

ODA3 INSTITUTE

TECHNICAL REPORT

AI-IRF™ v1.0 Technical Report

AI Security Incident Response Framework — Technical Requirements

ODA3-2026-06-TCR-HAI-001 | Public

© ODA3 Pvt Ltd

DOCUMENT CONTROL

Item	Value
Document ID	ODA3-2026-06-TCR-HAI-001
Doc Type	TCR — Technical Report
Framework Tag	AI-IRF™ v1.0
Publication History	Published 1 June 2026; regulatory correction issued 19 July 2026; source-verification, control-integrity and editorial correction issued 4 October 2026
Publication State	Published
Companion Document	AI-IRF™ v1.0 Executive Brief (ODA3-2026-06-EXB-HAI-002)
Review Cycle	Quarterly, with gap analysis against emerging threats, standards, and regulatory changes
Evidence Tier Legend	T1 — Primary Verified T2 — Secondary Verified T3 — Controlled Simulation / Academic Proxy T4 — Anecdotal / Unverified
Normative Language	SHALL = mandatory control SHOULD = recommended control
Classification	Public

Correction note (4 October 2026): This reissue corrects evidence sources, the standards crosswalk, regulatory references and control-register defects without changing the AI-IRF™ v1.0 framework version. Section 7 is authoritative: SAI-5 and MLC-5 sit in their domain tables, giving 79 unique controls; Phase 0 references rather than redefines the SAI controls; and validation methods were corrected where they did not test the stated requirement. No control IDs were added or removed.

Evidence Tier Definitions: All empirical claims in this document are assigned an evidence tier. **T1 — Primary Verified** = directly observed and validated. **T2 — Secondary Verified** = cross-referenced from two or more independent sources. **T3 — Controlled Simulation / Academic Proxy** = derived from controlled testing, simulation, or peer-reviewed academic work. **T4 — Anecdotal / Unverified** = single unverified report, illustrative scenario, or claim not independently confirmed; not presented as established fact.

SECTION 1

EXECUTIVE SUMMARY & POSITIONING



1.1 Executive Summary

AI-IRF™ v1.0 is an incident response framework engineered for agentic, machine-speed, and physically-consequential AI systems operating under evolving regulatory requirements and staged application timelines. [T2] It addresses nineteen critical operational gaps identified across existing industry frameworks and normative extension addenda by transitioning from traditional observe-and-respond models to a continuously verify, constrain, and recover paradigm — with mandatory pre-execution policy gates, cryptographic artifact governance, and adaptive automation.

LICENCE NOTICE — GEL v1.0 | ODA3-2026-06-TCR-HAI-001 | Public

This document is part of the GAISSE Ecosystem published by ODA3 Institute (ODA3 Pvt Ltd) under GEL v1.0 (effective 1 May 2026); it was adopted under GEL v1.0 on 1 June 2026. It is classified Public. Public classification governs disclosure only; it does not grant redistribution, derivative-work, or commercial-use rights beyond those set out in GEL v1.0. Commercial use — including embedding in training programmes, advisory services, or tool integrations — requires a separate written commercial licence from ODA3 Institute (GEL §3.4). AI training data use prohibited without licence (GEL §3.6). Evidence tiers: T1 = Primary Verified | T2 = Secondary Verified | T3 = Controlled Simulation/Academic Proxy | T4 = Anecdotal/Unverified. Financial figures in this document include an estimation formula and confidence range as required by the ODA3 Institute evidence standard. GAISSE™, UAIF™, AI-IRF™, AI IRF CERT™, and ODA3™ are trademarks of ODA3 Pvt Ltd, asserted on a use-in-commerce basis. Registration applications are pending. No statement in this document represents any mark as registered unless and until registration has been granted in the relevant jurisdiction (GEL §9.1). As of 19 July 2026, the AI IRF CERT™ scheme is under development and is not represented as accredited or operational. ODA3 Institute is the sole owner, proprietor, and governing body of the AI-IRF™ framework in perpetuity (GEL §§10.1, 10.5). Disclaimer: provided “AS IS” without warranties (GEL §12). Governing law: Republic of India; DIAC arbitration, New Delhi seat (GEL §14). All enquiries: <https://oda3.org/> — Attribution (GEL §6): AI-IRF™ v1.0 Technical Report © ODA3 Institute, GEL v1.0. Companion document: AI-IRF™ v1.0 Executive Brief (ODA3-2026-06-EXB-HAI-002).

The gap identification methodology analyzed: CoSAI V1.0, CoSAI AI IRF V1.0 (November 2025), NIST SP 800-61r3, OWASP LLM Top 10 v1.0 (2023), OWASP Top 10 for Agentic Applications 2026, OASIS CACAO v2.0, MITRE ATLAS, and ISO/IEC 42001:2023 against documented incident patterns and published research from mid-2025 to Q2 2026. [T2]

→ **COMPANION DOCUMENT:** For Board and CISO governance decisions, financial exposure modeling, and regulatory application-timeline implications, see the AI-IRF™ v1.0 Executive Brief.

Core Capability	Operational Impact	Gap Addressed
Protocol-Aware Security	MCP and A2A endpoints treated as primary attack surfaces with authenticated, inspected, and policy-enforced traffic	G1
Runtime Discovery	Continuous instance-aware visibility eliminating shadow AI blind spots	G2
Machine-Speed Response	Automated detection and containment at adversarial velocity (ms–seconds) with HITL gates for critical actions	G3

Extraction Defense	Behavioral distillation detection, dynamic rate limiting, and cryptographic output perturbation	G4
Cryptographic Provenance	Signed AI-BOM for all production artifacts enabling forensic reconstruction and regulatory audit	G5
Agentic Permissioning	Pre-execution validation, task-scoped least-privilege, and auto-expiring tokens	G6
Regulatory Decision Logic	Embedded jurisdictional decision trees mapping incident type to notification deadlines	G7
Safety-Security Convergence	Integrated cyber-physical incident command for AI systems with actuator control	G8
Lifecycle-as-Control	Immutable model identifiers, signed artifacts, and 15-minute rollback SLA	G9
Zero Trust Architecture	Identity-first, micro-segmented, continuously verified AI infrastructure	G10
Detection Reliability Engineering	Ensemble-validated auto-containment preventing false-positive cascades and single-detector dependency risk	G11
Agentic Threat Detection	Prompt injection, tool abuse, and memory poisoning detection distinct from the APA permission architecture	G12
Deterministic Severity Classification	Algorithmic, versioned severity scoring and reportability determination removing case-by-case judgement under time pressure	G13
Content Safety & Societal Harm	Harmful content filtering, copyright protection, synthetic media disclosure, bias/fairness monitoring, vulnerable population protections	G14
Federated Learning Security	Byzantine-robust aggregation, differential privacy, and client authentication for distributed AI training	G15
Model Merge Defense	Pre-merge provenance vetting and post-merge backdoor scanning for combined fine-tuned models	G16
Quantization Backdoor Screening	Pre/post-quantization behavioural baseline comparison preventing precision-reduction-triggered backdoors	G17
Hallucination Detection & Mitigation	RAG grounding verification, confidence calibration, cross-model consistency, and rate monitoring for factual AI failures	G18
Post-Quantum Cryptographic Migration	Phased readiness assessment and algorithm migration for AI-BOM signing, endpoint auth, and regulatory evidence integrity	G19

1.2 Framework Scope & Applicability

AI-IRF™ v1.0 applies to: large language models (LLMs), multimodal models, and foundation models in production; Retrieval-Augmented Generation (RAG) pipelines; agentic architectures with autonomous planning, tool execution, and memory persistence; AI systems with physical actuation capabilities; and hybrid cloud/edge/on-premises deployments.

NOTABLY ABSENT — Scope Limitation

- ◆ Traditional software incident response (addressed by NIST SP 800-61r3) — this framework does not replicate or replace general IR guidance.
- ◆ Non-AI data breach response governed by GDPR/NIST frameworks where AI is not a causal factor.
- ◆ Ethical AI governance and algorithmic bias remediation (addressed by ISO/IEC 42001 and the EU AI Act prohibited practices articles).
- ◆ Federated learning security and differential privacy implementation are addressed in the AI-IRF™ v1.0 normative extension documents (ODA3-ECO-013: FL-1–4 controls). These controls are part of the AI-IRF™ v1.0 control scheme as of 1 June 2026.
- ◆ Detection Reliability, Agentic Threat Detection, and Severity Classification (Domains 11–13: REL-1–3, APT-1–3, SEV-CLS/SEV-RPT) are part of the AI-IRF™ v1.0 control scheme and are specified in Section 7 of this document.
- ◆ Content Safety and Societal Harm (Domain 14: CSS-1–7) is part of the AI-IRF™ v1.0 control scheme via ODA3-ECO-011. Model Merge Defense and Quantization Security (Domains 16–17: MMD-1–3, QBS-1–3) are part of the AI-IRF™ v1.0 control scheme via ODA3-ECO-013.
- ◆ Hallucination Detection (Domain 18: HAL-1–5) is part of the AI-IRF™ v1.0 control scheme via ODA3-ECO-014. Post-Quantum Cryptography (Domain 19: PQC-1) operationalises the phased migration schedule in ODA3-ECO-015 across BOM-4, MCP-1, A2A-2, and REG-4.
- ◆ AI system design security (secure SDLC for AI) — this framework covers incident response, not pre-deployment design security.

