Independent Research

Cryptographic Enforcement Evidence for MCP Security

An Analysis of the CoSAI Workstream 4
MCP Threat Taxonomy

A T T E S T E D I N T E L L I G E N C E

Jack Brennan
admin@attestedintelligence.com

March 28, 2026

AttestedIntelligence.com

TABLE OF CONTENTS

KEY TERMS

Term	Definition
Attested Governance Artifact (AGA)	A sealed, cryptographically signed policy object encoding agent identity binding, authorized behavior, enforcement parameters, and measurement cadence
Subject Identifier	Cryptographic binding comprising hashes of the agent's normalized bytes and canonicalized metadata
Portal (Sentinel)	Runtime enforcement boundary that intercepts agent interactions and enforces sealed artifact constraints
Continuity Chain	Append-only sequence of signed events linked by structural metadata hashes, providing tamper-evident history
Evidence Bundle	Portable verification package containing artifact, receipts, Merkle proofs, and checkpoint references
Sealed Hash	Immutable reference value representing the agent's attested known-good state
Phantom Execution	Forensic capture mode where a compromised agent continues executing in a sandboxed environment while all outputs are recorded as signed receipts
DRIFT Receipt	Signed enforcement record generated when agent behavior diverges from sealed policy parameters

INTRODUCTION

The Coalition for Secure AI (CoSAI) published a comprehensive MCP security white paper in January 2026, approved by the CoSAI Project Governing Board on January 8, 2026. The white paper identifies 12 threat categories and nearly 40 distinct threats across all-local, single-tenant hybrid, and multi-tenant cloud deployments of the Model Context Protocol. It provides recommended mitigations for each category and explicitly notes that “follow on papers will provide reference implementations and recommendations for specific mitigation controls.”

This document examines the enforcement evidence gap across all 12 categories and maps how the Seal/Enforce/Prove governance architecture addresses each gap. For every category, we identify the current mitigation approach, the specific gap where no cryptographic proof exists that the mitigation was enforced, the mechanism that closes that gap, the corresponding patent claim reference, and the evidence artifact produced for verification.

This analysis is independent research. It does not represent a formal submission to the CoSAI working group. The enforcement evidence layer described here does not replace the mitigations the white paper recommends - it provides proof that they were applied.

CORE FINDING

Across all 12 CoSAI threat categories, the mitigations recommended are sound. The common gap is identical: no standard mechanism exists to produce cryptographic proof that the recommended mitigations were continuously enforced during operation. Every category recommends controls. No category specifies how to prove those controls were active at the time of every agent decision.

SUMMARY TABLE

#	Threat Category	Evidence Gap	AGA Mechanism	Claim
T1	**Improper Authentication & Identity Mgmt**	No proof authorization enforced per-invocation	Subject Identifier binding; Portal evaluates every tools/call against sealed identity-to-tool binding	**1(b)**
T2	**Missing or Improper Access Control**	No proof least-privilege maintained at runtime	Sealed artifact enumerates authorized tools via allowlist; Portal denies unlisted tools; DRIFT for unknowns	**1(a)**
T3	**Input Validation/ Sanitization Failures**	No proof validation applied to every request	Portal validates parameters per sealed policy; phantom execution captures attack sequences	**1(f), 11**
T4	**Input/Instruction Boundary Distinction**	No proof enforcement boundary active during injection	Behavioral drift detection against sealed baseline; anomalous sequences trigger DENY + DRIFT	**1(e), 11**
T5	**Inadequate Data Protection & Confidentiality**	Audit vs. privacy tension; no per-request proof	Structural metadata linking: chain integrity from metadata only; payload excluded from chain linking	**3**

