggered, financial_hold_applied, improvements Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for ai social engineering ir using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track AI-enabled phishing, impersonation/deepfake, identity-verification, social-engineering incident, external-attack, operator escalation and response outcomes relevant to this D7 control. Human-factors observations may be supporting evidence but do not redefine D7 as a generic human-factors or bias domain.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D7-CTL-H04 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D7-CTL-H05 - AI-ENHANCED EXTERNAL ATTACK DEFENSE

Defence interpretation: Adapt perimeter, identity, messaging and monitoring controls to attacks amplified by generative AI and automated reconnaissance.

Implementation actions: Implement ai-generated phishing detection + soc tuning + response automation within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with detection_time, quarantine_action, soc_alert_log Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for ai-enhanced external attack defense using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track AI-enabled phishing, impersonation/deepfake, identity-verification, social-engineering incident, external-attack, operator escalation and response outcomes relevant to this D7 control. Human-factors observations may be supporting evidence but do not redefine D7 as a generic human-factors or bias domain.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D7-CTL-H05 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D8 - REGULATORY ALIGNMENT & COMPLIANCE

D8-CTL-01 - EU AI ACT RISK TIER MAPPING

Defence interpretation: Record whether the EU AI Act is applicable and map only relevant obligations; do not infer legal applicability from technical similarity.

Implementation actions: Implement risk classification framework + conformity assessment within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with system_id, risk_tier, classification_justification, conformity_evidence Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for eu ai act risk tier mapping using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to eu ai act risk tier mapping.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D8-CTL-01 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D8-CTL-02 - ISO 42001 GAP ANALYSIS

Defence interpretation: Use ISO/IEC 42001 mapping to identify management-system evidence without treating crosswalk completion as certification.

Implementation actions: Implement gap analysis methodology + remediation tracking within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with clause_by_clause_status, remediation_plan, completion_tracker Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for iso 42001 gap analysis using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to iso 42001 gap analysis.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D8-CTL-02 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D8-CTL-03 - GPAI TECHNICAL DOCUMENTATION VERIFICATION

Defence interpretation: Maintain technical documentation needed to evaluate general-purpose AI use, modifications, limitations and downstream responsibilities.

Implementation actions: Implement technical documentation + training data summary + copyright attestation within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with model_id, technical_doc_hash, training_data_summary, copyright_attestation Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for gpai technical documentation verification using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to gpai technical documentation verification.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D8-CTL-03 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D8-CTL-04 - DORA ICT INCIDENT REPORTING (FINANCIAL SECTOR)

Defence interpretation: Retain the authoritative GAISSF title. For defence, apply the underlying incident-reporting intent to applicable defence cyber, classified, safety and operational reporting regimes; DORA itself applies only where legally in scope.

Implementation actions: Implement incident classification + notification workflow + sla monitoring within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with incident_id, classification, notification_timestamp, nca_submission Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for dora ict incident reporting (financial sector) using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to dora ict incident reporting (financial sector).

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D8-CTL-04 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D8-CTL-05 - NIST SP 800-218A COMPLIANCE CHECK

Defence interpretation: Use secure AI development practices to govern requirements, design, code, data, models, testing, release and vulnerability response.

Implementation actions: Implement secure development practices + attestation within a controlled defence architecture. Record the accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: JSON with `sp800_218a_practice_status`, `attestation_statement`, `continuous_monitoring_log`. Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for nist sp 800-218a compliance check using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track control-specific alerts, exceptions, bypass attempts, version/configuration changes, test regressions and incidents relevant to nist sp 800-218a compliance check.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D8-CTL-05 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D9 - PHYSICAL AI SAFETY

D9-CTL-01 - PHYSICAL HARM BOUNDARY ENFORCEMENT

Defence interpretation: Define enforceable physical action boundaries based on mission, safety, legal and command constraints.

Implementation actions: Implement an independent physical-harm boundary mechanism appropriate to the system architecture and hazard analysis. Enforce approved operating envelopes, command constraints and fail-safe responses outside the AI model itself where necessary. Record accountable owner, approved configuration, classified-data boundary, supplier dependency, fallback mode and change triggers.

Evidence expected: Hardware-in-the-loop test report (JSON with `command_id`, `value`, `blocked`, `monitor_response_time_ms`); safety-monitor assurance evidence appropriate to the system and hazard analysis. Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for physical harm boundary enforcement using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track sensor disagreement, command rejection, boundary approach, override use, safe-state entry, near misses and recovery success.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D9-CTL-01 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D9-CTL-02 - SAFE STATE AND GRACEFUL DEGRADATION

Defence interpretation: Ensure physical AI enters a verified safe or least-harm state when confidence, communications, sensors, power or control integrity is lost.

Implementation actions: Define and implement a verified safe or least-harm state for loss of confidence, sensor disagreement, communications failure, resource exhaustion and other relevant fault conditions. Demonstrate controlled degradation and recovery using mission-representative scenarios; record accountable owner, fallback authority and change triggers.

Evidence expected: Safe state test report (JSON with trigger_condition, transition_time_ms, safe_state_achieved, assessor_verification). Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for safe state and graceful degradation using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track sensor disagreement, command rejection, boundary approach, override use, safe-state entry, near misses and recovery success.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D9-CTL-02 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D9-CTL-03 - HUMAN OVERRIDE AND EMERGENCY STOP

Defence interpretation: Provide tested, accessible and authoritative override and emergency-stop mechanisms that operate within the available intervention window.

Implementation actions: Provide tested human override and emergency-stop mechanisms proportionate to the system and mission. Ensure the override path does not depend on the same AI function it is intended to stop, and document authority, accessibility, degraded-mode behavior and recovery controls.