SECTION 2

2026 SECURITY ASSUMPTIONS & THREAT BASELINE



2.1 Threat and Operational Assumptions

The following assumptions are not theoretical forecasts — they reflect documented incidents and published research from mid-2025 to Q2 2026 and published threat intelligence. [\[T2\]](#)

Assumption	Evidence	Tier
Adversaries use commercial LLMs to automate and scale intrusion operations, compressing the time defenders have to respond	Amazon Threat Intelligence (Feb 2026): a financially motivated actor used commercial generative AI services to compromise 600+ FortiGate devices across 55 countries between 11 January and 18 February 2026. Notably absent: no FortiGate vulnerability was exploited — access came through exposed management ports and weak single-factor credentials — and the actor was assessed as low-to-medium skilled.	T2 — Secondary Verified
MCP servers and agent-skill marketplaces are exposed and actively targeted attack surfaces	Trend Micro (Jul 2025): 492 network-exposed MCP servers running without client authentication or traffic encryption (a later Trend Micro update reported 1,467). Antiy CERT (Feb 2026): at least 1,184 malicious skills historically published on the ClawHub marketplace. Notably absent: these findings document exposure and supply-chain poisoning; they do not by themselves establish routine A2A lateral movement.	T2 — Secondary Verified
Large-scale capability extraction through legitimate-looking API access has been demonstrated	Anthropic disclosure (23 Feb 2026): DeepSeek, Moonshot and MiniMax allegedly generated more than 16 million exchanges with Claude through approximately 24,000 fraudulent accounts. OpenAI has made separate, less detailed allegations against DeepSeek. Notably absent: the named companies have not publicly confirmed these allegations, which are company-reported and not independently adjudicated.	T2 — Secondary Verified
Shadow AI is a likely organizational state where continuous discovery controls are absent	Censys (late Jan 2026): publicly exposed OpenClaw instances grew to roughly 21,000 within days; later scans reported higher counts. Notably absent: exposure counts alone do not establish that the deploying organizations were unaware of these deployments.	T2 — Secondary Verified
Agentic over-provisioning can cause serious incidents without any adversary involved	Meta internal AI agent (Mar 2026), reported by The Information and confirmed by Meta: an agent posted guidance to an internal forum without the requesting engineer's approval; acting on it left sensitive company and user data accessible to unauthorized employees for about two hours, classified internally as Sev 1. Notably absent: Meta stated that no user data was	T2 — Secondary Verified

	mishandled; no adversary was involved.	
Supply chain attacks target AI artifacts — training data, embeddings, MCP servers, orchestration dependencies	Vercel / Context.ai (Apr 2026), per both companies' disclosures: an infostealer infection at Context.ai led to theft of OAuth tokens; one token — a broad grant by a Vercel employee to Context.ai's AI Office Suite — was used to pivot into Vercel environments. Notably absent: a USD 2M forum listing of the data used the ShinyHunters name, but that group denied involvement; the listing's claims are unverified.	T2 — Secondary Verified
Regulators demand real-time structured notifications — narrative reports are insufficient	EU AI Act Article 73 serious-incident reporting is not a universal 72-hour rule: reporting is immediate once the required causal link or reasonable likelihood is established, subject to statutory outer limits of 15 days, two days for specified widespread infringements or serious incidents, and 10 days where a death occurs. Application timing depends on the system category and the operative amended timetable. CIRCI obligations depend on the final rule and its effective date	T1 — Primary Verified

NOTABLY ABSENT — Scope Limitation

- ◆ Post-quantum cryptography (PQC) readiness for AI model signing is addressed in ODA3-ECO-015 (PQC-1 control) as a normative Tier 3 requirement of the AI-IRF™ v1.0 control scheme. Quantum computing attacks on AI inference infrastructure remain post-2026 in threat horizon.
- ◆ AI-generated disinformation at scale is partially addressed through ODA3-ECO-011 (CSS-4: Disinformation and Manipulation Prevention). For full content safety governance, see the Content Safety addendum (ODA3-ECO-011, CSS-1–7 controls).
- ◆ AI system failures attributable to poor data quality or statistical drift without adversarial causation are excluded from this framework's incident taxonomy.

SECTION 3

CORE DESIGN PRINCIPLES



All seven design principles are normative: organizations implementing AI-IRF™ v1.0 **SHALL** comply with P1–P7 for any AI system within the framework's scope (see Section 1.2). **SHOULD** qualifiers are reserved for implementation methods where multiple valid approaches exist.

Principle	Operational Definition	Enforcement Mechanism	Normative Level
P1: Never Trust, Always Verify	Every agent, tool call, inference request, memory read, and A2A message is authenticated and authorized per interaction. No implicit trust based on network location or prior approval.	SPIFFE/SPIRE workload identity federation, per-request policy evaluation, continuous re-authentication	SHALL
P2: Machine-Speed Default, HITL for Critical	Automated detection and containment operate at adversarial velocity (ms–seconds). High-impact actions require explicit human approval via dual-control workflow.	LLM-as-a-Judge (LLM-J) scoring, confidence-thresholded automation, immutable approval audit trails	SHALL
P3: Cryptographic Provenance by Default	Every model weight, embedding, training dataset, prompt template, and MCP endpoint has an immutable, cryptographically signed manifest entry.	ECDSA/RSA signing, SHA-256 hashing; blockchain-anchored audit logs SHOULD be deployed for high-risk systems	SHALL
P4: Discovery Over Inventory	Continuous runtime discovery of AI instances across network edges, SaaS connectors, endpoints, and browser extensions — not periodic static CMDB updates.	Egress DLP with AI fingerprinting, EDR with AI process signatures, CASB with AI capability inventory	SHALL
P5: Safety-Security Convergence	Cyber incidents with physical harm pathways require integrated incident command with functional safety teams (ISO 26262 / IEC 61508).	Joint cyber-safety tabletop exercises (quarterly), fail-safe fallback modes, hardware-rooted trust attestation	SHALL for physical AI systems; SHOULD for all others
P6: Audit-Ready Telemetry	All AI interactions are logged with cryptographic integrity, structured schemas, and PII-redaction capabilities for regulatory submission and forensic reconstruction.	Immutable SIEM ingestion, signed log chains, automated redaction pipelines	SHALL
P7: Lifecycle-as-Security-Control	Model versioning, rollback, and retraining are enforced security controls with defined SLAs — not optional operational conveniences.	CI/CD gates, signed artifact deployment, 15-minute rollback capability (monthly drill required)	SHALL

SECTION 4

REFERENCE ARCHITECTURE & AI CONTROL PLANE



4.1 AI Control Plane — Five-Layer Security Model

The AI Control Plane is the authoritative enforcement layer responsible for identity-bound execution, policy-constrained action authorization, and verifiable observability across AI-initiated operations spanning one or more trust domains. All AI-IRF™ v1.0 controls map to one or more Control Plane layers.

CP Layer	Layer Name	Scope	Primary Control IDs
CP-1	Identity & Credentials	Workload identity for agents, tools, MCP servers; credential lifecycle; SPIFFE/SPIRE federation; per-request authentication	ZTA-1, ZTA-2, MCP-1, APA-2
CP-2	Permissions & Scoping	Pre-execution policy evaluation; task-scoped least-privilege; HITL authority; auto-expiring tokens; permission audit trails	APA-1, APA-2, APA-3, APA-4, ZTA-4
CP-3	Orchestration & MCP	MCP endpoint security (mTLS, semantic inspection); A2A authorization; agent-to-agent trust boundaries; micro-segmentation	MCP-1, MCP-2, A2A-1, A2A-2, A2A-3, ZTA-3
CP-4	Validation Gates	Pre-execution policy checks for all consequential actions; manifest verification; RAG provenance validation; extraction detection	APA-1, BOM-4, MEX-1, MEX-2, MEX-3
CP-5	Observability & Audit	Behavioral baseline monitoring; LLM-as-Judge scoring; immutable telemetry; cryptographic log integrity; regulatory evidence packaging	MSR-1, MSR-2, REG-4, APA-4, MLC-4

4.2 Logical Architecture Components

Layer	Components	Control Plane Layer	Primary Control IDs
User / Entity	Human users, other agents, external systems	CP-1	ZTA-1, ZTA-2
Application Interface	Chat UI, API gateway, CLI, browser extensions	CP-1, CP-2	MCP-1, SAI-1
Agent Orchestration	Planner, executor, memory controller, tool router	CP-2, CP-3, CP-4	APA-1, APA-2, APA-3
Memory System	Session memory, persistent KB, vector store, graph DB	CP-4, CP-5	BOM-2, MSR-1
Retrieval (RAG)	Retriever, embedder, reranker, source connectors	CP-4	BOM-3, APA-1

LLM Core	Base model, fine-tuned adapter, inference endpoint	CP-4, CP-5	MEX-1, MLC-1, MSR-1
Tools / Plugins	MCP servers, APIs, code executors, web search	CP-3, CP-4	MCP-1, MCP-2, A2A-1
Data Sources	Training data, documents, external feeds	CP-4	BOM-1, BOM-2, BOM-3
Infrastructure	Compute, network, cloud IAM, orchestration layer	CP-1, CP-2, CP-3	ZTA-1, ZTA-3, ZTA-4

4.3 Zero Trust Architecture for AI (ZTAI) — Mandatory Controls

ZTAI Principle	AI-Specific Implementation	Normative Req.	Validation Method
Identity-First	Every agent, tool, MCP server, and A2A peer has a workload identity (SPIFFE, JWT). No anonymous inference or tool execution permitted.	SHALL	Monthly identity audit; automated orphaned-identity detection
Continuous Verification	Per-request authorization for tool calls; session timeouts ≤15 minutes; re-authentication for high-risk actions.	SHALL	Runtime enforcement logs; session duration monitoring
Micro-Segmentation	Inference endpoints, vector stores, training pipelines, and orchestration layers reside in separate trust zones with explicit allow-lists.	SHALL	Network flow review; segmentation policy-as-code validation
Least-Privilege Runtime	Agent tokens auto-expire after task completion; no standing privileges; permissions are task-scoped and time-bound.	SHALL	Token lifetime monitoring; permission grant/deny audit trails
Encrypt Everywhere	All MCP/A2A traffic uses TLS 1.3 with mutual authentication; data at rest encrypted with customer-managed keys.	SHALL	TLS configuration scans; key rotation compliance audits

SECTION 5

INCIDENT SEVERITY CLASSIFICATION



AI-IRF™ v1.0 defines four incident severity levels. Severity determines automated response thresholds, HITL escalation requirements, regulatory notification obligations, and evidence preservation priority. Severity SHALL be assigned within 15 minutes of incident declaration and SHALL be reviewed at each triage milestone.

