MolTrust — NIST AI RMF Function Mapping

Function-by-Function Technical Mapping for Agentic AI Trust Infrastructure

Version 1.0 · April 2026

Lars Kersten Kroehl

MolTrust / CryptoKRI GmbH, Zurich

lars@moltrust.ch ·moltrust.ch

---

1. Scope and Positioning

This document maps the MolTrust Protocol — a W3C Verifiable Credentials and Decentralized Identifier trust infrastructure for autonomous AI agents — against the four core functions of the NIST AI Risk Management Framework (AI RMF 1.0, NIST AI 100-1, January 2023): **Govern**, **Map**, **Measure**, and **Manage**.

What this document is: a *technical function mapping* showing where MolTrust primitives operationally support the AI RMF functions, and how they integrate with the broader NIST agent-identity ecosystem — including the NIST Center for AI Standards and Innovation (CAISI) AI Agent Standards Initiative launched 17 February 2026, the National Cybersecurity Center of Excellence (NCCoE) concept paper *Accelerating the Adoption of Software and AI Agent Identity and Authorization* (5 February 2026), and the NIST Control Overlays for Securing AI Systems (COSAiS) project.

What this document is not: it is not a formal NIST compliance attestation, and it does not claim completeness against every AI RMF subcategory. The AI RMF is a framework for risk management governance at the organizational level; MolTrust operates at the infrastructure layer. The mapping identifies where infrastructure-level primitives provide structural support for organization-level governance processes.

Reference implementation scope: the MolTrust reference implementation is operated by a three-person team under CryptoKRI GmbH (Zurich). The scope of this mapping is the infrastructure layer; deployer organizations provide their own operational capacity, compliance attestation, and organizational-governance processes.

Scope of the mapping: the document focuses on the AI RMF subcategories most directly applicable to autonomous AI agents deployed in cross-organizational settings. The AI RMF Companion Playbook enumerates specific subcategories (GV-1.1 through MG-4.3) with suggested actions; this mapping references those subcategories explicitly and marks the ones

where MolTrust provides direct technical evidence.

Relationship to NIST agent-specific work: the AI RMF itself does not address autonomous agents specifically — it was written before production agentic AI deployment. The NIST CAISI Initiative and the NCCoE concept paper address the agent-specific gap. This document treats the AI RMF as the framing layer and the agent-specific NIST work as the operational extension.

Reference versions: NIST AI RMF 1.0 (NIST AI 100-1), January 2023; NIST AI RMF Playbook (February 2023); NIST NCCoE Concept Paper, February 2026; NIST CAISI AI Agent Standards Initiative announcement, 17 February 2026.

---

2. Methodology

For each of the four AI RMF core functions, the mapping is presented in two layers:

Layer 1 — Function-level alignment. How MolTrust primitives (DIDs, VCs, IPRs, AAE, Trust Score) contribute to the function's stated outcomes.

Layer 2 — Subcategory-level evidence. Specific AI RMF subcategories where MolTrust produces technical evidence artifacts that feed into organizational governance processes. Subcategories are referenced by their canonical codes (e.g. GV-1.1, MS-2.5).

Each mapping entry uses the following status taxonomy:

- *live*: deployed, operationally exercised, and verifiable as of the document date
- *live**: capability deployed in the reference implementation but not yet exercised against real adversarial events or operational incidents; the mechanism exists, its performance under stress is pending empirical validation (the asterisk marks the gap between deployment and adversarial validation)
- *partial*: some components are live while others require deployer-side processes or are on the roadmap
- *roadmap*: planned within a stated timeframe
- *gap*: outside MolTrust's scope and a deployer responsibility

Where a subcategory describes an organizational process that cannot be replaced by infrastructure (e.g. staff training, internal policy review), MolTrust is marked as *gap*.

---

3. Function 1 — GOVERN

Function outcome: "A culture of risk management is cultivated and present."

The Govern function defines policies, processes, roles, and accountability structures for AI risk management. It is the foundation for the other three functions. MolTrust does not replace organizational governance — it provides cryptographically verifiable primitives that governance processes can rely on.

3.1 Function-Level Alignment

MolTrust supports the Govern function in three structural ways:

Accountability chain clarity. The MolTrust Five-Party Trust Chain (Developer → Owner → Agent → Instructor → Counterparty, documented in arXiv Preprint v1.0 Section 3.2) formalizes the accountability relationships in autonomous agent deployments. The chain is cryptographically verifiable: each link carries its own DID and Verifiable Credentials, and the binding between links is signed. This gives governance functions a concrete basis for assigning responsibility when incidents occur.