Evidence expected: Override test report (JSON with override_type, test_condition, latency_ms, outcome). Physical emergency-stop design, test and assurance documentation appropriate to the system. Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for human override and emergency stop using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track sensor disagreement, command rejection, boundary approach, override use, safe-state entry, near misses and recovery success.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D9-CTL-03 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D9-CTL-04 - CYBER-PHYSICAL ATTACK DETECTION

Defence interpretation: Detect attacks that manipulate sensors, commands, timing, navigation, communications or actuators before unsafe physical effect.

Implementation actions: Implement layered detection and containment for cyber-physical manipulation of sensors, commands, navigation, timing, communications and actuation. Use mission-representative threat scenarios and maintain independent containment or safe-state mechanisms appropriate to the hazard analysis.

Evidence expected: Hardware-in-the-loop detection test report (JSON with `attack_type`, `injected_count`, `detected_count`, `detection_rate`, `false_positive_rate`, `mean_time_to_safe_state_ms`). Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for cyber-physical attack detection using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track sensor disagreement, command rejection, boundary approach, override use, safe-state entry, near misses and recovery success.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D9-CTL-04 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D9-CTL-05 - PHYSICAL ENVIRONMENT INTEGRITY MONITORING

Defence interpretation: Monitor environmental conditions and sensor integrity that can invalidate perception or control assumptions.

Implementation actions: Monitor sensor health, environmental conditions, calibration state and cross-sensor consistency to detect conditions that invalidate safe operation. Define thresholds and response logic from system-specific engineering analysis rather than universal GAISSF values.

Evidence expected: Sensor failure injection test report (JSON with `failure_mode`, `response_time_ms`, `system_response`, `assessor_verification`); calibration certificate register. Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for physical environment integrity monitoring using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track sensor disagreement, command rejection, boundary approach, override use, safe-state entry, near misses and recovery success.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D9-CTL-05 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D9-CTL-06 - ACTUATOR COMMAND VERIFICATION

Defence interpretation: Authenticate, validate, sequence and constrain commands before they reach actuators or mission-effecting interfaces.

Implementation actions: Authenticate, validate, sequence and constrain actuator commands before execution. Apply independent verification, rate/sequence checks and approval controls proportionate to the consequence of actuation; define acceptance criteria from the physical system specification and hazard analysis.

Evidence expected: Hardware-in-the-loop verification test report (JSON with `command_id`, `command_type`, `block_reason`, `gate_response_time_ms`, `dual_approval_test_result`). Design-assurance or functional-safety evidence appropriate to the system, hazard analysis and competent authority. Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for actuator command verification using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track sensor disagreement, command rejection, boundary approach, override use, safe-state entry, near misses and recovery success.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D9-CTL-06 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

D9-CTL-07 - PHYSICAL INCIDENT EVIDENCE PRESERVATION

Defence interpretation: Preserve tamper-evident sensor, command, model, configuration and human-decision records after physical-AI incidents.

Implementation actions: Preserve tamper-evident, time-correlated sensor, model, command, safety-monitor and human-override records sufficient to reconstruct physical incidents. Retention, signing, storage location and incident-record integration shall follow approved mission, classification, legal and evidence-preservation requirements. UAIF™ may be used where published, current and applicable; it is not automatically required.

Evidence expected: Incident evidence package (structured evidence bundle (OSCAL may be used as an optional external serialization where appropriate) with `sensor_stream`, `model_input_output_stream`, `actuator_command_stream`, `safety_monitor_log`, `human_override_log`, `cryptographic_chain_of_custody`). Retention shall follow the approved records schedule and applicable legal, contractual, mission, classification and evidence-preservation requirements; GAISSF conformance tiers do not define universal retention periods.

Test method: Execute controlled positive, negative, adversarial and regression testing for physical incident evidence preservation using mission-representative inputs, relevant degraded/failure conditions and independently controlled samples. Use organisation-approved tooling. Acceptance criteria, sample sizes, latency limits, error bounds and test duration shall be defined and approved for the specific system, mission, hazard analysis and assurance objective; values in earlier working drafts are not universal GAISSF pass criteria. An independent assessor or reviewer should be able to reproduce or challenge the method and inspect surviving evidence. Where any GAISSF test asset is published, current and applicable, it may be used as one method; no unreleased VTS, benchmark repository or test harness is required by this guide.

Monitoring: Track sensor disagreement, command rejection, boundary approach, override use, safe-state entry, near misses and recovery success.

Pass indicator: Authorised configuration is implemented; the control-specific test passes against approved mission criteria; exceptions and residual risk are recorded; rollback or containment is demonstrated.

Fail indicator: Uncontrolled or unsigned artefact, absent owner, supplier assertion without acquirer evidence, test failure, unreviewed material change, or operation outside approved scope.

Notably Absent: Evidence for D9-CTL-07 does not prove universal resistance, mission suitability, legality, safety or effectiveness outside the tested configuration and conditions.

Consolidated Notably Absent

- No legal, regulatory or IHL compliance claim.
- No AI output establishes target identity, command authority or lawful engagement.
- Supplier assurance does not replace acquiring-organisation testing.
- Testing, logging and human presence do not eliminate operational risk.

Publication Status

Draft for Publication — Public — External Review Open. ODA3 technical reconciliation has been completed against the locked GAISSF v1.0 baseline. Defence technical, cybersecurity/adversarial-ML, systems-safety, human-factors, operational T&E, autonomy, legal/IHL, national-law/export-control, classified-information-handling and final publication approvals remain open unless separately evidenced in the Publication Gate Register. No reviewer identity, approval or legal conclusion is implied by this technical-candidate status.