P0
 CRITICAL

Physical harm potential; critical infrastructure AI compromise; multi-system agentic cascade; confirmed high-risk system failure under EU AI Act. Automated containment initiated immediately. Safety team notified in parallel. Legal and Compliance engaged within 15 minutes. The internal escalation clock starts at P0 declaration; any statutory notification clock SHALL be calculated separately under the applicable law, actor role, incident definition, awareness trigger, causal-link test, and operative application date. Two-person HITL approval required for all actions.

P1
 HIGH

Active data exfiltration exceeding notification thresholds; confirmed MCP compromise or lateral movement; model extraction confirmed; confirmed supply chain compromise affecting production. Automated containment active. CISO notified within 30 minutes. Legal and Compliance consulted. Regulatory notification decision tree executed within 1 hour.

P2
 MEDIUM

Shadow AI discovered with sensitive data exposure; suspected but unconfirmed model extraction; policy violation by AI agent without confirmed exfiltration; anomalous A2A traffic pattern. Rate-limiting and session monitoring activated. Security architecture team engaged. Compliance notified for awareness. 24-hour investigation window.

P3
 LOW

Behavioral baseline deviation within normal operational range; configuration drift detected; single low-confidence anomaly alert; training data integrity warning without confirmed poisoning. Logged and queued for analyst review. No automated containment unless confidence threshold exceeded. 72-hour investigation window.

5.1 Severity Escalation Rules

Severity SHALL be escalated (P3 → P2 → P1 → P0) when any of the following are confirmed: additional systems involved; physical harm pathway identified; regulatory notification threshold crossed; HITL approval requested for high-impact action; containment fails to reduce anomaly signal within response SLA.

Severity SHALL NOT be de-escalated without documented evidence of containment and root cause determination. De-escalation requires security architect review and CISO acknowledgement for P0 and P1 incidents.

SECTION 6

AI-ADAPTED INCIDENT RESPONSE LIFECYCLE



Phase	Traditional Alignment	AI-IRF™ v1.0 Augmentation	Time Target
0: Continuous Discovery	Identify	Runtime shadow AI detection, instance fingerprinting, manifest registry sync	Continuous
1: Preparation	Govern, Protect	Manifest signing, ZTAI segmentation, adaptive playbook pre-approval, HITL authority designation, team readiness drills	Pre-incident
2: Detection & Triage	Detect	Machine-speed behavioral baselines, LLM-as-Judge scoring, confidence-thresholded alerting, severity classification	< 2 min (P0/P1)
3: Machine-Speed Containment	Respond	Automated isolation < 500ms for high-confidence threats; HITL gates for critical actions	< 500ms auto; < 5 min HITL
4: Eradication	Respond	Signed rollback, poisoned data quarantine, manifest reverification, credential rotation	< 15 min for critical
5: Recovery	Recover	Canary deployment of clean model, embedding integrity check, adversarial regression testing	Per recovery plan
6: Post-Incident & Regulatory	Improve	Regulatory decision tree execution, structured notification submission, audit evidence packaging	Per applicable law; use jurisdiction- and incident-specific triggers and deadlines

Phase 0: Continuous Discovery

Control ID	Authoritative Domain 2 Requirement	Phase 0 Implementation Guidance	Phase 0 Success Metric
SAI-1	Deploy instance-aware discovery across network edges, SaaS connectors, endpoints, and browser environments.	Use egress/proxy AI fingerprinting, endpoint/EDR discovery, browser-extension inventory, and CASB/API connector discovery as applicable.	Discovery coverage and latency reported; target 100% of known discovery surfaces continuously monitored.
SAI-2	Inspect all AI data egress points for sensitive-data exposure using DLP controls.	Apply DLP inspection to external AI APIs, SaaS AI features, browser-mediated AI traffic, and other identified AI egress paths.	Representative egress-path test set passes DLP policy checks; leakage simulations generate expected detections.
SAI-3	Automatically classify discovered AI instances as sanctioned, shadow, or rogue within 1 hour; log tier, timestamp, and rationale in CMDB.	Use automated classification based on data sensitivity, ownership, authorization status, and risk context. Measure accuracy against a labelled review set.	Classification completed within 1 hour; where the 95% accuracy target is used, dataset, window, and exception treatment are documented.
SAI-4	Quarantine rogue AI instances and declare an incident within 15 minutes of classification; page the AI Security Lead; declare at P2 minimum via SOAR.	Automate quarantine/access restriction, SOC/SOAR declaration, paging, and evidence capture; test end-to-end timing.	Quarantine and incident declaration verified within 15 minutes in controlled tests.
SAI-5	Unauthorized AI deployments classified as rogue SHALL trigger an automated SOC alert with risk classification within 5 minutes of rogue designation.	Integrate rogue designation with SOC/SIEM alerting and include risk tier, asset identity, owner context where available, and timestamp.	SOC alert generated within 5 minutes of rogue designation; alert content and routing verified.

Phase 1: Preparation

Machine-speed containment (Phase 3) fails without pre-approved playbooks, pre-established HITL authority, and verified manifest signing infrastructure. Organizations **SHOULD** complete all Phase 1 controls before declaring AI-IRF™ v1.0 operational readiness.

Control ID	Preparation Control	Normative Level	Completion Criterion
PREP-1	Sign cryptographic manifests (AI-BOM) for all production AI systems before deployment; unsigned artifacts blocked by CI/CD gates	SHALL	100% manifest coverage on production systems
PREP-2	Implement ZTAI micro-segmentation isolating inference, vector storage, and orchestration layers with documented allow-lists	SHALL	Architecture review passed; segmentation policy-as-code validated
PREP-3	Pre-approve adaptive playbooks for all six priority incident types (see Section 9.1); obtain CISO sign-off	SHALL	All six playbooks signed, tested, and accessible in < 30 seconds
PREP-4	Designate and document HITL approval authority for each consequential action category; establish dual-control workflow	SHALL	HITL authority matrix documented and tested in tabletop exercise
PREP-5	Conduct AI-specific incident response drills quarterly including at least one machine-speed containment simulation	SHALL	Drill completion log; response time benchmarks recorded
PREP-6	Establish behavioral baselines for all production AI systems; sign and version baselines for audit integrity	SHALL	7-day rolling baseline established; baseline signing verified
PREP-7	Verify rollback capability for critical AI services; demonstrate < 15-minute rollback execution in drill	SHALL	Monthly rollback drill passed; rollback artifacts cryptographically signed

Phase 2: Machine-Speed Adaptive Detection

Detection Layer	Technology	Response Time	FP Tolerance	Auto-Action Threshold
Statistical anomaly	Token usage spikes, latency deviations, error rate changes	< 100ms	High — triggers LLM-J validation	N/A — escalates only
LLM-as-Judge (LLM-J)	Fine-tuned classifier on prompt/response pairs for malicious intent	< 500ms	Medium — analyst review for medium confidence	> 95% confidence
Semantic similarity	Embedding drift from baseline; query pattern clustering	< 1s	Low — auto-contain if > 95% confidence	> 95% confidence
Graph-based	Cross-session attack chain	5s–30s	Very low — HITL	Mandatory HITL

correlation	reconstruction; A2A lateral movement detection		review mandatory	
-------------	--	--	------------------	--

Mandatory: Every production AI system SHALL maintain a 7-day rolling behavioral baseline with automated alerting on $> 3\sigma$ deviation. Baselines SHALL be versioned and signed for audit integrity.

Phase 3: Machine-Speed Containment with HITL Gates

Incident Type	Auto Action (0–500ms)	HITL Threshold	HITL Action
Prompt injection	Isolate session, rate-limit to 1 req/min, capture context snapshot	High confidence OR P1/P0	Revoke API key, quarantine user account, notify app owner
Model extraction	Dynamic rate limiting, cryptographic output perturbation, session throttling	Pattern confirmed across sessions	Revoke access, initiate legal hold, preserve query logs
Poisoned RAG entry	Remove from vector index, quarantine source document, invalidate cache	Before reindexing	Human review of source; reindex from trusted source only
Compromised MCP server	Network isolate endpoint, revoke all credentials issued to server, notify connected agents	All destructive actions	Rotate workload identities in trust domain; forensic capture (two-person)
Agent privilege abuse	Auto-expire tokens, disable agent execution graph, freeze memory writes	Permission scope remediation	Security architecture review; permission scope remediation sign-off
Physical safety threat	Trigger fail-safe mode, isolate control plane, notify safety team	ALL actions — no exceptions	Joint safety-engineering dual approval required

Mandatory: All automated containment actions SHALL be logged with justification (confidence score, triggered rule, timestamp). HITL override SHALL require two-person approval with immutable audit trail. Evidence capture SHALL NOT be bypassed by any automated action.

Phase 6: Post-Incident & Regulatory Response

Regulatory Notification Decision Tree

STEP 1 — Is the organisation a provider of a high-risk AI system placed on the Union market, or a deployer to whom Article 73 applies mutatis mutandis, and does the event meet the Article 3(49) definition of a serious incident?

YES → Apply Article 73 timing: report immediately after establishing the required causal link or reasonable likelihood, and no later than 15 days after awareness; no later than two days for a widespread infringement or an Article 3(49)(b) serious incident; and no later than 10 days where a death occurs. Confirm the operative application date and any sectoral reporting route. An incomplete initial report may be followed by a complete report where necessary.

NO → Record the determination and proceed to Step 2.

STEP 2 — Was personal data (PII or sensitive categories) accessed, exfiltrated, or exposed?

YES → Notify supervisory authority (DPA) within 72 hours (GDPR Art. 33). Notify data subjects "without undue delay" if high risk to their rights.

NO → Proceed to Step 3.

STEP 3 — Does the incident affect critical infrastructure or essential services?

YES → CIRCIA: verify final-rule status before relying on a federal reporting deadline. The CIRCIA statute/NPRM framework provides for covered cyber-incident reporting within 72 hours after the entity reasonably believes the incident occurred and ransom-payment reporting within 24 hours after payment, once implementing obligations are effective. NIS2: early warning within 24 hours, incident notification within 72 hours, and final report generally within one month under Article 23, subject to national transposition.

NO → Proceed to Step 4.

STEP 4 — Did covered automated decision-making technology materially influence a consequential decision about a consumer (Colorado SB 26-189)?