#	Threat Category	Evidence Gap	AGA Mechanism	Claim
T6	**Missing Integrity/ Verification Controls**	No proof runtime binary matches attested version	Agent identity hash compared to runtime hash on every measurement cycle; divergence triggers DRIFT	**1(b), 1(f)**
T7	**Session & Transport Security Failures**	No proof session boundaries enforced	Per-session sealed artifacts with TTL; expiry triggers re-attestation or safe-state transition	**6**
T8	**Network Binding/ Isolation Failures**	No proof isolation maintained during execution	Portal as PEP (NIST SP 800-207); two-process boundary with separate key material; TEE option	**1(e)**
T9	**Trust Boundary & Privilege Design Failures**	No proof constrained delegation maintained	Constrained sub-mandates: derived artifacts with TTL <= parent TTL, scope <= parent scope	**1**
T10	**Resource Mgmt/ Rate Limiting Absence**	No proof rate limits enforced	Rate limits sealed in artifact; Portal enforces at proxy layer; DENY receipt specifies violated limit	**1(a)**
T11	**Supply Chain & Lifecycle Security**	No proof supply chain controls active at runtime	Sealed artifact references attested binary hash; runtime measurement; per-tenant isolation	**1(b), 10**
T12	**Insufficient Logging, Monitoring & Audit**	Logs mutable, passive, producer-controlled	Hash-linked signed receipt chain; Merkle inclusion proofs; offline evidence bundles	**3, 9, 12**

DETAILED ANALYSIS

T1: Improper Authentication and Identity Management

CoSAI Mitigation. The white paper recommends end-to-end request traceability, SPIFFE/SPIRE for cryptographic workload identities, OIDC-based identity providers, Dynamic Client Registration, token exchange via RFC 8693, short-lived tokens with DPoP (RFC 9449), and Rich Authorization Requests (RFC 9396).

Enforcement Evidence Gap. These mitigations authenticate the agent at the point of entry and define access boundaries through token scopes. However, for autonomous AI agents, the gap between initial authorization and runtime behavior is where risk concentrates. An agent that is properly authenticated and initially authorized may subsequently attempt unauthorized tool invocations due to prompt injection, context poisoning, or emergent goal misalignment. No mechanism proves that least-privilege constraints were evaluated and enforced for every individual tool invocation throughout the session.

AGA Enforcement Mechanism. The sealed Policy Artifact binds the agent's cryptographic identity to authorized behavior through the Subject Identifier (Claim 1(b)). The Subject Identifier comprises at least one cryptographic hash of the agent's normalized bytes (executable image, model weights, container digest) and at least one cryptographic hash of canonicalized metadata (version, author, configuration manifest, creation timestamp). The enforcement Portal maintains a pinned public key for the policy issuer and rejects any artifact whose Ed25519 signature does not verify against the pinned key. Every `tools/call` request is evaluated against the sealed identity-to-tool binding.

> **Integration Point: SPIFFE/SPIRE.** The AGA Subject Identifier is complementary to SPIFFE/SPIRE. SPIRE handles node-to-workload identity: confirming that a specific container is running on a specific node in a specific cluster (transport-layer identity via SVID). AGA handles workload-to-intent governance: confirming that a specific agent instance has been attested against a specific policy and its runtime state matches its sealed reference (accountability-layer enforcement). In deployment, SPIRE issues the workload SVID; AGA binds governance parameters to that identity.

Continuous Authentication via TTL. Agent authorization is not a one-time event. The sealed artifact includes a time-to-live value (Claim 6). When TTL expires, the agent must re-attest to continue operating. Each re-attestation produces a new signed artifact and a corresponding continuity chain event. This implements continuous authentication at the governance layer: the agent's authority to act is periodically renewed, not permanently granted.

Evidence Artifact. Each receipt contains the agent's Subject Identifier reference, the tool invocation details, the policy artifact hash, and the PERMIT or DENY decision, all signed with Ed25519. The receipt chain provides per-invocation proof of authorization enforcement. The evidence bundle allows any third party to verify that identity-to-authorization binding was maintained for every action in the session.

T4: Input/Instruction Boundary Distinction Failure

CoSAI Mitigation. The white paper recommends input sanitization with guardrails and context isolation, clear boundaries between instructions and data, treating tool behavior descriptions as untrusted, and monitoring for hidden parameters or malicious defaults in schema definitions.

Enforcement Evidence Gap. Container isolation and process boundaries are runtime properties that produce no cryptographic evidence of their continuous operation. A container may start isolated and later have its boundaries weakened through configuration change, privilege escalation, or container escape. When isolation is breached through prompt injection or tool poisoning, current approaches provide no forensic capture mechanism: the compromised agent's actions are lost or mixed with legitimate activity in mutable logs.

AGA Enforcement Mechanism. The Portal operates as a Policy Enforcement Point (PEP) in NIST SP 800-207 (Zero Trust Architecture) terms. The agent cannot bypass the Portal to reach protected resources directly. Every tool invocation, API call, and data access passes through the Portal, which evaluates it against the sealed artifact's enforcement parameters (Claim 1(e)). The agent holds no signing keys and cannot self-authorize. For high-consequence deployments, the Portal itself should execute within a TEE such as Intel SGX or ARM TrustZone (Embodiment 6).

