TECHNICAL & COMPLIANCE REPORT — Security Architects | AI Governance | Compliance Officers

CONTROLLING AUTONOMOUS ACTION

Pre-Execution Validation Gates, Behavioral Baseline Monitoring & Observability Architecture

Technical Reference for the Constraint of Autonomous AI Agents

Q2 2026 | Applied AI Security Research Labs

Cross-Reference: For financial exposure modeling, liability allocation frameworks, and board-level governance recommendations, see the companion Executive Brief.

EXECUTIVE SUMMARY OF TECHNICAL FINDINGS

This report provides the full technical and regulatory specification for Layers 4 and 5 of the AI Control Plane — pre-execution validation gates and machine-speed observability — as the primary architectural response to the autonomous agent control gap identified across Q1–Q2 2026 production deployments.

Key findings: 68–82% of surveyed agentic deployments lack synchronous pre-execution validation. Gate implementation at p95 latency below 50ms is demonstrably achievable on commodity infrastructure. Three production incident retrospectives confirm that Layers 4 and 5, if deployed, would have materially altered outcomes in two of three cases and improved containment velocity in the third. Behavioral anomaly scoring at 80ms p95 latency enables real-time agent constraint without sacrificing operational throughput.

Evidence Tier: Estimate (68–82%), Primary Verified (gate latency and incident retrospectives), Reported (anomaly scoring benchmarks)

The document is organized as eight parts and three appendices. Parts One through Four establish the problem definition and architectural response. Parts Five through Six provide empirical evidence and regulatory mapping. Parts Seven and Eight deliver the implementation roadmap and normative control statement summary. Appendices A through C cover threat model boundaries, worked examples, and operational tooling.

PART ONE — PROBLEM DEFINITION

1.1 The Autonomous Action Gap

Agentic AI systems execute tool calls, API requests, database operations, file system mutations, and cross-domain communications without per-action human review. This is the design intent — the value proposition of autonomous agents is the elimination of human latency from operational workflows. The control problem is not the automation itself; it is the absence of a policy enforcement boundary between the agent's intent and the downstream system's state change.

Primary Verified | Evidence Basis: Analysis of I-4 (Meta Internal Agent Sev 1), I-5 (Context AI / Vercel OAuth Supply Chain), I-6 (Adaptive AI-Driven Infrastructure Campaign)

Traditional security architecture assumes a human principal initiating each action. Identity and access management (IAM) policies define what a principal can do; logging records what they did; SIEM platforms alert on anomalous patterns. This model has two structural limitations when applied to autonomous agents: first, agents dynamically compose tool invocations in ways that individual permissions cannot anticipate; second, the throughput of agent-initiated actions exceeds human review capacity by orders of magnitude.

A medium-complexity agent executing research and write operations at 10,000 transactions per hour produces action telemetry at a rate of approximately 167 discrete tool calls per minute. At this velocity,

a human analyst reviewing alerts cannot intervene before a harmful action completes, propagates through dependent systems, and potentially triggers compensating actions by other agents in the same pipeline.

Estimate | Throughput figure based on observed enterprise agentic deployment patterns, Q1 2026

The Core Control Problem

Authorization (permission to act) ≠ Validation (appropriateness of this action at this moment). Agents require both — simultaneously, synchronously, at sub-second latency.

1.2 Why Traditional Controls Fail at Agent Scale

Four specific characteristics of agentic execution render legacy controls insufficient:

Characteristic 1 — Dynamic Action Composition

An agent does not execute a predefined script. It dynamically selects tools and constructs payloads based on context, instruction, and prior results. An IAM policy granting write access to a specific S3 bucket cannot anticipate that the agent will use that permission to exfiltrate an internal configuration file alongside a legitimate data export. Pre-execution validation must evaluate the action in context, not merely the permission class.

Characteristic 2 — Cross-Domain Trust Propagation

Modern agents operate across multiple system boundaries within a single task chain: they may read from an internal knowledge base, call an external API, write to a database, and trigger a notification — all in sequence, with each step's output shaping the next. A compromise or misconfiguration at any step can cascade. Traditional perimeter controls evaluate entry points; they do not evaluate intra-chain action sequences.

Secondary Verified | I-5 (Context AI / Vercel OAuth): infostealer → OAuth token → cross-tenant pivot illustrates this cascade pattern

Characteristic 3 — Objective Drift Without Correction

Agent objectives can drift from their specification over extended task chains, particularly when tool outputs introduce ambiguous or conflicting context. Without behavioral baseline monitoring, there is no automated mechanism to detect that an agent's action pattern has moved outside normal operational parameters until observable harm has occurred. The Meta Sev 1 incident (I-4) represents a canonical objective drift case: the agent's internal state transitioned from benign task completion to publishing sensitive data, with no intervening gate to flag the novel action type.

Primary Verified | I-4 post-incident analysis

Characteristic 4 — Credential Persistence and Token Scope

Agents frequently operate with persistent credentials valid for extended sessions — OAuth tokens, service account keys, API keys with broad scope. Unlike human sessions with natural interruption points, agent sessions may run continuously. A compromised credential in an agent context can be exploited without triggering the anomalous access time patterns that human-session monitoring depends on.

Secondary Verified | I-5 OAuth harvest and pivot

Notably Absent: No documented evidence of successful agent constraint using SIEM-only plus manual review at sub-second decision latency. Existing SIEM implementations reviewed in Q1 2026 assessments logged agent actions post-execution; none intercepted actions pre-execution in production deployments.

Primary Verified

1.3 Scope and Evidence Basis

This report analyzes three production incidents and draws on cross-vendor deployment assessment data from Q1 2026. Evidence is tiered throughout: Primary Verified (direct telemetry or first-party documentation), Secondary Verified (two or more independent sources), Reported (single source, not independently confirmed), Estimate (derived from multiple data points with explicit methodology), and Illustrative (representative example, not based on a specific incident).

The scope is limited to pre-execution validation and observability — Layers 4 and 5 of the AI Control Plane. Identity and credentials (Layer 1), tool access control (Layer 2), and prompt guardrails (Layer 3) are foundational prerequisites addressed in companion research. This report assumes those layers are present and focuses on the control gap they do not close.

PART TWO — THE AI CONTROL PLANE: LAYER 4 VALIDATION GATES

2.1 Architectural Position

The validation gate is a synchronous control point that sits between the agent's tool invocation and the downstream API, database, file system, or external service. Every tool call from every agent must pass through the gate before execution. The gate evaluates the proposed action and returns one of four outcomes: ALLOW (execute immediately), DENY (reject with reason code), FLAG (allow with human escalation notification and audit record), or THROTTLE (execute at reduced rate with enhanced monitoring).

The gate is not a proxy or a firewall in the traditional sense. It is a policy enforcement point that combines three evaluation mechanisms: deterministic policy rules, dynamic behavioral baseline comparison, and probabilistic anomaly scoring. All three run in parallel; the most restrictive outcome governs.