YES → From 1 January 2027 (SB 26-189, which replaced SB 24-205), document the decision for post-adverse-outcome disclosure, data access and correction, and human-review rights; confirm current Attorney General rules. Engage compliance officer.

STEP 5 — Does the incident involve PHI or healthcare AI systems?

YES → HIPAA Breach Notification Rule: notify HHS and affected individuals within 60 days. If > 500 individuals, also notify prominent media in affected state.

STEP 6 — Does the incident involve AI-assisted financial transactions or AI capability disclosure (FINRA / SEC)?

YES → Engage securities counsel. Obligations are rule- and event-specific — for example FINRA Rule 4530 reporting for member firms, and Form 8-K Item 1.05 disclosure of material cybersecurity incidents for SEC registrants.

Mandatory Evidence Package for Regulators

Evidence Type	Content	Normative Req.
Immutable timeline	Signed timestamps from SIEM, orchestrator, safety systems, and HITL approval workflows; blockchain-anchored for P0 incidents	SHALL
Affected artifact manifest	Signed AI-BOM of all affected models, embeddings, data sources, and dependencies with provenance chain	SHALL
Telemetry sample	Sample of prompts/responses, tool calls, memory reads, retrieval results — automated PII redaction applied before submission	SHALL
Containment evidence	Automated action logs, confidence scores, triggered rules, HITL approval signatures, timestamps	SHALL
Root cause classification	Adversarial attack vs. structural permission failure vs. supply chain compromise; control gap mapping; remediation plan	SHALL
Regulatory mapping rationale	Jurisdictional determination with citation to applicable regulation and notification deadline calculation	SHALL

SECTION 7

CONTROL DOMAINS & TECHNICAL REQUIREMENTS



All controls use normative language: *SHALL* indicates a mandatory requirement. *SHOULD* indicates a recommended control where multiple valid implementations exist. *MUST* has been normalized to *SHALL* throughout this document for consistency. This Technical Report defines 79 controls across 19 domains, comprising the original 10 domains (Domains 1–10, 48 controls), three domains specified directly in this Technical Report (Domains 11–13: Detection Reliability, Agentic Threat Detection, Severity Classification — 8 controls), and six normative extension domains (Domains 14–19, drawn from ODA3-ECO-011/013/014/015 — 23 controls). This is the single authoritative control count for the AI-IRF™ v1.0 control scheme.

DOMAIN 1: MCP & A2A PROTOCOL SECURITY — CP-3

ID	Requirement	Normative	Validation Method
MCP-1	MCP endpoints SHALL require mutual TLS (mTLS) or workload identity federation. No unauthenticated endpoints permitted.	SHALL	Monthly port scan + configuration audit; automated certificate validation
MCP-2	All MCP message traffic SHALL be inspected at semantic level for injection patterns, malicious skill payloads, or protocol tampering.	SHALL	SIEM rule testing weekly; red-team semantic injection exercises
MCP-3	Unregistered MCP servers SHALL be automatically quarantined upon detection; notifications sent to orchestrator and platform owner within 5 minutes.	SHALL	Automated quarantine workflow testing; quarantine latency monitoring
MCP-4	MCP server manifests SHALL be registered with the agent orchestrator before endpoints go live; unregistered servers blocked at network layer.	SHALL	Orchestrator manifest registry audit; registration gate enforcement testing
A2A-1	Inter-agent authorization SHALL enforce least-privilege action scopes. Policy-as-code enforcement mandatory.	SHALL	Policy validation in CI/CD; runtime permission audit logs
A2A-2	All inter-agent messages SHALL be cryptographically signed (ECDSA) and integrity-verified before processing. Unsigned or tampered messages SHALL be rejected and logged with SIEM alert.	SHALL	Cryptographic message-integrity test: accept correctly signed messages; inject unsigned, incorrectly signed, and tampered messages; verify rejection, SIEM alerting, and signature-verification audit records.
A2A-3	A2A communication SHALL enforce per-agent-pair rate limits and anomaly detection to prevent agent loop amplification attacks. Circuit-breaker SHALL trigger on >3 anomalies per hour, freezing the amplifying agent and alerting SOC.	SHALL	Rate-limit and circuit-breaker test: exercise traffic at/below/above configured per-agent-pair limits; inject 3 then 4 qualifying anomalies within one hour; verify no premature trip, circuit-breaker activation at >3, agent freeze, SOC alert, and reset/recovery controls.

DOMAIN 2: SHADOW AI & CONTINUOUS VISIBILITY — CP-1, CP-5

ID	Requirement	Normative	Validation Method
SAI-1	Organizations SHALL deploy instance-aware discovery across network edges, SaaS connectors, endpoints, and browser environments.	SHALL	Weekly coverage report; automated discovery latency monitoring
SAI-2	DLP controls SHALL inspect all AI data egress points for sensitive data exposure.	SHALL	Monthly DLP policy testing; simulated data leakage drills
SAI-3	Shadow AI instances SHALL be automatically classified into risk tiers (sanctioned, shadow, rogue) within 1 hour of discovery. Classification logged to CMDB with tier assignment, timestamp, and rationale. Unclassified instances default to rogue tier.	SHALL	Workflow timing test and classification-quality audit: verify sanctioned/shadow/rogue tier assignment within 1 hour, CMDB logging, rationale capture, default-to-rogue behavior, and accuracy against a labelled sample.
SAI-4	Rogue AI instances SHALL trigger automated quarantine and incident declaration within 15 minutes of classification. AI Security Lead SHALL be paged automatically. Incident declared at P2 minimum via SOAR workflow.	SHALL	End-to-end containment test: introduce/simulate a rogue AI instance; verify quarantine, P2-or-higher incident declaration, AI Security Lead paging, evidence capture, and completion within 15 minutes.
SAI-5	Unauthorized AI deployments classified as rogue SHALL trigger automated SOC alert with risk classification within 5 minutes of rogue designation.	SHALL	SOC alerting workflow test; alert latency monitoring

DOMAIN 3: ADAPTIVE MACHINE-SPEED RESPONSE — CP-5

ID	Requirement	Normative	Validation Method
MSR-1	Response frameworks SHALL integrate LLM-as-Judge (LLM-J) metrics for real-time behavioral validation with confidence scoring.	SHALL	Detection latency SLA < 500ms; false positive rate monitoring
MSR-2	Playbooks SHALL support conditional branching based on live telemetry deltas with confidence-thresholded automation.	SHALL	Weekly playbook dry runs; purple-team branching logic tests
MSR-3	Behavioral baselines SHALL trigger automated containment for high-confidence attacks (confidence > 95%) without human delay.	SHALL	A/B testing with red team; containment efficacy metrics
MSR-4	HITL SHALL be required for actions with blast radius > 10 users, financial impact > \$10K, or safety-critical system involvement.	SHALL	Approval audit trail review; HITL compliance monitoring
MSR-5	MTTC (mean time to containment) for agentic threat events SHALL be 2 seconds or less.	SHALL	MTTC measurement across simulated agentic threat events

DOMAIN 4: MODEL EXTRACTION & DISTILLATION DEFENSES — CP-4

ID	Requirement	Normative	Validation Method
----	-------------	-----------	-------------------

		ve	
MEX-1	Behavioral distillation detection SHALL analyze query patterns for systematic extraction via embedding similarity tracking across sessions.	SHALL	Detection rule testing with emulated extraction campaigns
MEX-2	Dynamic rate limiting SHALL use per-user/session quotas with semantic clustering to detect sustained extraction patterns.	SHALL	Rate limit effectiveness review; extraction campaign simulation
MEX-3	Cryptographic output perturbation SHALL be applied for high-confidence suspected extraction attempts (> 85% similarity score across sessions).	SHALL	A/B testing detection rate; perturbation impact on legitimate use assessed
MEX-4	API key revocation workflow SHALL execute within 5 minutes of confirmed extraction detection. Legal hold preservation SHALL be initiated in parallel.	SHALL	Incident simulation quarterly; revocation latency monitoring

DOMAIN 5: AI-BOM & SUPPLY CHAIN INTEGRITY — CP-4

ID	Requirement	Normative	Validation Method
BOM-1	Every production AI system SHALL have a cryptographically signed manifest before deployment; unsigned artifacts blocked by CI/CD gates.	SHALL	Pre-deployment gate enforcement; signature validation audits
BOM-2	Training data and model weights SHALL be cryptographically hashed before ingestion; hashes SHALL be recorded and verified at load time.	SHALL	Integrity check on data load; hash mismatch alerting
BOM-3	External data sources SHALL undergo source authenticity checks including vendor attestation before integration.	SHALL	Quarterly vendor assessment; attestation validation testing
BOM-4	Incident playbooks SHALL include supply chain rollback and vendor notification SLA (≤ 24 hours from incident declaration).	SHALL	Tabletop exercise; vendor escalation path testing
BOM-5	Runtime BOM validation SHALL occur on every model load; mismatch between loaded model and registered BOM blocks inference.	SHALL	Deployment pipeline log showing BOM validation gate executed

DOMAIN 6: AGENTIC PERMISSION ARCHITECTURE (APA) — CP-2

ID	Requirement	Normative	Validation Method
APA-1	Pre-execution validation gates SHALL enforce policy checks before any consequential action. No bypass permitted.	SHALL	Policy enforcement audit; pre-execution bypass testing
APA-2	Agents SHALL operate under least-privilege, task-	SHALL	Token lifetime monitoring; permission

	scoped permissions with auto-expiration (≤ 15 minutes from task completion).		scope violation alerts
APA-3	HITL SHALL be required for: data export > 1K records, payment initiation, production delete, PII access, safety override.	SHALL	HITL approval logs; dual-control compliance monitoring
APA-4	All permission grants, denials, and overrides SHALL be logged with immutable timestamps and cryptographic signing. SIEM ingestion mandatory; retention minimum 3 years.	SHALL	Permission-decision evidence audit: sample grants, denials, and overrides; verify completeness, immutable timestamping, cryptographic signature/integrity verification, SIEM ingestion, tamper detection, retention configuration for at least 3 years, and retrieval of retained records.
APA-5	Tool use SHALL require a pre-execution approval gate, separate from the general APA-1 action validation gate, specific to tool invocation.	SHALL	Tool invocation gate configuration audit