Policy as machine-readable artifact. Agent Authorization Envelopes (AAE) encode governance decisions (permitted actions, constraints, oversight thresholds) as machine-readable, cryptographically signed artifacts. This moves policy from PDF documents read by humans to structured data consumed by runtime systems — which is a prerequisite for any governance approach that operates at the scale of autonomous agent deployment.

Audit trail immutability. Interaction Proof Records anchored on Base Layer 2 provide an immutable record of agent behavior. Governance processes that depend on historical behavior analysis (incident investigation, periodic review, compliance reporting) can reference IPRs as ground truth rather than relying on logs controlled by the party being reviewed.

3.2 Subcategory-Level Evidence

GV-1.1 — Legal and regulatory requirements involving AI are understood, managed, and documented.

MolTrust contribution: the protocol supplies primitives that deployers can reference when producing their own legal-regulatory documentation — DID-based agent identity for per-agent regulatory traceability, AAE for machine-readable encoding of jurisdictional constraints (regulatory-boundary-aware validity), and on-chain anchoring of published specifications for integrity-verifiable reference artifacts. The organizational understanding, management, and documentation of legal requirements remains a deployer responsibility; MolTrust supplies components, not the compliance artefact itself. *Status: partial.*

GV-1.4 — The risk management process and its outcomes are established through transparent policies, procedures, and other controls based on organizational risk priorities.

MolTrust contribution: the AAE structure encodes organizational risk priorities into machine-readable authorization objects. Risk priorities that would otherwise remain as internal policy documents become enforceable constraints. Examples: financial thresholds encoding risk appetite for autonomous transactions, jurisdictional restrictions encoding regulatory-boundary-aware risk limits, counterparty minimum trust score encoding counterparty-risk tolerance. *Status: live.*

GV-1.7 — Processes and procedures are in place for decommissioning and phasing out AI systems safely and in a manner that does not increase risks or decrease the organization's trustworthiness.

MolTrust contribution: credential revocation via CAEP-compatible endpoint with 60-second propagation in the reference implementation supports the phase-out process at the credential level. Principal-DID-linked Violation Records persist across agent re-registrations, which prevents the class of phase-out-and-rebrand attacks where a decommissioned agent returns under a new identity. *Status: live*.*

GV-3.1 — Decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team.

MolTrust contribution: gap. This is an organizational composition requirement that no infrastructure layer can satisfy. MolTrust documentation acknowledges this explicitly.

GV-4.1 — Organizational policies and practices are in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize potential negative impacts.

MolTrust contribution: organizational culture cannot be engineered through infrastructure. However, the AAE default-deny semantics (any action not explicitly permitted is denied, and explicit denial takes precedence over any allow) encode a safety-first default into the authorization layer. *Status: partial — construction defaults only.*

GV-5.1 — Organizational policies and practices are in place to address AI risks and benefits arising from third-party software and data and other supply chain issues.

MolTrust contribution: the W3C Verifiable Credentials infrastructure supports attestations from third-party issuers with cryptographic chain-of-custody. Supply-chain attestations (provenance, quality, ownership) can be issued by relevant third parties and verified independently. `ProductProvenanceCredential` is one of the enumerated credential verticals. *Status: live.*

GV-6.1 — Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third party's intellectual property or other rights.

MolTrust contribution: gap on IP policy, live on third-party verification. MolTrust provides the technical primitive for verifying that a third party is who they claim to be (DID resolution + VC verification). IP policy itself remains an organizational commitment.

---

4. Function 2 — MAP

Function outcome: "Context is established and understood. Risks to the AI system are identified."

The Map function establishes the context in which an AI system operates and identifies the risks specific to that context. Map produces the risk register that Measure and Manage operate against.

4.1 Function-Level Alignment

MolTrust supports Map through four primitives:

Context as explicit metadata. AAE CONSTRAINTS blocks encode operational context explicitly: jurisdictions where the agent is permitted to operate, time windows of validity, financial thresholds above which additional oversight applies, counterparty minimum trust score. Context that would otherwise be implicit (or documented only in prose) becomes structured and queryable.

Scope boundaries as first-class fields. The MANDATE block enumerates permitted purposes

(`commerce`, `data_read`, `data_write`, `communication`, `delegation`, `administration`), allowed action URI patterns, and explicitly denied actions (which take precedence). Risk identification becomes the process of comparing required operational scope against permitted scope — and when the comparison produces a gap, the gap is visible at the infrastructure level.