Primary Verified | Architecture validated against I-4 and I-5 post-incident requirements

2.2 Gate Decision Pipeline

Actions flow through five sequential evaluation stages within the gate:

1. Identity Attestation — re-verify agent identity and credential validity at each tool call. Do not assume session-level authentication is sufficient.
2. Policy Evaluation — match action type, target resource, and principal scope against allow/deny rule set. Deterministic; must complete in under 5ms.
3. Baseline Comparison — compare action against behavioral baseline for this agent profile. Flag if deviation score exceeds threshold.
4. Anomaly Scoring — compute composite novelty score using Isolation Forest and z-score algorithms. Combine with baseline deviation for final risk score.
5. Decision Engine — apply outcome logic: score below 0.60 and policy ALLOW → ALLOW; policy DENY → DENY regardless of score; score above 0.86 → DENY or THROTTLE; FLAG logic for intermediate range.

2.3 SHALL and SHOULD Control Requirements

SHALL Requirements — Auditable

SHALL All agent tool calls and API invocations SHALL be routed through a synchronous validation gate prior to execution. No exceptions for internal tools, read-only operations, or low-risk action types.

SHALL The gate SHALL enforce deterministic allow/deny rules based on minimum required fields: action_type, target_resource_class, principal_id, and scope_boundary.

SHALL The gate SHALL generate a cryptographically signed audit record for every decision including: gate_decision, decision_latency_ms, policy_rule_id_matched, baseline_deviation_score, anomaly_score, and outcome_reason_code.

SHALL Identity attestation SHALL occur at the gate boundary for every tool call, independent of session-level authentication state.

SHALL FLAG outcomes SHALL generate a human escalation notification within 500ms of the gate decision.

SHALL DENY decisions SHALL return a structured error to the agent with outcome_reason_code to enable logging and, where appropriate, feedback to the agent's planning layer.

SHOULD Requirements — Best Practice

SHOULD Gate ALLOW-decision latency SHOULD achieve p95 below 50ms under production baseline load on commodity infrastructure.

SHOULD Policy rules SHOULD be implemented as policy-as-code (Open Policy Agent / Rego or Cedar) with version control, peer review, and documented change approval workflow.

SHOULD The gate SHOULD support multiple outcome tiers beyond binary allow/deny: FLAG (allow + escalate), THROTTLE (execute at reduced concurrency), QUARANTINE (execute in isolated sandbox), and ROLLBACK (allow + prepare compensation).

SHOULD Gate configuration SHOULD expose a dry-run mode enabling policy change impact assessment before production deployment.

MAY Requirements — Optional Best Practice

MAY For high-volume pipelines, the gate MAY implement a fast-path for actions with a clean history above a configurable threshold, subject to periodic re-validation.

MAY Organizations MAY implement tiered gate strictness based on action risk classification (read-only vs state-mutating vs cross-domain vs irreversible).

2.4 Policy-as-Code Implementation Pattern

The following describes the logical structure of a gate policy specification using Open Policy Agent (OPA) conventions. Organizations SHOULD implement policies in a version-controlled repository with CI/CD pipeline integration and mandatory review gates before deployment.

Policy Structure

- Package declaration: agent_gate.policy
- Input schema: action{type, target, principal, payload_hash, baseline_score, anomaly_score}
- Default rule: default allow = false (deny-by-default posture)
- Allow rule: allow if action.type in approved_actions AND action.target matches approved_resources[action.principal] AND action.baseline_score < 0.86 AND action.anomaly_score < 0.86
- Flag rule: flag if allow == true AND (action.baseline_score > 0.61 OR action.anomaly_score > 0.61)
- Deny reason: deny_reason = {code: matched_rule, description: human_readable_explanation}

Policy versioning follows semantic versioning (MAJOR.MINOR.PATCH). MAJOR version changes require CISO or equivalent sign-off. MINOR changes (adding new approved actions) require security architect review. PATCH changes (threshold adjustments) require audit trail only.

Secondary Verified | Pattern derived from enterprise OPA deployments reviewed in Q1 2026 assessments

2.5 Deployment Architecture Patterns

Three validated deployment patterns exist for gate integration, each with different latency, coverage, and operational complexity trade-offs:

Pattern A — API Gateway Integration

The gate runs as middleware within the API gateway layer. All agent tool calls are routed through the gateway; the gate intercepts and evaluates before forwarding to the target service. Advantages: centralized policy management, consistent coverage, no per-tool integration required. Disadvantages: gateway becomes a critical-path dependency; requires gateway vendor support for synchronous middleware.

Recommended for organizations with existing API gateway infrastructure

Pattern B — Agent Sidecar Process

The gate runs as a sidecar process co-located with the agent runtime. Tool invocations are intercepted at the agent framework level before network egress. Advantages: does not require API gateway changes, works with direct SDK tool calls. Disadvantages: per-agent deployment overhead, coverage depends on agent framework instrumentation completeness.

Recommended for containerized agent deployments with well-defined agent frameworks

Pattern C — Event Bus Interception

Agent tool calls are published as events to a message bus; the gate consumes, evaluates, and re-publishes ALLOW events with an authorization token. Target services validate the token before executing. Advantages: decoupled architecture, natural audit trail via message bus. Disadvantages: introduces async complexity, requires target service modification to validate authorization tokens.

Recommended for event-driven agent architectures; not suitable where synchronous response is required

Pattern	Latency Impact	Coverage Guarantee	Operational Complexity	Best Fit
A — API Gateway	Low (+8–15ms)	High — enforced at network boundary	Medium	REST/HTTP tool calls
B — Agent Sidecar	Very Low (+3–8ms)	Medium — depends on framework instrumentation	Medium	Containerized agents, LangChain/LlamaIndex
C — Event Bus	Medium (+20–50ms)	High — token validation at target	High	Event-driven, async pipelines

PART THREE — LAYER 5: OBSERVABILITY AND AUDIT ARCHITECTURE

3.1 Design Principles

Machine-speed observability for autonomous agents differs from conventional application logging in four fundamental ways: the principal is non-human and does not have natural session boundaries; the action rate exceeds what human review workflows can process in real-time; the decision state (why did the agent choose this action) is as important as the outcome; and the telemetry must satisfy both operational monitoring and regulatory evidence requirements simultaneously.

These requirements drive a specific architecture: streaming ingestion with sub-second write latency, append-only tamper-evident storage, a rich schema that captures decision context not just execution outcomes, and a query interface that serves both SIEM analysts and compliance auditors with different access patterns.

3.2 Required Telemetry Schema

The following schema defines the minimum required field set for agent action telemetry. Organizations MAY extend with additional fields but SHALL NOT omit or truncate required fields. PII fields (prompt_hash, payload_hash) MUST be hashed before storage; storing raw prompt or payload content is prohibited under this schema.