DOMAIN 7: REGULATORY & COMPLIANCE OPERATIONS

ID	Requirement	Normative	Validation Method
REG-1	Incident playbooks SHALL include a regulatory notification decision tree mapping blast radius, data type, and AI risk tier to jurisdiction-specific deadlines.	SHALL	Playbook validation with legal counsel; deadline simulation testing
REG-2	Legal and Compliance SHALL be embedded as Consulted/Accountable roles in initial triage within 1 hour of incident declaration.	SHALL	RACI adherence audit; triage timeline review
REG-3	Organizations SHALL maintain incident-to-regulation mapping tables reviewed quarterly with compliance team.	SHALL	Quarterly review documentation; mapping accuracy validation
REG-4	Audit-ready documentation SHALL be preserved with cryptographic integrity for minimum 3 years (GDPR baseline); 5 years for critical infrastructure (NIS2); 10 years for health AI (HIPAA/HITECH).	SHALL	Retention policy audit; evidence chain integrity verification

DOMAIN 8: PHYSICAL AI & SAFETY INTEGRATION — CP-2, CP-5

ID	Requirement	Normative	Validation Method
PHY-1	Separate incident playbooks SHALL exist for physical-harm scenarios with explicit escalation paths to safety teams.	SHALL	Playbook library review; safety scenario coverage audit
PHY-2	Safety and cybersecurity teams SHALL operate in integrated incident command structures with joint tabletop exercises quarterly.	SHALL	Exercise attendance logs; joint command effectiveness metrics
PHY-3	Systems SHALL mandate fail-safe fallback modes upon cyber anomaly detection or attestation failure.	SHALL	Safety validation testing; fail-safe activation drills

	Human-safe-stop states are the default.		
PHY-4	Hardware-rooted trust SHALL default to human oversight or safe-stop states if cryptographic attestation fails. No autonomous action permitted post-attestation failure.	SHALL	Architecture review; attestation failure simulation testing

DOMAIN 9: MODEL LIFECYCLE SECURITY — CP-4, CP-5

ID	Requirement	Normative	Validation Method
MLC-1	Every production model SHALL have a unique immutable identifier and cryptographic signature; unsigned models blocked from deployment.	SHALL	CI/CD gate enforcement; signature validation at inference time
MLC-2	Automated rollback capability SHALL execute within 15 minutes of compromise detection for critical services. Monthly drills required.	SHALL	Monthly recovery drills with documented results; rollback latency monitoring
MLC-3	Recovery playbooks SHALL include rollback validation gates confirming integrity before traffic is restored.	SHALL	Post-recovery audit; validation gate compliance monitoring
MLC-4	Model version changes SHALL be logged in immutable audit trail with signer identity, timestamp, and approval record.	SHALL	Audit log review; change management compliance validation
MLC-5	Drift detection thresholds SHALL trigger a dual-approval retraining workflow; retraining without dual approval is blocked.	SHALL	Drift-triggered retraining workflow audit

DOMAIN 10: ZERO TRUST ARCHITECTURE FOR AI (ZTAI) — CP-1, CP-2, CP-3

ID	Requirement	Normative	Validation Method
ZTA-1	AI infrastructure SHALL adopt explicit zero-trust design with documented architecture reviewed at least annually.	SHALL	Architecture review; ZTAI compliance assessment
ZTA-2	Continuous verification SHALL require re-authentication every 15 minutes or per high-risk action; no persistent sessions permitted.	SHALL	Session duration audit; re-authentication enforcement testing
ZTA-3	Micro-segmentation SHALL isolate inference, vector storage, and orchestration layers with explicit allow-lists enforced in policy-as-code.	SHALL	Network flow analysis; segmentation policy compliance scans
ZTA-4	Runtime least-privilege enforcement SHALL include per-request authorization for tool execution with full audit trail.	SHALL	Policy decision point logs; permission grant/deny audit trails

DOMAIN 11: DETECTION RELIABILITY**PROVENANCE NOTE**

Specified directly in this Technical Report. No separate ODA3-ECO addendum applies to this domain.

ID	Requirement	Normative	Validation Method
REL-1	Auto-containment SHALL require ≥ 3 ensemble signals, agreement ≥ 0.7	SHALL	Detection architecture documentation showing ≥ 3 independent signal sources
REL-2	Confidence scores SHALL decay 10% per consecutive similar alert	SHALL	Decay function configuration and rolling window parameters
REL-3	Circuit-breaker SHALL disable automation after 3 FP within 1h	SHALL	Circuit breaker configuration showing 3-FP/1-hour trip threshold

DOMAIN 12: AGENTIC THREAT DETECTION**PROVENANCE NOTE**

Specified directly in this Technical Report. APT controls form the threat-detection layer that triggers the APA permission gates (APA-1, APA-5): APA governs the permission architecture, while APT detects when that architecture is being abused or attacked.

ID	Requirement	Normative	Validation Method
APT-1	Prompt injection detection $\geq 90\%$ across all channels	SHALL	Injection detection test report showing $\geq 90\%$ detection rate across channel types
APT-2	Tool abuse prevention via pre-execution gate on every tool call	SHALL	Tool abuse pattern library documentation
APT-3	Memory poisoning detection with automated quarantine $\leq 60s$	SHALL	Memory write scanning configuration and pattern library

DOMAIN 13: SEVERITY CLASSIFICATION**PROVENANCE NOTE**

Specified directly in this Technical Report. Severity levels follow the P0–P3 scheme defined in Section 5. The scoring structure is aligned with the UAIF™ severity presentation model for cross-framework consistency.

ID	Requirement	Normative	Validation Method
SEV-CLS	Deterministic severity scoring algorithm implemented and versioned	SHALL	Severity scoring algorithm documentation with version history
SEV-RPT	Reportability thresholds evaluated deterministically	SHALL	Reportability threshold table covering all applicable jurisdictions and regulations

DOMAIN 14: CONTENT SAFETY

PROVENANCE NOTE

CSS-1, CSS-3 and CSS-5 are fully specified in ODA3-ECO-011. For CSS-2, CSS-4, CSS-6 and CSS-7, ODA3-ECO-011 provides the control name, tier and regulatory mapping; the requirements below state the control obligation, with fuller implementation guidance maintained separately.

ID	Requirement	Normative	Validation Method
CSS-1	Harmful content detection pipeline SHALL achieve >99.9% block rate for CSAM (zero-tolerance); alert and log all blocked outputs.	SHALL	See ODA3-ECO-011 full specification
CSS-2	Maintain a sufficiently detailed summary of content used to train the GPAI model, published per EU AI Act Art. 53(1)(d) and the AI Office template. Implement output filtering to detect verbatim or near-verbatim reproduction of training content.	SHALL	Verify publication of the required training-content summary and test output filtering with a representative corpus for verbatim/near-verbatim reproduction; retain threshold, exception, and review evidence.
CSS-3	AI-generated synthetic media SHALL carry machine-readable provenance marker (C2PA recommended) and human-facing disclosure label.	SHALL	See ODA3-ECO-011 full specification
CSS-4	Design review SHALL confirm the AI system does not employ subliminal techniques (content presented below the threshold of conscious perception) or techniques designed to exploit known vulnerabilities	SHALL	Design review documentation confirming absence of subliminal/exploitative techniques
CSS-5	Monthly statistical analysis of AI outputs across protected demographic groups SHALL be conducted; cumulative_bias_index > 0.05 triggers Chronic Harm Feedback Loop.	SHALL	See ODA3-ECO-011 full specification
CSS-6	Conduct a quarterly Demographic Impact Assessment (DIA) for all high-risk AI systems. The DIA SHALL evaluate: demographic parity difference (<0.10 threshold), equalised odds difference (<0.10 threshold).	SHALL	Quarterly DIA reports showing demographic parity difference and equalised odds difference results
CSS-7	Tier 1 + documented vulnerable population impact assessment, conducted alongside CSS-6's DIA, specifically evaluating outcomes for identified	SHALL	Vulnerable population identification and use-case mapping documentation

vulnerable groups rather than aggregate demographic categories.

DOMAIN 15: FEDERATED LEARNING SECURITY

ID	Requirement	Normative	Validation Method
FL-1	Federated clients authenticated via mutual TLS with client-specific certificates; unregistered/revoked clients rejected.	SHALL	Client certificate registry; update acceptance/rejection logs (ODA3-ECO-013)
FL-2	Byzantine-robust aggregation (FedAvg with clipping; Krum/Bulyan for high-threat); gradients >3 sigma from epoch mean flagged; clients with anomaly rate >5% rejected.	SHALL	Gradient norm distribution logs; anomaly flag rate (ODA3-ECO-013)
FL-3	Differential privacy (DP-SGD) applied; epsilon <=1 for high-sensitivity data, epsilon <=8 maximum ceiling for lower-sensitivity data; epsilon=10 not recommended for production.	SHALL	DP configuration audit; privacy budget consumption log (ODA3-ECO-013)
FL-4	Global model signed with ECDSA (or CRYSTALS-Dilithium per PQC-1) after each aggregation round; signed checkpoint stored in AI-BOM registry.	SHALL	Signed model checkpoint registry; aggregation audit log (ODA3-ECO-013)

DOMAIN 16: MODEL MERGE DEFENSE

ID	Requirement	Normative	Validation Method
MMD-1	Pre-merge source model vetting: AI-BOM provenance, absence of known malicious fine-tuning datasets, training data manifest completeness verified; incomplete provenance rejected.	SHALL	Pre-merge vetting checklist; source model AI-BOM records (ODA3-ECO-013)
MMD-2	Post-merge backdoor scan (Neural Cleanse or ABS); semantic consistency testing across 500+ prompt templates; embedding space anomaly analysis.	SHALL	Backdoor scan results with tool version; anomaly cluster analysis (ODA3-ECO-013)
MMD-3	Model merge operations performed in isolated compute environment with no internet access; all artifacts hash-verified entering and leaving.	SHALL	Merge environment isolation configuration; artifact hash verification log (ODA3-ECO-013)

DOMAIN 17: QUANTIZATION SECURITY

ID	Requirement	Normative	Validation Method
----	-------------	-----------	-------------------

QBS-1	Pre-quantization behavioural baseline established on 1,000+ representative inputs; stored as signed artefact in AI-BOM.	SHALL	Baseline output corpus; signed baseline artefact (ODA3-ECO-013)
QBS-2	Post-quantization differential analysis against baseline; divergence >5% flagged; divergence >10% rejects model.	SHALL	Differential analysis report; divergence rate by input category (ODA3-ECO-013)
QBS-3	Only trusted, audited quantization toolchains used (ONNX Runtime, TensorRT, PyTorch quantization); toolchain integrity hash-verified; third-party pre-quantized models screened via QBS-1/QBS-2.	SHALL	Toolchain hash verification log (ODA3-ECO-013)

DOMAIN 18: HALLUCINATION DETECTION

PROVENANCE NOTE

HAL-1 to HAL-5 are specified in ODA3-ECO-014; all five are part of the AI-IRF™ v1.0 control scheme.