Behavioral baseline for risk detection. The Trust Score and its component signals (direct endorsements, propagated endorsements, cross-vertical diversity, interaction bonus, sybil penalty, inactivity penalty) provide a behavioral baseline against which anomalies can be detected. An agent whose Trust Score drops significantly, or whose score is withheld due to insufficient cross-vertical endorsements, is a risk signal that feeds into the Map function's risk identification output.

Third-party attestation ingestion. Verifiable Credentials issued by external parties — including skill verification credentials, provenance attestations, regulatory compliance attestations — are all cryptographically verifiable inputs to the risk-mapping process.

4.2 Subcategory-Level Evidence

MP-2.3 — AI system's knowledge limits and how system output may be utilized and overseen by humans is documented.

MolTrust contribution: AAE `deniedActions` and `obligations.toolAllowlist` make the agent's operational scope explicit. Human oversight thresholds are encoded in CONSTRAINTS (autonomous threshold, step-up threshold, approval threshold) with machine-readable semantics. *Status: live.*

MP-3.4 — Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from Govern function.

MolTrust contribution: the approval threshold mechanism in AAE CONSTRAINTS provides the enforcement point for human oversight. Crossing the threshold without human authorization produces a Violation Record. Assessment and documentation of the oversight process itself remains a deployer responsibility, but the technical enforcement gate is standardized and auditable. *Status: partial — live for enforcement, gap on assessment/documentation process.*

MP-4.1 — Approaches for mapping AI technology and legal risks of its components — including the use of third-party data or software — are in place, followed, and documented, as are risks of infringement of a third party's intellectual property or other rights.

MolTrust contribution: DID-based identity for agents and their components (including third-party libraries attested via VCs) supports technology component mapping. Legal risk mapping remains an organizational responsibility. *Status: partial.*

MP-5.1 — Likelihood and magnitude of each identified risk based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.

MolTrust contribution: Trust Score components (endorsements, interaction proofs, violation records) provide quantitative inputs into risk likelihood assessment. The endorsement graph can be queried to identify similar agents in similar contexts. Interaction Proof Records provide one component of past-use data (interaction records) in cryptographically verifiable form; operational

telemetry, user feedback, and public incident reports remain deployer-provided. *Status: partial.*

MP-5.2 — Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.

MolTrust contribution: gap. Stakeholder engagement practices are organizational. MolTrust's endorsement mechanism provides a structured way for third parties to signal trust or concerns about an agent, but converting that signal into organizational feedback loops remains a deployer responsibility.

---

5. Function 3 — MEASURE

Function outcome: "Identified risks are assessed, analyzed, or tracked."

The Measure function quantifies risks identified in Map and tracks them over time. It is the function most naturally aligned with cryptographically verifiable evidence production.

5.1 Function-Level Alignment

Measure is where MolTrust's production-deployed evidence infrastructure provides the most direct operational support:

Quantitative behavioral metrics. Trust Score (0–100, A–F grade) is a continuous quantitative measure of agent behavior. Sub-components (direct endorsements, propagated endorsements, cross-vertical diversity bonus, interaction bonus) provide decomposable risk signals.

Verifiable event records. Interaction Proof Records provide timestamped, dual-signed records of agent interactions. They are the raw material for behavioral analysis and statistical monitoring.

On-chain anchoring for longitudinal analysis. Base L2 anchoring ensures that IPRs, AAE changes, and Violation Records cannot be modified retrospectively. Longitudinal studies of agent behavior can rely on on-chain evidence as ground truth.

Multi-source risk signals. Trust Score, Violation Records, Skill Audit results, and A2A discovery scan results are independent signals that can be correlated for more robust risk measurement. Correlation is the deployer's analytical work; MolTrust provides the signals in machine-consumable form.

5.2 Subcategory-Level Evidence

MS-1.1 — Approaches and metrics for measurement of AI risks enumerated during the Map function are selected for implementation starting with the most significant AI risks.

MolTrust contribution: Trust Score formula is publicly documented (arXiv Preprint v1.0 Section 3.1, Technical Specification v0.8 Section 4). Weights are explicit and reasoning is published. Deployers can evaluate whether the formula components map to their identified risks, and can configure threshold values (e.g. counterpartyMinScore in AAE) based on their risk tolerance. *Status: live.*

MS-2.1 — Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented.