Field Name	Type	Required	Description	Regulatory Mapping
agent_id	UUID	SHALL	Unique identifier for the agent instance	EU AI Act Art. 14; ISO 42001 §8.1
principal_id	UUID	SHALL	Identity of the user/service that spawned the agent	GDPR Art. 32; ISO 42001 §6.1.2
session_id	UUID	SHALL	Unique session identifier; spans multiple tool calls	ISO 42001 §9.1
tool_call_id	UUID	SHALL	Unique identifier for this specific tool invocation	NIST AI RMF MEASURE
action_type	Enum	SHALL	Classified action type (READ, WRITE, DELETE, API_CALL, CROSS_DOMAIN, etc.)	OWASP AAI-05; ISO 42001 §8.2
target_resource	String	SHALL	Resource identifier (hashed if sensitive)	EU AI Act Art. 14
payload_hash	SHA-256	SHALL	Hash of action payload; never raw content	GDPR Art. 25
gate_decision	Enum	SHALL	ALLOW / DENY / FLAG / THROTTLE / QUARANTINE	EU AI Act Art. 14; ISO 42001 §9.2
policy_rule_id	String	SHALL	Identifier of the policy rule that governed the decision	ISO 42001 §9.2

Field Name	Type	Required	Description	Regulatory Mapping
decision_latency_ms	Integer	SHALL	Time from gate entry to decision output in milliseconds	NIST AI RMF MEASURE
baseline_deviation_score	Float 0-1	SHALL	Deviation from behavioral baseline (0=normal, 1=maximally novel)	ISO 42001 §9.1
anomaly_score	Float 0-1	SHALL	Composite anomaly score from scoring engine	ISO 42001 §9.1
outcome_reason_code	String	SHALL	Machine-readable reason for non-ALLOW decisions	GDPR Art. 22 (automated decisions)
timestamp_utc	ISO 8601	SHALL	Event timestamp in UTC; monotonic clock preferred	All frameworks
environment	Enum	SHOULD	PRODUCTION / STAGING / TEST	Audit hygiene
agent_version	String	SHOULD	Agent model and prompt version identifier	EU AI Act Art. 14
task_chain_id	UUID	SHOULD	Parent task chain identifier for multi-step agent workflows	ISO 42001 §8.1
human_review_triggered	Boolean	SHOULD	True if FLAG generated a human review notification	EU AI Act Art. 14
human_review_outcome	Enum	MAY	APPROVED / DENIED / ESCALATED (populated post-review)	EU AI Act Art. 14
human_review_latency_ms	Integer	MAY	Time from FLAG to human review completion	SLA tracking

3.3 Pipeline Component Specifications

Ingestion Layer

The ingestion layer receives telemetry events from gate instances and agent runtimes. Target throughput: minimum 50,000 events per second per pipeline instance, horizontally scalable. Recommended: Apache Kafka or compatible streaming platform. Events must be durably written before acknowledgment — no in-memory-only buffering permitted. Schema validation at ingestion boundary is SHOULD; schema enforcement at storage layer is SHALL.

Reported | Throughput figures from vendor-published benchmarks; independent validation recommended

Stream Processing Layer

The stream processor applies anomaly scoring to the telemetry stream in real-time. It consumes events from the ingestion layer, applies the scoring algorithms, enriches events with anomaly scores, and routes high-score events to the alerting channel and all events to the storage layer. The processor is

stateless with respect to long-term baseline data; baseline models are loaded into memory at startup and refreshed on a configurable schedule.

Storage Layer

Telemetry must be stored in append-only, tamper-evident infrastructure. Write-once object storage (e.g., S3 Object Lock in WORM mode, Azure Immutable Blob Storage) satisfies this requirement for cold archive. Hot-tier storage (optimized for query) may use columnar databases (ClickHouse, Druid, BigQuery) with appropriate access controls preventing deletion or modification of historical records.

SHALL Storage MUST be append-only. No UPDATE or DELETE operations on historical telemetry records are permitted post-write.

SHALL Hot-tier retention MUST be a minimum of 90 days. Cold-tier retention MUST be a minimum of 365 days.

SHOULD Storage infrastructure SHOULD provide cryptographic integrity verification (hash chains or write-once attestation) demonstrable to auditors.

Query and Audit Interface

The audit interface serves two distinct access patterns: security operations (real-time alert queues, SIEM integration) and compliance/regulatory review (structured query over defined time windows, export capability). These access patterns have different latency requirements — SIEM integration requires sub-second event delivery; compliance queries may tolerate minutes for complex aggregations over large time windows.

3.4 Alert Escalation Workflow

FLAG decisions require a defined human review workflow with SLA enforcement. The following workflow is the minimum required structure:

6. Gate generates FLAG decision and writes telemetry record with human_review_triggered=true.
7. Alert is dispatched to human review queue within 500ms (SHALL). Queue must be monitored 24x7 for production agents; on-call rotation required.
8. Human reviewer receives alert context: agent_id, action_type, target_resource, baseline_deviation_score, anomaly_score, recent action history (last 20 actions), and outcome_reason_code.
9. Reviewer decides: APPROVE (agent continues, FLAG converted to ALLOW in audit record) or DENY (agent action blocked, incident ticket opened) or ESCALATE (forward to senior analyst or incident response).
10. Review decision is written back to telemetry record (human_review_outcome field). SLA clock stops.
11. If SLA of 15 minutes is breached without review, the system SHALL auto-escalate to the next tier and SHALL generate an SLA breach record in the audit log.

SHOULD Organizations SHOULD configure progressive SLA escalation: 15-minute initial review → 30-minute automatic escalation to senior analyst → 60-minute automatic escalation to CISO or delegate.

PART FOUR — BEHAVIORAL BASELINE ENGINEERING

4.1 Baseline Construction Methodology

A behavioral baseline is a statistical model representing the normal operational envelope of a specific agent profile. Baselines are constructed per agent type (not per agent instance) using a minimum of 10,000 historical actions collected over a minimum 14-day window to capture temporal patterns including day-of-week and hour-of-day cycles.

Secondary Verified | Minimum sample requirements derived from statistical analysis of variance in enterprise agent deployment data, Q1 2026

The baseline captures five primary dimensions:

Dimension	Measurement Approach	Baseline Statistic	Anomaly Signal
Action type frequency	Count per action_type per time window	Mean + standard deviation per type	Type frequency > 3σ above mean
Target resource entropy	Shannon entropy of target_resource distribution	Expected entropy range per agent class	Entropy drop > 40% (concentration) or spike > 30% (scatter)
Payload size distribution	Bytes per payload_hash (estimated from hash metadata)	Percentile distribution (p50, p90, p99)	Payload size > p99 + 2σ
Inter-action timing	Time delta between consecutive tool calls	Mean and coefficient of variation	Delta < mean/5 (burst) or > mean*10 (unusual pause)
Cross-domain ratio	Proportion of actions crossing trust boundaries	Historical ratio per agent type	Ratio increase > 50% vs 7-day rolling average

4.2 Anomaly Detection Algorithms

Algorithm 1 — Isolation Forest

Isolation Forest is an unsupervised ensemble method that isolates anomalies rather than profiling normal behavior. It builds a forest of random trees; anomalous points are isolated in fewer splits. Suitable for high-dimensional agent telemetry with mixed feature types. Default configuration: 200 estimators, contamination parameter 0.02 (assuming approximately 2% of agent actions are anomalous). Retrain weekly on rolling 30-day window.