ID	Requirement	Normative	Validation Method
HAL-1	RAG factual grounding verification: all factual claims grounded in retrieved source documents; citation extraction and source verification; ungrounded claims flagged for human review.	SHALL	D2-CTL-04 alignment evidence (ODA3-ECO-014)
HAL-2	Model confidence scoring implemented; outputs below 0.7 confidence for consequential decisions require human review before action; uncertainty surfaced to end users.	SHALL	D5-CTL-03 alignment evidence (ODA3-ECO-014)
HAL-3	For high-stakes outputs (medical, legal, financial, safety-critical): parallel inference with second model or rule-based verifier; inconsistencies >10% divergence flagged for human review.	SHALL	D2-CTL-06 alignment evidence (ODA3-ECO-014)
HAL-4	Hallucination rate tracked via human correction feedback loop; alert at >3% of evaluated outputs; escalate to AI-IRF corrective action at >5% (trust_erosion_rate threshold).	SHALL	UAIF trust_erosion_rate field evidence (ODA3-ECO-014)
HAL-5	For regulated domains (medical diagnosis, legal analysis, financial advice): external fact-checking API or knowledge graph verification integrated for domain-specific claims.	SHALL	D6-CTL-03 Model Card alignment evidence (ODA3-ECO-014)

DOMAIN 19: POST-QUANTUM CRYPTOGRAPHY

PROVENANCE NOTE

PQC-1 operationalises the phased migration schedule established in ODA3-ECO-015 (Alignment Note) across four existing controls (BOM-4, MCP-1, A2A-2, REG-4). This control does not mandate immediate PQC implementation beyond what ECO-015 itself specifies — it mandates the readiness assessment and adherence to the documented timeline (Q1 2027 / Q3 2027 / Q3 2028).

ID	Requirement	Normative	Validation Method
PQC-1	PQC readiness assessment SHALL be completed covering BOM-4, MCP-1, A2A-2, and REG-4 cryptographic use cases. Migration to CRYSTALS-Dilithium (BOM-4, A2A-2) at Tier 3 from Q1 2027; hybrid Kyber-1024 (MCP-1) from Q3 2027; hybrid ECDSA+Dilithium (REG-4, all tiers) from Q3 2027.	SHALL	PQC readiness assessment document; migration plan with dated milestones (ODA3-ECO-015)

Telemetry Integration Requirement: AI-IRF™ v1.0 requires SIEM/SOAR/XDR integration for all control domains with structured event schemas. Telemetry retention SHALL support forensic reconstruction and regulatory audit (minimum 3 years; see REG-4 for sector-specific requirements).

SECTION 8

REGULATORY & COMPLIANCE OPERATIONS



8.1 Comprehensive Regulatory Mapping

Regulation	Applicable AI Incident Types	Timing / Operational Note	Compliance / Enforcement Note	AI-IRF™ v1.0 Controls
EU AI Act Art. 73	A serious incident meeting Article 3(49), within Article 73 scope; malfunction, safety event, or near-miss is not automatically reportable	Report immediately after the causal-link / reasonable-likelihood threshold is established; outer limits: 15 days generally, 2 days for specified widespread/Article 3(49)(b) cases, 10 days where death occurs.	Article 73 reporting does not itself create a standalone penalty tariff; enforcement depends on the breached AI Act obligation and applicable Article 99 / Member-State rules.	REG-1, REG-2, PHY-2
GDPR Art. 33–34	Personal data breach from AI inference, training data exposure, embedding exfiltration	Notify the competent supervisory authority within 72 hours where Article 33 applies; communicate to affected data subjects without undue delay where Article 34 high-risk threshold is met.	Supervisory measures and administrative fines depend on the infringed GDPR provisions and case facts; verify current Article 83 applicability.	REG-3, BOM-2, SAI-2
NIS2 Directive	Critical infrastructure AI incident, supply chain compromise affecting essential services	Article 23: early warning within 24 hours; incident notification within 72 hours; final report generally within one month, subject to national transposition and competent-authority rules.	Enforcement and fine ceilings depend on entity class and Member-State transposition; verify the applicable national law.	REG-4, ZTA-1, BOM-4
Colorado AI Act	Covered automated decision-making technology used to materially influence consequential decisions; not a general cyber-incident reporting regime.	Colorado SB26-189 requirements begin 1 January 2027. Post-adverse-outcome disclosure, data-access/correction, and human-review rights may apply; confirm operative rules at the time of use.	Enforced through the Colorado Consumer Protection Act as specified by the enacted law; cure/enforcement details are time- and fact-dependent.	REG-1 with impact assessment, APA-3
CIRCA (US)	Critical infrastructure AI compromise (covered entity)	Check final-rule effectiveness before treating reporting as operative. Statutory/NPRM framework: covered cyber incident within 72 hours after reasonable belief the incident occurred; ransom payment within 24 hours after payment.	CISA enforcement mechanisms depend on the implementing rule and statutory process; do not treat this table as a substitute for current rule review.	REG-4, ZTA-1, PHY-3
HIPAA / HITECH	PHI exposure via AI inference pipeline, training data, RAG retrieval, or agent action	60 days — HHS and affected individuals; media notification if > 500 in state	Civil monetary penalties are tiered and subject to statutory/regulatory adjustment; verify current HHS amounts and breach-notification applicability.	REG-1, REG-4, BOM-2, SAI-2
FINRA / SEC	AI-assisted transaction anomaly, unauthorized algorithmic action, AI capability disclosure failure	Rule- and event-specific reporting/escalation requirements apply; engage securities counsel/compliance rather than applying a single AI-specific deadline.	Enforcement consequences are rule- and fact-specific; verify the applicable SEC/FINRA obligation and current guidance.	REG-1, REG-2, APA-3, MSR-4
ISO/IEC 42001	AI-management-system incidents within an organisation operating an ISO/IEC 42001 AIMS.	Follow the organisation's AIMS incident, corrective-action, and management-review procedures; ISO/IEC 42001 is not a statutory incident-notification regime.	Any certificate action is determined by the applicable certification body and scheme; this report does not confer or operate ODA3 certification.	All controls; audit-ready documentation

8.2 Retention Requirements by Sector

Sector / Regulation	Minimum Retention	Notes
General (GDPR baseline)	3 years minimum	All AI incidents involving EU data subjects

Critical infrastructure (NIS2)	5 years	Essential services and critical infrastructure operators
Health AI (HIPAA / HITECH)	10 years	AI systems touching PHI or clinical decision support
Financial AI (FINRA / SEC)	6 years (Books and Records Rule 17a-4)	Trade records, AI decision logs, capability disclosures
Criminal investigation	Indefinite — legal hold	All evidence preservation obligations supersede standard retention

8.3 Ethical Guardrails

Principle	Requirement	Normative Level
Proportional Response	Automated containment SHALL NOT exceed the blast radius necessary to contain the identified threat. Disproportionate containment is a governance failure.	SHALL
Human Accountability	HITL decisions SHALL be attributed to specific individuals with cryptographic sign-off. No anonymous approvals permitted.	SHALL
Non-Discrimination	Incident classification SHALL NOT use protected characteristics (race, religion, gender, national origin, etc.) as classification factors.	SHALL
Transparency	Automated containment decisions SHALL be explainable: the triggering rule, confidence score, and evidence basis SHOULD be available for audit and appeal within 24 hours.	SHALL
Beneficence	Response actions SHALL prioritize minimizing harm to individuals over operational continuity when these goals conflict.	SHALL

SECTION 9

NEXT-GENERATION PLAYBOOK FRAMEWORK



9.1 Adaptive Playbook Library — Q2 2026 Priority Incidents

Playbook	Trigger	Branch Logic	HITL Required?	Escalation Path
PB-Agent-Privilege-Abuse	Tool call outside permission scope; anomalous write/export pattern	High confidence (>95%): auto-disable agent, revoke tokens. Medium: rate-limit + request review. Low: log + baseline update	Payment initiation, data export > 1K records, production delete, PII access	Service owner → IAM team → Compliance officer
PB-MCP-Compromise	Unauthenticated MCP request; malicious skill payload detected; semantic anomaly in A2A traffic	Immediate: isolate broker, revoke credentials, notify connected agents, forensic snapshot	All destructive actions require two-person approval	Platform owner → Vendor security → SOC lead
PB-Model-Extraction	Query similarity > 0.85 across sessions + sustained high volume + embedding convergence	Dynamic throttling → output perturbation → session isolation → legal hold if pattern confirmed	API key revocation, legal hold initiation	Abuse desk → SOC → Product owner → Legal
PB-Poisoned-RAG	Retrieved content fails provenance check; embedding outlier + user flag; index integrity mismatch	Auto-remove from vector index; quarantine source; invalidate cache	Human review before reindexing; source validation required	Data owner → Platform team → Supplier security
PB-Shadow-AI-Discovery	Unregistered inference endpoint detected; sensitive data egress via unsanctioned AI	Quarantine endpoint (block or rate-limit); notify administrator; preserve evidence	Approval required to allowlist; data exposure assessment	App owner → DLP team → Compliance
PB-Physical-Safety-Trigger	Combined cyber anomaly + safety sensor alert; unsafe actuator command detected	Emergency stop → fail-safe mode → isolate control plane → notify safety team	Override requires joint safety-engineering dual approval	Safety commander → Operations → Regulator (if required)

9.2 Playbook Integration Requirements

All playbooks SHALL expose REST APIs or MCP endpoints for SOAR integration with standardized input/output schemas.