> **Phantom Execution for Post-Injection Forensics (Claim 11).** When behavioral drift is detected through anomalous tool invocation patterns, the Portal severs all connections to protected resources while the agent continues executing in a sandboxed environment. The agent receives no signal that it has been transitioned to phantom execution. Every unauthorized tool call, every attempted data access, every state change is signed and appended to the continuity chain as forensic receipts. This transforms a breach into an intelligence-gathering opportunity without exposing production systems.

Evidence Artifact. Each Portal mediation produces a signed receipt recording that the request passed through the enforcement boundary, was evaluated against the sealed policy, and received a PERMIT or DENY decision. During phantom execution, forensic receipts capture the full attack sequence. The evidence bundle packages both operational and forensic receipts.

T5: Inadequate Data Protection and Confidentiality

CoSAI Mitigation. The white paper recommends sandboxing and least privilege execution, cryptographic MACs and digital signatures, fine-grained authorization, Trusted Execution Environments (Intel TDX, AMD-SEV/SNP), and Confidential Containers with remote attestation.

Enforcement Evidence Gap. A fundamental tension exists between comprehensive governance auditing and data confidentiality. Logging every agent interaction to prove governance compliance potentially exposes sensitive model outputs, proprietary business logic, or PII contained in agent interactions. Current approaches force a choice: comprehensive audit (privacy risk) or privacy protection (audit gap).

AGA Enforcement Mechanism. The architecture resolves this tension through structural metadata linking (Claim 3, privacy-preserving disclosure). The leaf hash for each continuity chain event is computed using only structural metadata: schema version, protocol version, event type, event identifier, sequence number, timestamp, and previous leaf hash. Payload data is deliberately excluded from the chain-linking computation. A third-party auditor can verify the complete integrity of the enforcement chain - confirming that every measurement was performed on schedule, every enforcement action was properly executed, and no receipts were omitted or reordered - all without ever seeing the contents of any agent interaction.

Evidence Artifact. The evidence bundle supports tiered disclosure. The structural verification layer (chain integrity, receipt sequence, Merkle proofs) is shareable with any auditor. The payload verification layer (individual event contents) is accessible only to authorized parties. Tool-level access controls sealed in the policy artifact determine which verification tier each auditor receives. This enables compliance verification

without confidentiality compromise in multi-tenant deployments.

T8: Network Binding and Isolation Failures

CoSAI Mitigation. The white paper recommends network segmentation, localhost binding, DNS rebinding protection, payload limits, integrity checks, mutual TLS, CORS controls, and CSRF protection. Stdio transport is recommended for local MCP deployments.

Enforcement Evidence Gap. Network isolation is enforced by infrastructure and produces no application-layer governance evidence. If an agent connects to a shadow MCP server through DNS rebinding, the legitimate server's logs show nothing. Network-layer controls cannot prove that every agent interaction occurred with the authorized endpoint.

AGA Enforcement Mechanism. The Portal is the sole network path between the agent and MCP servers (Claim 1(e)). The sealed artifact specifies the authorized upstream by identity (URL, TLS certificate fingerprint, service mesh identity). The Portal validates upstream identity on every connection. Even if network isolation fails, the Portal refuses traffic to any server not matching the sealed reference. For Kubernetes deployments, the admission webhook (failurePolicy: Fail) prevents pods from starting without a sealed artifact.

Evidence Artifact. Signed mediation receipts prove every I/O request passed through the Portal enforcement boundary and was directed to an authorized upstream. The evidence bundle includes both Portal-layer and sidecar-layer receipts, enabling independent verification that network isolation was maintained at both the application and infrastructure levels.

T12: Insufficient Logging, Monitoring, and Auditability

CoSAI Mitigation. The white paper recommends comprehensive logging at MCP host, client, and server levels; immutable audit records; centralized logging via gateways/proxies; identity provider token exchange for accountability; and OpenTelemetry for end-to-end linkability.