MolTrust contribution: CONFORMANCE.md v1.0 documents five reproducible test vectors (TV-001 through TV-005) that any party can run against the live endpoint to verify protocol compliance. Test artifacts are version-pinned and SHA-256 anchored. *Status: live.*

MS-2.5 — The AI system to be deployed is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.

MolTrust contribution: the arXiv Preprint v1.0 Section 7 ("Discussion and Limitations") explicitly documents what the three Claim A capabilities provide and what they do not. This is the analog at the infrastructure layer of the validity documentation required at the AI system layer. *Status: live.*

MS-2.7 — AI system security and resilience — as identified in the MAP function — are evaluated and documented.

MolTrust contribution: Skill Audit infrastructure evaluates nine specific security properties (secrets scanning CWE-798, A2A discovery integrity, endpoint reachability, signature validity, etc.) and documents results in machine-readable form via `GET /guard/audit/checks`. *Status: live.*

MS-2.8 — Risks associated with transparency and accountability — as identified in the MAP function — are examined and documented.

MolTrust contribution: the transparency primitives (DID Document, AgentCard, AAE public inspection, Trust Score publicly queryable) directly address the transparency subcategory. Accountability is supported through the Five-Party Trust Chain (documented, machine-readable, cryptographically verifiable). *Status: live.*

MS-2.13 — Effectiveness of the employed TEVV metrics and processes in the MEASURE function are evaluated and documented.

MolTrust contribution: partial. The conformance test vectors themselves (TV-001 through TV-005) are evaluated on every release via automated CI. Effectiveness evaluation of the metrics as measurement tools for actual risk reduction requires longitudinal deployment data that is still accumulating. *Status: roadmap — measurement effectiveness report planned Q3 2026.*

MS-3.1 — Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts.

MolTrust contribution: Behavioral Anomaly Detection in the Swarm Intelligence Protocol flags agents whose confidence distribution shows systematic inflation (mean > 0.95 or standard deviation < 0.02). The Jaccard cluster detection heuristic identifies coordinated endorsement patterns that may indicate Sybil activity. These detection mechanisms are deployed in the reference implementation but have not yet been exercised against real adversarial adaptation; as the arXiv Preprint Section 7.2 states, robustness at adversarial scale is pending empirical validation. *Status: live*.*

MS-4.2 — Measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle are informed by input from domain experts and other relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.

MolTrust contribution: the cross-vertical endorsement requirement (minimum three distinct verticals for a non-seed agent's score to be published) supplies a structured proxy for domain-expert input — endorsements from distinct verticals raise the structural diversity of attestations, but are not substitutes for formal domain-expert review. Agents lacking this diversity have their scores withheld. *Status: partial.*

---

6. Function 4 — MANAGE

Function outcome: "Risks are prioritized and acted upon based on a projected impact."

The Manage function operates on the outputs of Measure: it decides which risks require action, allocates resources to mitigation, and monitors the effect of mitigations over time.

6.1 Function-Level Alignment

MolTrust supports Manage through four mechanisms:

Automated enforcement as primary mitigation. Many mitigations in autonomous agent deployment must be automated because the pace of agent action exceeds human oversight capacity. The three-layer enforcement model (cryptographic, API, kernel) provides automated mitigation across the agent lifecycle, with Layer 3 (Falco eBPF) providing enforcement below the agent process boundary.

Revocation as a first-class operation. CAEP-compatible revocation propagates to verifiers within 60 seconds in the reference implementation. Credential revocation is the primary mitigation mechanism for compromised agents, and its operational characteristics (propagation time, failure mode — fail-closed by default — notification pattern) are specified explicitly.

Violation persistence as behavioral consequence. Principal-DID-linked Violation Records persist across agent re-registrations. An agent cannot shed behavioral consequences by rotating identities. This strengthens mitigation effectiveness by making "start fresh" a non-option.

Partner ecosystem for mitigation composition. MolTrust integrates with payment protocol mitigation gates (x402, MPP), cross-protocol verification harnesses (APS, qntm Authority Constraints), and A2A governance consumer-side vocabularies. Mitigation at the MolTrust layer composes with mitigations at adjacent layers.

6.2 Subcategory-Level Evidence

MG-1.1 — A determination is made as to whether the AI system achieves its intended purpose and stated objectives and whether its development or deployment should proceed.

MolTrust contribution: partial. Trust Score and Violation Records provide evidence about whether an agent is operating consistently with its stated authorization. The go/no-go decision itself is organizational. *Status: partial.*

MG-2.3 — Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.