Automation Safeguard	Requirement	Normative
----------------------	-------------	-----------

Evidence preservation	Automated containment SHALL never bypass evidence capture; all actions logged with immutable timestamps before execution	SHALL
AI recommendation validation	AI-generated recommendations for containment SHALL be validated by LLM-J scoring before executing high-impact actions	SHALL
Rollback signing	Rollback procedures SHALL be cryptographically signed and tested monthly against production-equivalent environments	SHALL
Playbook version control	All playbooks SHALL be version-controlled with signed commits; unauthorized modifications blocked by CI/CD gates	SHALL

IMPLEMENTATION ROADMAP & MATURITY MODEL

10.1 Phased Adoption Roadmap

Phase	Timeline	Priority Gaps	Key Activities	Success Criteria
1: Critical Foundations	0–60 days	G1, G2, G6	Deploy continuous AI discovery. Enforce mTLS for all MCP endpoints. Implement pre-execution validation gates. Register HITL approval authority. Complete PREP-1 through PREP-7.	100% AI instances discovered < 24h. All MCP endpoints authenticated. Pre-execution policy blocks validated. Phase 1 preparation controls complete.
2: Regulatory Readiness	60–120 days	G7, G3, G5	Embed regulatory decision trees. Deploy LLM-as-Judge detection. Sign manifests for top 20 production models. Tabletop exercises with Legal embedded.	Documented EU AI Act operational-readiness posture established against the applicable actor role, system category, current application timetable, and unresolved dependencies; this criterion does not assert legal compliance. Auto-containment latency < 500ms. Manifest signing rate 100% for critical models.
3: Advanced Capabilities	Q3–Q4 2026	G4, G9, G10	Implement behavioral distillation detection. Achieve < 15-min automated rollback. Complete ZTAI architecture documentation. Integrate physical AI safety playbooks.	Extraction detection rate > 90%. Recovery drill passed with < 15-min rollback. ZTAI architecture review passed.
4: Autonomous Security	2027	G8, advanced automation	Full cyber-safety integrated incident command. Autonomous HITL exception handling with formal verification. Cross-organizational threat intelligence sharing.	< 2-min safety escalation verified in drill. Regulatory audit with zero critical findings.

10.2 Capability Maturity Model

Level	Name	Key Capabilities	MTTD Target	MTTC Target	Auto-Containment
Level 1	Legacy Baseline	Static playbooks; inventory-centric visibility; human-speed triage	> 60 min	> 4 hours	< 10%
Level 2	Managed	Shadow AI discovery; basic MCP authentication; pre-execution gates	< 15 min	< 60 min	40%

Level 3	Advanced	LLM-J detection; automated containment (< 500ms); signed manifests; regulatory DTs	< 2 min	< 10 min	> 80%
Level 4	Autonomous	Predictive detection; autonomous HITL exception handling; cyber-safety convergence	< 30 sec	< 2 min	> 95%

10.3 KPI Dashboard

KPI	Target (Level 3)	Measurement Method	Frequency
Mean Time to Detect (MTTD)	< 2 minutes	SIEM detection-to-alert time; LLM-J scoring latency	Real-time dashboard
Mean Time to Contain (MTTC)	< 10 minutes	Playbook execution logs; containment action timestamps	Post-incident report
Automated containment rate	> 80% of incidents	IR records; automation trigger vs. HITL override ratio	Monthly review
False positive rate	< 5% for auto-containment	Post-incident review; analyst validation of automated actions	Quarterly audit
Regulatory submission accuracy	100% within deadline	Compliance audit; notification receipt verification	Per-incident
Manifest coverage	100% production models	CI/CD gate logs; signature validation reports	Weekly
Shadow AI discovery latency	< 24h from first activity	Discovery tool reports; instance registration timestamps	Daily
MCP endpoint authentication	100% mTLS	Port scan + configuration audit; certificate validation	Monthly
Rollback execution time	< 15 min for critical services	Recovery drill logs; rollback completion timestamps	Monthly drill
HITL compliance rate	100% for high-impact actions	Approval audit trail; dual-control verification logs	Per-incident

SECTION 11

NOTABLY ABSENT — FRAMEWORK SCOPE LIMITATIONS

This section documents what AI-IRF™ v1.0 explicitly does not cover and why. Documenting scope limitations is as important as documenting scope inclusions. It prevents threat inflation, sets accurate expectations for practitioners, and identifies areas requiring complementary frameworks or future research.

Out-of-Scope Area	Rationale	Complementary Framework
AI system design security (Secure SDLC for AI)	Pre-deployment design security requires a separate framework; incident response assumes production deployment	OWASP ASVS, NIST SSDF, ISO/IEC 27034
Broad algorithmic fairness remediation and model-design governance	AI-IRF™ includes incident-relevant bias monitoring and impact-assessment controls (for example CSS-5/CSS-6), but does not replace end-to-end fairness engineering, model-design remediation, or governance programmes.	ISO/IEC 42001, EU AI Act Art. 9, NIST AI RMF GOV function
Federated-learning platform design beyond the defined incident-response controls	Domain 15 (FL-1–FL-4) is in scope for AI-IRF™ v1.0. Broader federated-learning architecture, privacy engineering, and algorithm selection beyond those controls remain outside this Technical Report.	AI-IRF™ Domain 15; ODA3-ECO-013; specialist FL/privacy guidance
Content-authenticity governance beyond AI-IRF™ synthetic-media controls	AI-IRF™ includes CSS-3 provenance/disclosure requirements. Broader content-authenticity ecosystems, media policy, and provenance governance outside incident-response use remain separate.	AI-IRF™ Domain 14; C2PA; applicable transparency law
Enterprise-wide post-quantum migration beyond defined AI-IRF™ cryptographic use cases	Domain 19 (PQC-1) is in scope for identified AI-IRF™ cryptographic use cases. Organisation-wide PQC programme design and migration outside those use cases remain separate.	AI-IRF™ Domain 19; ODA3-ECO-015; NIST/CISA PQC guidance
AI model interpretability and explainability	Explainability is a design and regulatory compliance requirement; AI-IRF™ v1.0 requires explainability of containment decisions, not of model internals	EU AI Act Art. 13, NIST AI RMF MANAGE
Non-AI data breaches where AI is incidental	If AI is not a causal factor in the breach, NIST SP 800-61r3 and GDPR breach response frameworks apply	NIST SP 800-61r3, GDPR Art. 33–34
Criminal investigation and digital forensics standards	Chain of custody for criminal proceedings requires jurisdiction-specific forensic standards beyond this framework's scope	ACPO Digital Evidence principles, ISO/IEC 27042
Certification / assurance service operation	This Technical Report defines controls and validation guidance. It does not represent that an ODA3 certification, accreditation, audit, or independent assurance service is currently operating, and use of the framework does not confer a certification or compatibility mark.	GEL v1.0; applicable ODA3 trademark/conformance governance

SECTION 12

STANDARDS CROSSWALK



Note: Identifiers verified on 4 October 2026 against NIST AI 100-1 (AI RMF 1.0), ISO/IEC 42001:2023 clause and Annex A numbering, the MITRE ATLAS matrix, the OWASP Top 10 for LLM Applications (2023 edition) and the OWASP Top 10 for Agentic Applications 2026 (ASI01–ASI10). Mappings indicate topical correspondence, not equivalence or conformance. ATLAS technique names change periodically; re-validate at atlas.mitre.org before operational use.

Control Domain	NIST AI RMF 1.0	ISO/IEC 42001:2023	MITRE ATLAS	OWASP LLM Top 10 (2023)	OWASP Agentic 2026
MCP/A2A Security	MANAGE 3.1; MEASURE 2.7	A.10.3 (Suppliers); A.6.2.6 (AI system operation and monitoring)	AML.T0051 (LLM Prompt Injection); AML.T0053 (LLM Plugin Compromise)	LLM07: Insecure Plugin Design	ASI02: Tool Misuse and Exploitation; ASI07: Insecure Inter-Agent Communication
Shadow AI Visibility	GOVERN 1.6	A.4.2 (Resource documentation)	No direct technique (defensive inventory control)	N/A	No direct equivalent
Machine-Speed Response	MANAGE 2.4; MANAGE 4.1	8.1 (Operational planning and control); A.6.2.6	N/A (response, not attack)	LLM08: Excessive Agency	ASI08: Cascading Failures; ASI10: Rogue Agents
Model Extraction Defense	MEASURE 2.7	8.2 (AI risk assessment)	AML.T0040 (AI Model Inference API Access); AML.T0005 (Create Proxy AI Model)	LLM10: Model Theft	N/A (model-level threat)
AI-BOM & Supply Chain	GOVERN 6.1; MANAGE 3.1	A.7.5 (Data provenance); A.10.3 (Suppliers)	AML.T0010 (AI Supply Chain Compromise)	LLM05: Supply Chain Vulnerabilities	ASI04: Agentic Supply Chain Vulnerabilities
Agentic Permissions	MAP 3.5	A.9.2 (Processes for responsible use of AI systems)	AML.T0053 (LLM Plugin Compromise)	LLM08: Excessive Agency	ASI03: Identity and Privilege Abuse
Physical AI Safety	MEASURE 2.6	A.5.2 (AI system impact assessment process)	AML.T0043 (Craft Adversarial Data)	N/A	N/A