Enforcement Evidence Gap. This category has the most direct and consequential evidence gap. The white paper recommends "immutable records of actions and authorizations" and "comprehensive logging," but standardized audit logging across MCP implementations does not yet exist. More fundamentally, logs have three structural limitations: (1) Logs are mutable - an attacker who compromises the logging infrastructure can alter entries without detection. (2) Logs are passive - they record events but do not prove an enforcement boundary existed. (3) Logs are producer-controlled - the system generating the log determines its completeness.

AGA Enforcement Mechanism. Signed enforcement receipts replace mutable logs with a cryptographic proof structure. Each receipt is signed by the Portal's Ed25519 private key over the SHA-256 hash of the JCS-canonicalized (RFC 8785) receipt content. Receipts are hash-linked to their predecessors via SHA-256 of structural metadata, forming an append-only continuity chain (Claim 3, Claims 3(d–f)). Modification of any receipt invalidates all subsequent hash links. Omission of a receipt is detectable via Merkle inclusion proofs (Claim 9). The governed agent cannot forge receipts because it does not hold the signing keys.

> **Merkle Checkpoint Anchoring (Claims 3(d–f)).** Receipt chains are periodically checkpointed by computing a Merkle root over batched events (Claim 12). The Merkle root serves as a compact

commitment to the complete set of receipts in the batch. If an attacker gains full access to the local system and attempts to rewrite history, the anchored checkpoint proves what the chain contained before the compromise.

Evidence Artifact. The evidence bundle packages the sealed artifact, complete receipt chain, Merkle inclusion proofs, checkpoint references, and the Portal's public key into a portable verification unit (Claim 9). Verification uses standard Ed25519 signature checking and SHA-256 hash computation. No network callbacks. No proprietary tooling. The verification is a deterministic mathematical operation producing the same PASS/FAIL result regardless of where or when it is executed.

STANDARDS ALIGNMENT

The enforcement evidence mechanisms described in this mapping align with the following frameworks. This alignment is architectural, not certification. The AGA framework does not claim compliance with these standards. It implements mechanisms that support the objectives these standards describe.

Standard	Alignment Point
NIST SP 800-207 (Zero Trust)	Portal operates as Policy Enforcement Point (PEP); sealed artifact serves as PDP payload
NIST SP 800-218 (SSDF)	Automated runtime integrity verification (PS.3); forensic data collection during incidents (RV.1)
NIST SP 800-204 (Microservices)	Extends service mesh security from communication channel to agent intent governance
NIST AI RMF	Operationalizes Measure function (continuous verification) and Manage function (autonomous enforcement)
SLSA	Sealed artifacts extend build-time provenance to runtime governance; receipt chains extend attestation post-deployment
in-toto Attestation Framework	Governance receipts expressible as in-toto attestations with runtime governance predicate type
OWASP MCP Top 10	Addresses MCP02 (scope), MCP03 (tool poisoning), MCP07 (authentication), MCP08 (audit trails)

Algorithm Agility. The architecture uses Ed25519 and SHA-256 as current defaults. The cryptographic layer is designed for algorithm agility: post-quantum signature schemes (ML-DSA, SLH-DSA) and hash functions are drop-in replacements within the existing sealed artifact and receipt structures. No architectural changes are required for PQC migration.

CONCLUSION

The CoSAI WS4 threat taxonomy provides a comprehensive foundation for MCP security. The cryptographic enforcement evidence layer described in this mapping does not replace the mitigations the white paper recommends - it provides proof that they were applied.

For each of the 12 threat categories, the gap between "mitigation recommended" and "mitigation provably enforced" can be closed with sealed policy artifacts, signed enforcement receipts, and offline-verifiable

evidence bundles using standard cryptographic primitives (Ed25519, SHA-256/BLAKE2b, Merkle trees).

The white paper notes that “follow on papers will provide reference implementations and recommendations for specific mitigation controls.” This mapping is intended as input toward that objective. We welcome feedback on the category mappings and are prepared to develop detailed implementation guidance for any categories the working group prioritizes.

A reference implementation with 231+ automated tests and an independent verifier (zero framework imports, standard cryptographic libraries only) is available for evaluation. The four-phase laboratory demonstration proposed in our NCCoE submission (attestation, authorized operation, simulated compromise with phantom execution, offline audit) provides a concrete validation path for any category in this mapping.

Attested Intelligence Holdings LLC

USPTO App. No. 19/433,835 (Patent Pending)

USPTO Trademark Serial No. 99677085

attestedintelligence.com