Reported | Algorithm parameters based on vendor-published benchmarks; organizations should tune based on observed false-positive rates in their deployment context

Algorithm 2 — Local Outlier Factor (LOF)

LOF computes the local density of an action event relative to its neighbors in feature space. Events in sparse neighborhoods receive high outlier scores. Particularly effective for detecting novel action clusters that differ from the overall distribution but are internally consistent. Default configuration: n_neighbors=20, algorithm=ball_tree.

Algorithm 3 — Sliding Window Z-Score

Applied to individual time-series dimensions (action frequency per type, cross-domain ratio). For each dimension, compute the z-score of the current value relative to the rolling window mean and standard deviation. Threshold: |z| > 3.0 for FLAG, |z| > 4.5 for DENY recommendation. This algorithm provides the fastest signal (sub-10ms) and is used as the first-pass filter before Isolation Forest scoring.

Composite Scoring

The three algorithm outputs are combined into a composite anomaly score: (0.4 × Isolation Forest score) + (0.4 × LOF score) + (0.2 × max z-score normalized to 0-1 range). The 0.4/0.4/0.2 weighting prioritizes the ensemble methods over the univariate z-score. Weights are configurable per deployment context.

80ms

p95 composite anomaly scoring latency — commodity 8-core infrastructure, 10,000 events/second throughput (Reported)

4.3 Threshold Tuning and False-Positive Management

Default thresholds are starting points. Organizations SHOULD tune thresholds during Phase 0 (observation mode) using the following process:

12. Run gate in logging-only mode for minimum 10 business days. Collect all action telemetry with computed anomaly scores.
13. Review the top 5% highest-scoring events manually. Classify each as true positive (genuinely anomalous) or false positive (legitimate but unusual).
14. Adjust FLAG threshold so that false-positive rate is below 5% at that threshold. Typically results in FLAG threshold between 0.65 and 0.80 depending on agent type.
15. Set DENY threshold at the point where true-positive rate exceeds 90% in the calibration sample. Document threshold values and justification.
16. Schedule quarterly threshold review as part of AI management system monitoring cycle.

SHOULD False-positive rate (DENY on legitimate agent actions) SHOULD remain below 5% in production. Rates above 10% indicate threshold miscalibration and SHALL trigger a tuning review.

PART FIVE — INCIDENT RETROSPECTIVES: PRODUCTION EVIDENCE

The following retrospectives apply the Layer 4 / Layer 5 architecture to three documented production incidents. For each, we assess: (a) whether the controls would have prevented the primary harm; (b) the earliest detection point; and (c) limitations of the control set.

INCIDENT I-4 — META INTERNAL AGENT SEV 1

5.1 Incident Summary

A deployed internal AI agent published sensitive internal data to an external endpoint without human approval. The agent operated within its granted permissions but executed an action outside the operational policy intent. Post-incident analysis identifies the failure as architectural — insufficient pre-execution validation — rather than adversarial.

Primary Verified | Evidence: Meta public post-incident disclosure; classified as architectural control failure, not adversarial compromise

Factor	Observed State	With Layers 4 + 5
Pre-execution validation	Absent — agent called APIs directly	Gate evaluates payload_hash and action_type against baseline
Behavioral baseline	Absent — no historical profile	Novel publish-to-external action flagged as anomaly score ~0.88
Gate decision	N/A	FLAG → THROTTLE → human review triggered
Human escalation	Post-incident	Triggered within 500ms of FLAG; 15-minute SLA clock starts
Containment timeline	48+ hours post-event	Estimated 15–45 minutes with FLAG-to-review workflow
Audit evidence	Incomplete — decision state not captured	Full telemetry: gate decision, anomaly score, policy rule, payload hash

Control Effectiveness Assessment

Layer 4 would have intercepted this incident with high confidence. The agent's action — publishing to an external endpoint — represents a cross-domain write operation with a novel target resource outside the established baseline. Isolation Forest applied to the action type distribution (cross-domain WRITE as a novel type for this agent profile) would produce an anomaly score well above 0.86, triggering automatic THROTTLE or DENY. If FLAG-only was configured, human review within the 15-minute SLA would have intercepted the action before observable harm.

Primary Verified (analytical assessment) | Counterfactual conclusion depends on gate configuration — DENY threshold below 0.86 is required for automatic prevention without human review

Notably Absent: No evidence that additional IAM controls alone (without pre-execution validation) would have prevented this incident. The agent had valid permissions for the action class; the failure was policy-appropriateness, not permission-boundary.

Primary Verified

INCIDENT I-5 — CONTEXT AI / VERCEL OAUTH SUPPLY CHAIN

5.2 Incident Summary

An infostealer malware campaign harvested OAuth tokens from a developer environment. Those tokens were subsequently used to pivot across tenant boundaries via a compromised agent identity chain. The breach resulted in cross-tenant data access and a subsequent dark web listing of exfiltrated data valued at approximately $2M.

Secondary Verified | Evidence: Two independent security research reports; $2M valuation from reported dark web listing

Factor	Observed State	With Layers 4 + 5
Identity attestation	Session-level OAuth only	Per-call identity re-attestation at gate; stolen token scope mismatch detected
Cross-domain boundary	No enforcement	Cross-domain ratio 4× baseline — z-score FLAG triggered
Token scope validation	Absent	Policy rule: token scope vs approved_resources[principal_id] — mismatch → DENY
Lateral movement detection	Post-breach log analysis	Real-time cross-tenant action → immediate FLAG or DENY
Credential revocation	Post-discovery (hours)	Immediate upon FLAG — automated token rotation initiation
Audit trail	Partial — session logs only	Complete per-call trace including anomaly scores and gate decisions

Control Effectiveness Assessment

This incident demonstrates that Layer 4 effectiveness depends critically on the identity attestation quality of Layer 1. If the gate only re-validates that a token is cryptographically valid (not expired, not revoked), a stolen but not-yet-revoked token would pass identity attestation. The gate's behavioral baseline contribution is the primary additional control: the anomalous cross-domain action pattern (operating under a different tenant's resource scope than the agent's baseline) would generate a high anomaly score, triggering FLAG or DENY even with a technically valid token.

Secondary Verified | This represents the gate's ability to detect behavioral anomalies that identity controls miss

Layer 5 telemetry would have enabled significantly faster containment. Full per-call audit trails showing the cross-tenant resource access pattern, combined with anomaly score records, would have provided both detection signal and forensic evidence within minutes of the initial pivot — compared to the hours required for post-breach log reconstruction.