MolTrust contribution: credential revocation (API endpoint, CAEP-compatible, 60-second prop-

agation) provides the supersede/disengage mechanism. Responsibility assignment — who decides to revoke — remains organizational. *Status: live*.*

MG-2.4 — Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI actors.

MolTrust contribution: on-chain anchoring of every version update (Whitepaper, Technical Specification, arXiv Preprint) provides the integrity record of continual improvement activity. Version-pinned conformance test vectors provide the regression baseline for update assessment. *Status: live.*

MG-3.1 — AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.

MolTrust contribution: third-party credential issuers are monitored through the Trust Score mechanism applied to issuer DIDs. Credentials from issuers with degraded trust scores can be filtered by deployers. *Status: partial.*

MG-4.1 — Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.

MolTrust contribution: multiple primitives contribute:

- Input capture: endorsement mechanism (structured third-party input)
- Appeal/override: revocation with fail-closed default
- Decommissioning: credential revocation with Violation Record persistence
- Incident response: IPR queries provide timestamped interaction history
- Recovery: agent DID rotation with explicit migration support (principal continuity maintained)
- Change management: AAE versioning with issuance/expiry semantics

Status: live.*

MG-4.3 — Mechanisms are in place and applied to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes consistent with harm or risk to people.

MolTrust contribution: see MG-2.3. Additionally, the Violation Record mechanism creates a permanent audit trail of disengagement events, supporting accountability after the fact. *Status: live*.*

---

7. Integration with NIST Agent-Specific Work

The AI RMF does not directly address autonomous agents. Three NIST initiatives extend the framework into the agent-specific space, and MolTrust integrates with each.

7.1 NIST CAISI AI Agent Standards Initiative (17 February 2026)

Organized around three pillars:

Pillar 1 — Industry-led standards development. MolTrust is already built on W3C standards

(DIDs, VCs) and uses the widely-deployed Ed25519 and RFC 8785 canonicalization. Integration with A2A, x402, and MPP payment protocols is live. This positions MolTrust as a standards-conformant reference implementation that U.S. industry can point to when contributing to international standards bodies.

Pillar 2 — Community-led open-source protocol development. The MolTrust reference implementation is published at `github.com/MoltyCel/moltrust-protocol`. Conformance specifications are public and SHA-256 anchored. The `@moltrust/sdk`, `@moltrust/x402`, `@moltrust/mpp`, `@moltrust/aae`, and `@moltrust/openclaw` packages are published on npm under open-source licensing.

Pillar 3 — Fundamental research in agent security and identity infrastructure. The MolTrust arXiv Preprint v1.0 contributes to this pillar with a production-deployment-first perspective. The paper is peer-review-submitted (cs.AI, endorsed via arXiv endorsement process).

7.2 NCCoE Concept Paper on Agent Identity and Authorization (5 February 2026)

The concept paper identifies that "AI agents are commonly treated as generic service accounts without dedicated identity, authorization, or accountability controls." MolTrust directly addresses this gap:

- **Identity:** did:moltrust (or any W3C-conformant DID method with Ed25519 support) provides per-agent identity distinct from service-account credentials.
- **Authorization:** AAE provides machine-evaluable, scoped, time-bounded authorization structured around MANDATE/CONSTRAINTS/VALIDITY blocks.
- **Accountability:** IPRs + Violation Records + Principal-DID continuity provide the accountability chain.

The paper identifies OAuth 2.0, OpenID Connect, and SPIFFE/SPIRE as foundation identity standards requiring extension. MolTrust is architecturally compatible with all three but is not a drop-in replacement. Concrete integration points: AAE plays the same operational role as an OAuth 2.0 bearer token (a bounded, time-scoped authorization object issued by an authority), with richer MANDATE/CONSTRAINTS structure; DID Document resolution supplies the identifier-to-key mapping that OpenID Connect Discovery supplies for IdPs; a SPIRE-bridge for exchanging MolTrust Verifiable Credentials for SPIFFE IDs is planned for Protocol v0.9.

7.3 NIST Control Overlays for Securing AI Systems (COSAiS)

The COSAiS project develops SP 800-53 control overlays specifically for single-agent and multi-agent AI deployments. As of the version date of this document, COSAiS overlays have not been finalized. MolTrust's mapping to SP 800-53 controls will be published once the COSAiS overlays are available, to ensure the mapping references the authoritative agent-specific control interpretations rather than the general SP 800-53 text. *Status: roadmap — formal SP 800-53 mapping pending COSAiS publication.*

---

8. Honest Gaps and Deployer Responsibilities

Organizational governance processes. The AI RMF is fundamentally about organizational governance. Several Govern function subcategories describe processes (team diversity, cultural commitments, policy review cadence) that cannot be satisfied by infrastructure. MolTrust documentation does not claim otherwise.