Regulatory Operations	GOVERN 1.1; MANAGE 4.3	A.8.4 (Communication of incidents)	N/A	N/A	N/A
Model Lifecycle	MANAGE 4.1	A.6.2.5 (AI system deployment); A.6.2.8 (AI system recording of event logs)	AML.T0020 (Poison Training Data)	LLM03: Training Data Poisoning; LLM05: Supply Chain Vulnerabilities	ASI04: Agentic Supply Chain Vulnerabilities
Zero Trust (ZTAI)	MEASURE 2.7	8.3 (AI risk treatment)	AML.T0016 (Obtain Capabilities)	LLM08: Excessive Agency	ASI03: Identity and Privilege Abuse; ASI07: Insecure Inter-Agent Communication

SECTION 13

GAP RESOLUTION MATRIX



Gap ID	Gap Description	Root Cause in Existing Frameworks	AI-IRF™ v1.0 Resolution	Validation Method
G1	MCP & A2A protocol security blind spot	Treated as observable telemetry, not primary attack surface	MCP-1 through MCP-4, A2A-1 through A2A-3: mTLS, semantic inspection, peer authorization, broker isolation, manifest registration	Red-team MCP/A2A compromise exercise; denied-lateral-movement testing
G2	Shadow AI & visibility gaps	Inventory-centric model assumes visibility rather than mandating discovery	SAI-1 through SAI-5: Continuous discovery, DLP at egress, risk-tier classification, automated quarantine, SOC alerting on rogue designation	Discovery coverage KPI reporting; unauthorized AI detection testing
G3	Static vs. adaptive response	Human-speed bounded playbooks; static IOC matching; no machine-speed automation	MSR-1 through MSR-5: LLM-J scoring, conditional branching, confidence-thresholded automation, HITL thresholds, agentic MTTC target	Purple-team timing tests on containment latency; FP rate monitoring
G4	Model extraction & distillation — not operationalized	Threat acknowledged but no detection methodology or response workflow	MEX-1 through MEX-4: Behavioral extraction analytics, dynamic rate limiting, output perturbation, automated key revocation	Simulated distillation campaign detection rate; extraction containment efficacy
G5	AI-BOM & supply chain integrity	Traditional SBOM mindset; no cryptographic provenance for AI artifacts	BOM-1 through BOM-5: Signed manifest, hashed artifacts, supplier attestation, supply chain rollback SLA, runtime BOM validation	Provenance audit drills; rollback reconstruction testing
G6	Agentic permission architecture	Assumes adversarial intent only; missing pre-execution validation gates	APA-1 through APA-5: Pre-execution policy checks, task-scoped permissions, auto-expiration, HITL for high-impact, tool-invocation gate	Agent overreach simulation testing; policy-block evidence collection
G7	Regulatory mapping not operational	Regulation-aware but no decision logic, deadlines, or embedded compliance roles	REG-1 through REG-4: Jurisdictional decision tree, embedded legal/compliance roles, audit-ready evidence packages	Tabletop exercises with mapped deadlines; RACI adherence audits
G8	Physical AI & safety-security excluded	Entirely absent or explicitly excluded from all reviewed frameworks	PHY-1 through PHY-4: Joint cyber-safety command, fail-safe fallback, hardware-rooted trust	Safety incident simulation drills; fail-safe activation testing

G9	Model lifecycle as operational — not security control	Treated as operational process, not enforceable security control with SLAs	MLC-1 through MLC-5: Immutable signed lifecycle, 15-min rollback SLA, integrity verification gates, dual-approval drift retraining	Integrity-verification testing; rollback-time drills
G10	Zero trust not systematized for AI	Individual controls present but not unified under zero-trust paradigm	ZTA-1 through ZTA-4: Identity-first, continuous verification, micro-segmentation, runtime least-privilege	Architecture review against ZTAI principles; control enforcement validation
G11	Detection reliability not engineered	Automated containment specified without ensemble validation or false-positive cascade safeguards	REL-1 through REL-3: ≥3-signal ensemble agreement for auto-containment, confidence decay on repeated alerts, false-positive circuit breaker	Detection architecture review; circuit-breaker trip testing
G12	Agentic threats not detected as distinct events	Prompt injection, tool abuse and memory poisoning treated as model risks rather than detection requirements	APT-1 through APT-3: ≥90% prompt-injection detection across channels, pre-execution gate on every tool call, memory-poisoning quarantine ≤60s	Injection detection test report; memory-write scanning tests
G13	Severity and reportability left to judgement	Severity and reportability decided case by case under time pressure	SEV-CLS and SEV-RPT: Versioned deterministic severity scoring; deterministic reportability thresholds	Scoring algorithm version history; reportability threshold table review
G14	Content safety and societal harm outside IR	Content-safety and societal-harm failures handled outside security incident response	CSS-1 through CSS-7: Harmful-content blocking, training-content summary, synthetic-media provenance, manipulative-technique design review, bias monitoring, demographic and vulnerable-population impact assessment	Domain 14 validation methods; ODA3-ECO-011
G15	Federated learning security not addressed	Distributed training not covered by incident response guidance	FL-1 through FL-4: Client mTLS authentication, Byzantine-robust aggregation, differential privacy, signed global checkpoints	Client certificate registry; gradient anomaly logs; DP configuration audit
G16	Model merge not treated as a supply-chain event	Combined fine-tuned models not vetted before or after merging	MMD-1 through MMD-3: Pre-merge provenance vetting, post-merge backdoor scan, isolated merge environment	Pre-merge vetting checklist; backdoor scan results; isolation configuration review
G17	Quantization not treated as behaviour-changing	Precision reduction not screened for triggered backdoors	QBS-1 through QBS-3: Signed pre-quantization baseline, post-quantization differential analysis, trusted toolchains	Differential analysis report; toolchain hash verification log

G18	Hallucination not linked to incident response	Factual failures not tied to thresholds or corrective action	HAL-1 through HAL-5: RAG grounding verification, confidence thresholds, cross-model consistency, hallucination-rate escalation, domain fact-checking	Grounding verification evidence; hallucination-rate tracking
G19	No post-quantum path for IR cryptography	No dated migration path for the cryptographic evidence underpinning AI incident response	PQC-1: Readiness assessment and dated migration milestones across BOM-4, MCP-1, A2A-2 and REG-4	PQC readiness assessment; dated migration plan

RACI MATRIX

Activity	CISO	AI Sec Lead	SOC Analyst	AI/ML Eng	Legal/ Comp	Safety Eng	Product Owner
Declare Incident	A	R	R	C	C	C (PHY)	I
Execute Auto-Containment	I	C	R	R	I	R (PHY)	I
HITL Approval (Critical)	A	R	C	C	C	R (PHY override)	C
Forensic Investigation	I	R	R	R	C	C (PHY)	I
Manifest Verification	I	A	C	R	C	I	I
Regulatory Notification	A	C	C	I	R	C (PHY)	I
Model Rollback Execution	I	C	I	R	I	I	C
Post-Mortem & Remediation	A	R	C	R	C	C	R
Severity Classification	A	R	R	C	C	C (PHY)	I
HITL Authority Designation (PREP-4)	A	R	I	I	C	C	I

Legend: R = Responsible | A = Accountable | C = Consulted | I = Informed | PHY = applies only to physical AI incidents

SECTION 15

GLOSSARY



Term	Definition
A2A (Agent-to-Agent)	Protocol for inter-agent communication, trust establishment, and orchestration. A primary attack surface in agentic systems for lateral movement and trust-boundary bypass.
AI-BOM / Manifest	Cryptographically signed manifest documenting model lineage, training data sources, embeddings, dependencies, MCP endpoints, and provenance chain. Equivalent to SBOM for AI systems.
Agentic Permission Architecture (APA)	Framework enforcing pre-execution validation, task-scoped least-privilege permissions, and auto-expiring tokens for autonomous AI agents. Addresses both adversarial and structural permission failures.
CASB	Cloud Access Security Broker. Security enforcement layer for SaaS applications used in AI-IRF™ v1.0 for shadow AI detection across SaaS connectors.
CI/CD	Continuous Integration / Continuous Deployment pipeline. In AI-IRF™ v1.0, CI/CD gates enforce manifest signing requirements and prevent unsigned AI artifacts from reaching production.
CIRCIA	Cyber Incident Reporting for Critical Infrastructure Act (US). Requires covered entities to report significant cyber incidents to CISA within 24 hours.
CMDB	Configuration Management Database. In AI-IRF™ v1.0, CMDB must reflect continuously discovered AI instances, not periodic static inventory.
Consequential Action	Any agent action with significant potential impact: data export > 1K records, payment initiation, production delete, PII access, safety-relevant command. Requires HITL approval.
HITL (Human-in-the-Loop)	Mandatory human approval workflow for high-impact automated actions. Requires dual-control sign-off and immutable audit trail. Cannot be bypassed by automated systems.
LLM-as-Judge (LLM-J)	Fine-tuned language model used to classify prompt/response pairs, tool calls, or behavioral patterns as malicious or benign with confidence scoring. Enables machine-speed triage.
Machine-Speed Response	Automated detection and containment operating at adversarial velocity (milliseconds to seconds). Defined in contrast to human-speed triage which is structurally inadequate for agentic attacks.
MCP (Model Context Protocol)	Protocol for LLM-agent tool calling, context exchange, and orchestration. Treated in AI-IRF™ v1.0 as a primary security boundary requiring authentication and semantic inspection.
mTLS	Mutual TLS — both client and server authenticate via certificates. Mandatory for all MCP endpoint traffic under AI-IRF™ v1.0 (MCP-1).
RAG (Retrieval-Augmented Generation)	Architecture pattern combining LLM inference with real-time retrieval from external knowledge sources. Retrieval pipeline is an attack surface for data poisoning and provenance attacks.

Shadow AI	Unmanaged or unsanctioned AI instances including personal deployments, SaaS AI upgrades, embedded AI features, or browser AI extensions operating outside organizational control.
SPIFFE/SPIRE	Secure Production Identity Framework for Everyone / SPIRE Runtime Environment. Open standards for workload identity federation used in AI-IRF™ v1.0 to authenticate agents, tools, and MCP servers.
ZTAI (Zero Trust Architecture for AI)	Application of zero-trust principles specifically to AI systems. Defined in AI-IRF™ v1.0 as: identity-first authentication for every AI component, per-request authorization, micro-segmentation of AI infrastructure layers, runtime least-privilege enforcement, and continuous verification.