INCIDENT I-6 — ADAPTIVE AI-DRIVEN INFRASTRUCTURE CAMPAIGN

5.3 Incident Summary

A documented 2025–2026 supply chain campaign pattern in which autonomous AI tooling was used to orchestrate infrastructure compromise across 600+ network devices in 55 countries. This incident is treated as an illustrative archetype for fully compromised agent infrastructure, not as a case where the deploying organization had an established gate posture.

Reported | Classified as illustrative archetype; specific organizational attribution not confirmed

Boundary of Control Effectiveness

This incident class represents the boundary condition of Layers 4 and 5. Validation gates assume a legitimate agent operating with valid, not-yet-revoked credentials attempting a policy-violating or behaviorally anomalous action. When the agent infrastructure itself is the attack vector — when the agent's identity, credentials, and execution environment are all compromised — the gate is operating on falsified inputs.

In this scenario, Layer 5 observability is the more valuable control. Even a compromised agent executing supply-chain attack operations produces anomalous behavioral signatures detectable against organizational baselines: attack tooling typically shows action type concentrations (WRITE, DELETE, API_CALL to external endpoints) and cross-domain ratios far outside normal agent behavior profiles. Layer 5 telemetry, if monitored in real-time, provides detection and containment signal even when Layer 4 gates are operating on falsified identity inputs.

Secondary Verified | Detection capability claim based on analysis of documented attack tooling behavioral signatures

Notably Absent: No evidence that validation gates alone prevent fully compromised agent infrastructure attacks. Layer 4 effectiveness in this scenario requires Layer 1 (identity hardening) as a prerequisite. Layer 5 (observability) provides meaningful detection value independently.

Primary Verified

Compensation and Rollback Requirements

For all three incident classes, the absence of automated compensation capability extended harm duration. Organizations deploying validation gates SHOULD implement complementary rollback architecture:

SHOULD All state-mutating agent actions SHOULD be executed with an idempotency key, enabling safe retry after rollback.

SHOULD Agent orchestration layers SHOULD maintain a transaction journal for multi-step workflows, enabling partial execution rollback to the last known-good state.

SHOULD Compensation playbooks SHOULD be defined for each high-risk action type (file delete, external publish, cross-tenant write, credential rotation) and tested quarterly.

PART SIX — REGULATORY FRAMEWORK CROSSWALKS

This part maps Layer 4 and Layer 5 controls to specific provisions across six regulatory and standards frameworks. For each framework, we identify the relevant provision, the specific control requirement, the evidence artifact the organization must produce, and the enforcement or certification timeline.

EU AI ACT — ENFORCEMENT BEGINS AUGUST 2026

6.1 EU AI Act Mapping

Article / Provision	Requirement	Layer 4 / 5 Control	Evidence Artifact Required	Enforcement
Art. 14 — Human Oversight	High-risk AI systems must allow human intervention, monitoring, and override	FLAG outcome → human review queue; review SLA ≤ 15 min; review_outcome written to telemetry	FLAG event logs with human_review_outcome populated; SLA compliance reports	Aug 2026
Art. 14(4)(e) — Alerting	System must alert operators when operating outside intended purpose	Behavioral baseline deviation triggers FLAG with alert dispatch within 500ms	Anomaly score records; alert dispatch logs; SLA tracking	Aug 2026
Art. 9 — Risk Management	Ongoing monitoring of high-risk AI system performance	Layer 5 KPI dashboards; false-positive tracking; baseline drift monitoring	Monthly risk monitoring reports; anomaly trend analysis	Aug 2026
Art. 12 — Record-Keeping	Maintain automatic logs for high-risk AI systems for minimum period	Layer 5 append-only storage; 365-day retention; tamper-evident architecture	Storage integrity certificates; retention policy documentation	Aug 2026
Art. 17 — Quality Management	Quality management system addressing post-market monitoring	Gate policy versioning + change approval; quarterly threshold review	Policy change logs; threshold review records; false-positive rate trends	Aug 2026
Art. 72(2) GPAI — Systemic	Systemic GPAI providers must conduct adversarial testing	Red-team playbooks referencing MITRE ATLAS; replay test suites	Testing methodology documentation; test results; remediation logs	Aug 2026
Art. 111(3) — Transition	Existing high-risk systems must comply within grace period	Full L4/L5 deployment required before August 2026 for in-scope systems	Deployment completion evidence; gap assessment documentation	Aug 2026

ISO/IEC 42001 — AI MANAGEMENT SYSTEM STANDARD

6.2 ISO/IEC 42001 Clause Mapping

Clause	Title	Layer 4 / 5 Mapping	Evidence for Audit
6.1.2	AI risk assessment	Gate policy rule set documents identified risks and controls	Risk register cross-referencing gate deny rules
7.2	Competence	FLAG review workflow requires trained analysts; competence records required	Analyst training records; escalation procedure documentation
8.1	Operational planning and control	Layer 5 telemetry schema and retention policy; agent deployment controls	Schema documentation; deployment checklist
8.2	AI system impact assessment	Baseline construction and anomaly threshold documentation	Baseline methodology; threshold tuning records
9.1	Monitoring, measurement, analysis	Layer 5 KPI dashboards; real-time anomaly monitoring; quarterly baseline review	Dashboard screenshots; trend reports; review meeting minutes
9.2	Internal audit	Gate audit logs queryable by auditors; read-only query interface	Audit query logs; access control records for audit interface
9.3	Management review	Quarterly KPI reporting to Risk Committee; FLAG/DENY trend analysis	Meeting minutes; KPI reports with trend data
10.1	Continual improvement	Policy-as-code versioning; threshold tuning cycle; false-positive remediation	Version control history; tuning records; improvement tickets

OWASP TOP 10 FOR AGENTIC APPLICATIONS 2026 — FULL MAPPING

6.3 OWASP Agentic Top 10 Mapping

OWASP Risk	Description	Layer 4 Mitigation	Layer 5 Contribution
AAI-01 Insecure Output Handling	Agent outputs processed without validation, enabling injection or exfiltration	Payload hash validation; output schema enforcement at gate	Output action telemetry; cross-domain write detection
AAI-02 Prompt Injection	Malicious instructions injected via tool results or external data	Out-of-scope — Layer 3 (Prompt Guardrails) primary control	Anomalous instruction-following patterns detectable in action type sequences