Measurement effectiveness evaluation. MS-2.13 requires evaluation of whether TEVV metrics are effective at reducing risk. This requires longitudinal deployment data from the metrics' application. MolTrust has been in production since March 2026; meaningful effectiveness evaluation requires additional deployment duration and is on the Q3 2026 roadmap.

Adversarial validation at scale. Subcategories marked *live** (GV-1.7, MS-3.1, MG-2.3, MG-4.1, MG-4.3) fall into this class: the capabilities are deployed in the reference implementation but have not yet been exercised against real adversarial events or coordinated attacks. The arXiv Preprint Section 7.2 explicitly identifies adversarial validation as pending work, and the Sybil-Resistance Methodology Note (companion document) details the three structural mechanisms whose empirical stress-testing is scheduled for Q3 2026.

Incident response workflow. MG-4.1 mentions incident response as part of the monitoring plan. MolTrust provides the evidence primitives for incident response (IPR queries, Violation Records, revocation mechanism) but does not prescribe an incident response workflow. The workflow — including roles, escalation paths, communication protocols, and remediation procedures — remains the deployer's responsibility.

Revocation-endpoint availability. CAEP-compatible revocation with fail-closed default is a reference-implementation design choice, not a cryptographic guarantee. A denial-of-service attack against the revocation endpoint degrades "fail-closed" into "deny-all" — converting an integrity mechanism into an availability-attack surface. Deployers must operate the revocation endpoint with availability SLAs appropriate to their threat model, or explicitly accept the fail-closed-as-deny-all failure mode as an operational risk.

Formal compliance attestation. This document is a technical mapping, not a formal attestation of NIST AI RMF compliance. Deployers using MolTrust as part of an AI RMF-aligned governance program must produce their own compliance documentation integrating infrastructure-level evidence with organizational governance evidence.

---

9. References

NIST reference documents:

- NIST AI 100-1, Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, January 2023.
- NIST AI RMF Playbook. National Institute of Standards and Technology, February 2023.
- NIST NCCoE Concept Paper, *Accelerating the Adoption of Software and AI Agent Identity and Authorization*. National Institute of Standards and Technology, 5 February 2026.
- NIST CAISI AI Agent Standards Initiative announcement. National Institute of Standards and Technology Center for AI Standards and Innovation, 17 February 2026.

- NIST SP 800-53 Rev. 5, Security and Privacy Controls for Information Systems and Organizations. National Institute of Standards and Technology, September 2020 (with ongoing COSAiS overlay development).
- NIST SP 800-162, Guide to Attribute Based Access Control (ABAC) Definition and Considerations. National Institute of Standards and Technology, January 2014.

Parallel regulatory frameworks:

- Regulation (EU) 2024/1689 (EU AI Act). Official Journal of the European Union, 12 July 2024. See MolTrust EU AI Act Mapping (companion document).
- Infocomm Media Development Authority (IMDA) of Singapore, Model AI Governance Framework for Agentic AI, 22 January 2026.

MolTrust references:

- MolTrust Protocol Technical Specification v0.8, SHA-256 anchored on Base L2 Mainnet, Block 44745864.
- MolTrust Protocol Whitepaper v0.8, SHA-256 anchored on Base L2 Mainnet, Block 44507827.
- MolTrust arXiv Preprint v1.0, SHA-256 `c9c349852dbc77b80c4d8d3b0f9e3db60244a2bc60345b6fc43de1dc1` anchored on Base L2 Mainnet, TX `0x49a548346f12ee0a273cf9d4d60f00685777d88050df06ef77d92caaff` Block 45,037,732, tag `MolTrust/arXiv/v1.0`.
- CONFORMANCE.md v1.0 (MolTrust GitHub repository).

Technical standards:

- W3C Decentralized Identifiers (DIDs) v1.0, W3C Recommendation, 19 July 2022.
- W3C Verifiable Credentials Data Model v2.0, W3C Recommendation, 15 May 2025.
- RFC 8785, JavaScript Object Notation (JSON) Canonicalization Scheme (JCS). IETF, June 2020.
- RFC 9396, OAuth 2.0 Rich Authorization Requests. IETF, May 2023.
- Ed25519Signature2020, W3C Credentials Community Group.
- SPIFFE (Secure Production Identity Framework For Everyone), CNCF.