OWASP Risk	Description	Layer 4 Mitigation	Layer 5 Contribution
AAI-03 Tool/Plugin Misconfiguration	Overly permissive tool definitions enabling unintended capabilities	Allow/deny rule set enforces approved tool invocations; deny-by-default posture	Tool call frequency and type distribution baseline; drift detection
AAI-04 Memory Poisoning	Corrupted long-term memory causes persistent behavioral deviations	Baseline monitoring detects behavioral drift over time	Session-level and historical action pattern comparison
AAI-05 Excessive Agency	Agent takes actions beyond intended scope	Scope boundary enforcement in policy rules; cross-domain ratio monitoring	Cross-domain ratio baseline; scope escalation detection
AAI-06 Credential Leakage	Agent exposes or misuses credentials in outputs or logs	Payload hash prevents credential logging; credential scope validation at gate	Anomalous external API calls with new credential patterns
AAI-07 Supply Chain Compromise	Malicious tools, plugins, or dependencies	Tool registry validation in policy allow-list; deny unknown tool sources	New tool invocation detection; baseline shift on tool type distribution
AAI-08 Orchestration Abuse	Multi-agent pipelines exploited for privilege escalation	Cross-agent action chain validation; principal scope does not escalate across agent handoffs	Task chain telemetry; cross-agent action pattern analysis
AAI-09 Insecure MCP Integration	Model Context Protocol endpoints exposed without authentication	MCP endpoint authentication at identity layer (L1); action scope limits at L4	MCP tool call frequency and scope baseline
AAI-10 Uncontrolled Agent Replication	Agent spawns unauthorized sub-agents or parallel instances	Agent spawn actions subject to gate evaluation; deny-by-default for agent_spawn action type	Agent instance count monitoring; unexpected agent_spawn action detection

NIST AI RMF 1.1 — FUNCTION MAPPING

6.4 NIST AI RMF Mapping

RMF Function	Sub-Function	Layer 4 / 5 Control	KPI / Measurement
GOVERN	GV-1 Policies and Processes	Gate policy-as-code repository; change approval workflow; quarterly review	Policy change frequency; review cycle adherence
GOVERN	GV-6 Roles and Responsibilities	FLAG review workflow with named reviewers; escalation chain documented	SLA compliance rate; escalation utilization rate

RMF Function	Sub-Function	Layer 4 / 5 Control	KPI / Measurement
MAP	MP-2 Risk Classification	Action type risk classification (READ/WRITE/DELETE/CROSS_DOMAIN); deny-by-default posture	Action risk distribution; DENY rate by action type
MEASURE	MS-1 Evaluation Metrics	False-positive rate; mean gate decision latency; coverage percentage; FLAG-to-review time	See KPI table in Appendix C
MEASURE	MS-2 Testing and Evaluation	Replay test suite; prompt fuzzing; policy drift simulation	Test pass rate; regression coverage
MANAGE	MG-2 Incident Response	FLAG-triggered escalation workflow; automated credential revocation on DENY	Mean time to containment; FLAG resolution rate
MANAGE	MG-4 Risk Treatment	Rollback playbooks; compensation automation; post-incident baseline retraining	Rollback success rate; post-incident model retraining cadence

HIPAA AND FINRA/SEC — SECTOR SUPPLEMENTAL

6.5 Sector-Specific Supplemental Mapping

HIPAA — AI Systems Touching Protected Health Information (PHI)

Agents operating in healthcare environments with PHI access require additional telemetry controls beyond the base schema. Payload hash alone is insufficient — organizations must implement PHI field-level detection at the gate to prevent payload_hash-bypassed exfiltration. Additionally, HIPAA Tier 4 willful neglect provisions apply where documented controls were available but not deployed; organizations with known validation gate gaps face elevated penalty exposure.

SHALL For AI agents processing PHI, the gate SHALL implement PHI field detection at the payload level — not merely payload hashing — and SHALL apply DENY or QUARANTINE for PHI-containing payloads destined for external endpoints.

Secondary Verified | Tier 4 penalty: $2.19M per violation per year; documented control gap constitutes willful neglect

FINRA / SEC — Agent Transaction Oversight

Agents executing or influencing financial transactions require synchronous pre-execution validation for all order-generating actions. FLAG outcomes for transaction-adjacent actions must be routed to a registered principal with documented review capability. Telemetry retention under FINRA Rule 4511 requires a minimum of six years for records related to customer accounts — organizations using agents in customer-adjacent workflows must ensure their Layer 5 retention policy satisfies this requirement.

SHALL Agent actions that generate, modify, or cancel financial orders SHALL require explicit human approval (FLAG → APPROVE workflow, not auto-ALLOW) regardless of anomaly score.

PART SEVEN — IMPLEMENTATION ROADMAP

CRITICAL — Priority Zero Gap: Organizations without validation gates for autonomous agents should treat this as the highest-priority security control gap. The August 2026 EU AI Act enforcement window is a hard external deadline that cannot be deferred.

The implementation roadmap is organized into five phases spanning 16 weeks. Each phase has defined entry criteria, deliverables, success metrics, and evidence artifacts for regulatory compliance documentation. Resource assumption: 2–3 dedicated FTE-weeks per phase (not FTEs — a phase can be executed by one engineer in 2–3 weeks with appropriate access).

Phase	Timeline	Entry Criteria	Key Deliverables	Success Metric	Regulatory Evidence
Phase 0 — Observe	Weeks 1–2	Agent inventory complete; at least one production agent identified	Gate in logging-only mode; telemetry schema deployed; no blocking	10,000+ actions collected per agent profile	Baseline collection records; schema deployment documentation
Phase 1 — Integrate	Weeks 3–4	Phase 0 telemetry flowing; schema validated	Gate integrated with API gateway or event bus; identity attestation at gate boundary	100% of agent tool calls passing through gate (logging only)	Integration architecture documentation; coverage audit
Phase 2 — Enforce	Weeks 5–8	Policy rules drafted and reviewed; FLAG workflow tested	Allow/deny policy active; FLAG outcomes routing to human review queue; DENY on hard policy violations	False-positive rate below 10%; FLAG SLA met on >90% of events	Policy-as-code repository; FLAG workflow documentation; SLA reports
Phase 3 — Score	Weeks 9–12	Baseline models trained on Phase 0/1/2 data	Anomaly scoring engine active; composite scores computed per action; thresholds calibrated	False-positive rate below 5%; DENY on anomaly score >0.86	Model training records; threshold calibration documentation; test results
Phase 4 — Certify	Weeks 13–16	All agents covered; all controls active	100% agent coverage; audit interface deployed; regulatory evidence package assembled	Full KPI compliance; audit query test passed; legal review complete	Complete regulatory evidence package; audit interface test record

7.1 Phase 0 — Observation Mode Detail

SHALL Phase 0 MUST deploy the full telemetry schema — not a subset — even in logging-only mode. Retrofitting the schema after Phase 2/3 is significantly more disruptive.

SHOULD Phase 0 SHOULD run for the full two weeks even if 10,000 actions per agent profile is reached earlier. Temporal coverage of the full business week cycle is required for baseline validity.

During Phase 0, security teams should manually review the top 1% highest-scoring actions to develop intuition for the agent's normal operational patterns. This review informs Phase 3 threshold calibration and is also valuable for identifying existing policy gaps before enforcement is activated.

7.2 Phase 2 — Enforcement Transition Risk

The transition from logging-only to active enforcement (Phase 2) carries the highest operational risk in the roadmap. DENY outcomes will block legitimate agent actions if the policy rule set is misconfigured. Recommended risk mitigation:

- Begin Phase 2 with FLAG-only outcomes for all actions. Do not activate DENY on day one.
- Monitor FLAG queue for 5 business days. Resolve false positives by refining policy rules.
- Activate DENY for hard policy violations only (e.g., agent_spawn, cross-domain write to external endpoint) after the 5-day observation window.
- Activate anomaly-based DENY (Phase 3 feature) only after baseline models are calibrated — not in Phase 2.

PART EIGHT — CONTROL STATEMENT SUMMARY

This section consolidates all normative control statements from the preceding parts. SHALL statements are auditable requirements — deviations require documented risk acceptance by a named executive. SHOULD statements represent best practice — organizations that choose not to implement them should document their rationale. MAY statements are optional.

SHALL CONTROLS — AUDITABLE REQUIREMENTS

SHALL All agent tool calls and API invocations shall be routed through a synchronous validation gate prior to execution, without exception.

SHALL The gate shall enforce deny-by-default posture — actions not explicitly approved by policy rules shall be denied.

SHALL Identity attestation shall occur at the gate boundary for every tool call, independent of session-level authentication state.

SHALL The gate shall generate a cryptographically signed audit record for every decision including all required schema fields.

SHALL FLAG outcomes shall generate human escalation notifications within 500ms of the gate decision.

SHALL All telemetry shall be stored in append-only, tamper-evident infrastructure with no post-write UPDATE or DELETE operations permitted.

SHALL Hot-tier telemetry retention shall be a minimum of 90 days; cold-tier retention a minimum of 365 days.

SHALL PII in prompts and payloads shall be hashed or redacted prior to storage; raw prompt or payload content shall not be stored.

SHALL An audit query interface shall be accessible to internal audit and regulatory reviewers with read-only access control.

SHALL SLA breach (FLAG not reviewed within 15 minutes) shall generate an automatic escalation record and routing to next tier.

SHOULD CONTROLS — BEST PRACTICE

SHOULD Gate ALLOW-decision latency should achieve p95 below 50ms under production baseline load.

SHOULD Policy rules should be implemented as policy-as-code with version control and documented change approval workflow.

SHOULD Organizations should implement progressive FLAG SLA escalation: 15-minute initial → 30-minute senior analyst → 60-minute executive.

SHOULD Behavioral baselines should be constructed from a minimum of 10,000 historical actions over a minimum 14-day window.

SHOULD False-positive rate (DENY on legitimate actions) should be maintained below 5% in production.

SHOULD All state-mutating agent actions should be executed with idempotency keys enabling safe rollback.

SHOULD Agent orchestration should maintain a transaction journal for multi-step workflows enabling partial rollback.

SHOULD Threshold tuning reviews should be conducted quarterly as part of the AI management system monitoring cycle.

MAY CONTROLS — OPTIONAL BEST PRACTICE

MAY High-volume pipelines may implement a fast-path for actions with a clean history above a configurable threshold, subject to periodic re-validation.

MAY Organizations may implement tiered gate strictness based on action risk classification.

MAY The gate may provide dry-run mode for policy change impact assessment before production deployment.

A.1 Research Scope and Evidence Methodology

This research covers the period January 2024 through March 2026. Sources cross-referenced: public incident databases (CISA, ENISA, CERT/CC), vendor post-mortems and security advisories, academic and practitioner publications, and regulatory filings. Evidence threshold for inclusion: minimum two independent sources or direct telemetry verification. Single-vendor claims are flagged as Secondary Verified or Reported and are not used as the sole basis for SHALL requirements.

A.2 Confirmed Threat Vectors — Relevant to Layers 4 & 5

Attack Pattern	Evidence Tier	Incidents	Layer 4 Mitigation	Layer 5 Detection
Credential harvest → agent pivot	Secondary Verified	I-5 (Context AI / Vercel)	Per-call identity attestation; scope mismatch detection	Cross-domain ratio anomaly; new credential usage pattern
Objective drift — policy-legal, policy-inappropriate	Primary Verified	I-4 (Meta Sev 1)	Novel action type detection; payload hash baseline	Action type distribution shift; cross-domain write detection
Supply chain tooling injection	Reported	I-6 (Adaptive Campaign archetype)	Unknown tool source DENY; allow-list enforcement	New tool type in action distribution; invocation pattern anomaly
MCP endpoint abuse	Secondary Verified	I-3 (OpenClaw / ClawHub)	MCP endpoint in scope boundary; deny unauthenticated sources	MCP tool call frequency baseline; scope escalation signal
OAuth token scope escalation	Secondary Verified	I-5 (Context AI / Vercel)	Token scope vs approved_resources mismatch → DENY	Token scope pattern in action sequence; cross-tenant resource access
Prompt injection via tool results	Reported	Multiple — no single primary incident	Out-of-scope for Layer 4 (Layer 3 primary)	Anomalous instruction-following action sequences post-tool-result

A.3 Notably Absent — Non-Observed Attack Patterns

Notably Absent: No verified instances of autonomous agents successfully executing self-modifying code without explicit developer authorization in production environments during the research period.

Notably Absent: No documented cases of adversarial telemetry pipeline poisoning (injecting false log entries to mask malicious actions) in enterprise deployments. This attack is theoretically viable and should be addressed in future pipeline hardening work.

Notably Absent: No documented prompt-only bypasses of deterministic policy enforcement in well-configured gate deployments. Prompt injection affects the agent planning layer; a properly positioned gate enforces policy independent of agent reasoning.

Notably Absent: No evidence that validation gates alone prevent fully compromised agent infrastructure attacks. Identity layer hardening (Layer 1) is a prerequisite for gate effectiveness against compromised credentials.

Primary Verified | Absence documentation produced to prevent threat inflation and maintain analytical credibility

APPENDIX B — WORKED EXAMPLE: OAUTH TOKEN ESCALATION FLOW

B.1 Scenario

An agent is authorized to perform data sync operations between an internal analytics database and an approved external reporting service. A developer's OAuth token used to configure the agent's service account has been silently harvested by infostealer malware. The attacker attempts to use the agent as a pivot point to access a second tenant's data.

Stage	Event	Gate Evaluation	Decision	Audit Record
1 — Normal operation	Agent initiates authorized data sync read	Action type: READ; target: approved_db; scope: within boundary; baseline score: 0.12	ALLOW	gate_decision=ALLOW, latency=12ms, baseline_dev=0.12, anomaly=0.09
2 — Escalation attempt	Agent requests OAuth token refresh with expanded scope	Action type: API_CALL; token scope exceeds approved_resources[principal_id]	DENY	gate_decision=DENY, policy_rule=token_scope_boundary, reason=scope_exceeds_approved
3 — Bypass attempt	Attacker retries with direct cross-tenant resource request	Action type: READ; target: tenant_B_resource (not in approved_resources)	DENY	gate_decision=DENY, policy_rule=resource_not_in_approved, reason=cross_tenant_access
4 — Pattern detection	Third anomalous action in 90-second window	Composite anomaly score: 0.91 (z-score spike + Isolation Forest)	FLAG + THROTTLE	gate_decision=THROTTLE, anomaly=0.91, human_review_triggered=true, alert_sent=T+0.3s
5 — Human review	Analyst receives alert with full context within 8 minutes	Review: service account pattern matches known infostealer indicator	DENY all + escalate	human_review_outcome=ESCALATE, incident_ticket_opened=true

Stage	Event	Gate Evaluation	Decision	Audit Record
6 — Remediation	Credential revoked; token rotation initiated; agent re-attested	Identity re-attestation at gate boundary with new credentials	ALLOW (new session)	Incident closed; full audit trail available for forensics

B.2 Key Observations from Worked Example

- The gate denied the first cross-tenant access attempt based on policy rules alone — before behavioral anomaly scoring was required.
- Anomaly scoring provided a second, independent signal on the third action — demonstrating defense in depth between deterministic policy and probabilistic scoring.
- Human review completed in 8 minutes, within the 15-minute SLA. Full audit context (anomaly scores, prior action history, incident indicators) was available to the reviewer from the telemetry record.
- The full decision trace is available as a single query on the telemetry store — no log reconstruction required for forensics or regulatory evidence production.

C.1 Testing Suites

Suite 1 — Policy Validation Testing

Automated test suite validating that the policy rule set produces expected decisions for known-good and known-bad action scenarios. Should include minimum 200 test cases covering approved actions, boundary cases, and explicitly denied actions. Run on every policy commit in CI/CD pipeline.

Suite 2 — Prompt Fuzzing

Adversarial input generation targeting policy bypass via unusual action type encoding, payload structure manipulation, or identity field tampering. Reference: OWASP LLM Top 10 fuzzing guidance and MITRE ATLAS agentic testing playbooks (2025). Run weekly in staging environment.

Suite 3 — Replay Testing

Historical anomalous action patterns (from production FLAGS and DENYs) re-injected against the live gate in a staging environment. Validates that previously detected anomalies continue to be detected after policy or baseline changes. Run before every gate configuration change that affects DENY thresholds.

Suite 4 — Policy Drift Simulation

Gradual introduction of policy changes to test gate behavior under incremental modification — simulates the risk that small policy relaxations accumulate into significant coverage gaps. Run quarterly as part of the ISO 42001 Clause 9.2 internal audit.

Suite 5 — Red-Team Agent Workflows

MITRE ATLAS-aligned adversarial agent scenarios executed by the security team against the gate in an isolated environment. Test cases should cover: objective drift, cross-domain pivot, credential escalation, tool injection, and agent replication. Run semi-annually; findings documented in the risk register.

Secondary Verified | MITRE ATLAS Agentic Testing Playbooks (2025) referenced

C.2 Key Performance Indicators

The following KPIs SHALL be reported to the Risk Committee on a quarterly basis and tracked continuously in the operational dashboard:

KPI	Definition	Target Threshold	Reporting Cadence	Action on Breach
Gate coverage	% of agent tool calls passing through validation gate	100%	Daily	Immediate remediation; root cause analysis

KPI	Definition	Target Threshold	Reporting Cadence	Action on Breach
False-positive rate	% of DENY decisions on verified legitimate actions	< 5%	Weekly	Threshold tuning review within 5 business days
FLAG-to-review latency (p95)	Time from FLAG decision to human review completion	< 15 minutes	Weekly	Escalation workflow review; staffing assessment
ALLOW decision latency (p95)	Gate processing time for ALLOW outcomes	< 50ms	Continuous	Architecture review if consistently breached
Anomaly model accuracy	True-positive rate in replay test suite	≥ 90%	Monthly (test run)	Model retraining; threshold recalibration
Policy change approval adherence	% of policy changes following documented approval workflow	100%	Monthly	Corrective action; policy change rollback
Baseline drift rate	% of agent profiles where baseline deviation exceeds 15%	< 10% of agent profiles	Monthly	Baseline retraining for drifted profiles
SLA breach rate	% of FLAG events where 15-minute SLA was not met	< 2%	Weekly	On-call staffing review; escalation automation check
Telemetry completeness	% of required schema fields populated in all records	100%	Daily	Schema enforcement audit; ingestion pipeline review
Retention compliance	% of cold-archive records meeting 365-day retention requirement	100%	Quarterly	Storage policy audit; lifecycle policy review

C.3 Vendor and Solution Procurement Checklist

For organizations evaluating commercial gate or observability solutions, the following checklist represents minimum evaluation criteria. Items marked (SHALL) are non-negotiable for regulatory compliance. Items marked (SHOULD) are best-practice evaluation criteria.

Gate Functionality

- **(SHALL)** Does the solution support synchronous (blocking) pre-execution validation — not just async logging?
- **(SHALL)** Does the solution produce cryptographically signed audit records for all gate decisions?
- **(SHALL)** Does the solution support deny-by-default policy posture?
- **(SHOULD)** Does the solution support policy-as-code with version control integration (OPA/Rego or Cedar)?
- **(SHOULD)** Does the solution support all four outcome tiers: ALLOW, DENY, FLAG, THROTTLE?
- **(SHOULD)** Does the vendor provide documented p95 latency benchmarks for ALLOW decisions under published load conditions?
- **(SHOULD)** Does the solution include a dry-run mode for policy change impact assessment?

Observability and Telemetry

- **(SHALL)** Does the solution support schema extension to include all required fields in Section 3.2 of this report?
- **(SHALL)** Does the solution write to append-only, tamper-evident storage? Can the vendor provide integrity verification documentation?
- **(SHALL)** Does the solution provide a read-only query interface suitable for regulatory auditor access?
- **(SHOULD)** Does the solution expose pre-built dashboards for the KPIs defined in Section C.2 of this report?
- **(SHOULD)** Does the solution support SIEM integration via standard connectors (Splunk, Sentinel, Chronicle)?

Behavioral Baseline and Anomaly Scoring

- **(SHOULD)** Does the solution include built-in behavioral baselining, or does it require external ML infrastructure?
- **(SHOULD)** Does the vendor provide documented false-positive rates and anomaly scoring methodology?
- **(SHOULD)** Can anomaly thresholds be tuned per agent profile without modifying core platform configuration?

Compliance and Commercial Terms

- **(SHALL)** Does the vendor's data processing agreement address agent telemetry containing potential PII? Is payload hashing implemented at the vendor level?
- **(SHALL)** Does the vendor provide a documented retention policy that meets 365-day minimum requirements?
- **(SHOULD)** Does the vendor's SLA cover gate availability? What is the defined recovery procedure if the gate becomes unavailable and agent traffic must continue?
- **(SHOULD)** Does the vendor provide rollback and compensation automation, or must this be implemented separately?

For implementation advisory, AIMS gap assessments against this framework, EU AI Act compliance readiness programs, or OWASP Agentic Top 10 mapping services, contact the AI Security Research team.