MolTrust — Sybil-Resistance Methodology Note

Version 1.3 · April 2026

Lars Kersten Kroehl

MolTrust / CryptoKRI GmbH, Zurich

lars@moltrust.ch ·moltrust.ch

Companion to: MolTrust — EU AI Act Mapping and MolTrust — NIST AI RMF Mapping

---

Abstract

This methodology note documents the Sybil resistance design of the MolTrust Protocol: three compounding mechanisms — dual-signature Interaction Proof Records, cross-vertical endorsement diversity gating, and principal-DID-linked violation persistence — complemented by Jaccard cluster detection and an economic audit-trail layer. The note accompanies the MolTrust EU AI Act Mapping and NIST AI RMF Mapping documents and draws its formal architecture from Section 6 of the MolTrust arXiv preprint v1.0. The canonical preprint (SHA-256 prefix `c9c34985`, suffix `1c11d946`) is anchored on Base Layer 2 Mainnet at Block 45,037,732; the full hash is recoverable from the anchoring transaction at that block and is also served from `moltrust.ch/publications/integrity.html`. This note is intended for analysts, implementers, and reviewers who require a standalone account of the Sybil resistance argument without reading the full preprint. Limitations are stated explicitly and are bounded by the current operational scale (54 agents, 57 endorsements, two months of production operation). This note is not a security proof and does not constitute an independent adversarial evaluation.

---

1. Scope and Positioning

This methodology note covers one specific slice of the MolTrust Protocol: the set of mechanisms designed to resist identity-fabrication and endorsement-manipulation attacks — collectively referred to as Sybil resistance after Douceur's 2002 formulation of the problem in permissionless peer-to-peer networks [1]. The note is a standalone companion to two sibling documents: the EU AI Act Mapping, which maps MolTrust primitives to the regulation's high-risk system obligations; and the NIST AI RMF Mapping, which maps the same primitives to the four functions of the NIST AI Risk Management Framework.

Sybil resistance is relevant across both regulatory documents. Under the EU AI Act [13], Article 14 (Human Oversight) and Article 15 (Cybersecurity, Accuracy and Robustness) imply that the underlying trust signals — including reputation-derived scores — must themselves be resistant to coordinated manipulation. Under the NIST AI RMF and the adjacent NIST CAISI AI Agent Standards Initiative [12], the MEASURE function requires meaningful metrics; a trust score that can be inflated through coordinated fake endorsements is not a meaningful metric. A third convergent regulatory reference point is the Singapore IMDA Model AI Governance Framework for Agentic AI [11], published in January 2026, which

shares the structural concern with coordinated-manipulation resistance even though IMDA's specific vocabulary differs from the EU AI Act and NIST AI RMF frameworks. This note therefore addresses a shared structural concern across all three mappings rather than replacing any of them.

The note is **not** a security proof. It does not establish formal bounds on adversarial success probability, and it does not report results from red-team campaigns conducted against the live registry. What it documents is the architecture, the design rationale behind each mechanism, and the honest limitations of the current deployment. Where mechanisms are operational as of April 2026, they are labeled *live*. Where mechanisms are specified and scheduled but not yet deployed, they are labeled *roadmap*. Where capability is explicitly outside the scope of the current design, it is labeled *gap*. This taxonomy is identical to the companion documents and permits line-item cross-referencing.

The target audience is threefold: institutional analysts and research bodies evaluating MolTrust against the converging regulatory requirements described in Section 2.1 of the MolTrust preprint [2]; developers implementing MolTrust-conformant verifiers who need to understand why specific thresholds exist; and reviewers auditing the protocol's claims against its published conformance specification. The note assumes familiarity with W3C Decentralized Identifiers v1.0 [3] and the Verifiable Credentials Data Model 2.0 [4] but does not require prior reading of the arXiv preprint — all architectural components necessary for the Sybil resistance argument are summarized where referenced.

Finally, the note does not constitute legal advice and is not a conformity assessment. Deploying organizations remain responsible for the operational, legal, and accountability requirements that apply to their own agent deployments. The MolTrust Protocol is an infrastructure layer, not a substitute for organizational compliance processes.

1.1 Positioning Against Established Sybil Defenses

The Sybil resistance mechanisms described in this note occupy one position in a design space populated by several well-established approaches. The comparative typology below, structured along the lines of the inter-agent trust-model taxonomy in Hu and Rong [10], positions MolTrust against those approaches across four dimensions: the type of cost the approach imposes on a Sybil adversary, whether the approach is decentralized, whether it is compatible with autonomous agent identities as distinct from human user identities, and the principal known limitation of the approach.

Approach	Sybil Cost Mechanism	Decentralized	Agent-Compatible	Principal Limitation
Proof-of-Work (Bitcoin)	Computational cost per unit of influence	Yes	No — not an identity primitive	Energy-intensive; identity-free by design
Proof-of-Stake (Ethereum)	Capital at risk per validator	Yes	Partial — stake held per agent	Plutocratic; stake-slashing does not map to reputation
SybilGuard / SybilLimit	Random walks on a trusted social graph	Yes	No — requires human social graph	Not applicable to new agent ecosystems
Proof-of-Personhood (Worldcoin)	Biometric uniqueness	Partial	No — human identity only	Not applicable to software agents

Approach	Sybil Cost Mechanism	Decentralized	Agent-Compatible	Principal Limitation
Gitcoin Passport	Stamp accumulation across issuers	Partial	Partial — credential ecosystem human-centric	Issuer trust concentrated; stamp cost adversary-dependent
MolTrust	**Layered: cryptographic + structural + audit**	**Anchored on Base L2**	**Yes — designed for autonomous agents**	**Scale unvalidated (Section 6)**

No single approach dominates across all dimensions. MolTrust's positioning is specifically in the agent-identity quadrant: Proof-of-Work and Proof-of-Personhood do not apply; SybilGuard/SybilLimit requires prior social structure that agent ecosystems do not inherit; Proof-of-Stake mechanisms are difficult to align with trust reputation as distinct from economic staking. This positioning does not imply that MolTrust's mechanisms are strictly stronger than the alternatives listed; it implies that they address a problem setting the alternatives do not natively target.

---

2. Methodology and Status Taxonomy

Each mechanism described in this note is annotated with the same status taxonomy used in the companion EU AI Act and NIST AI RMF Mapping documents:

- *live*: deployed and verifiable against the reference endpoint `api.moltrust.ch` as of April 2026*
- *roadmap*: specified in the Technical Specification or on the published roadmap, with a defined target milestone
- *gap*: outside the scope of the current design and explicitly not covered by MolTrust

* The *live* label indicates operational status in the reference deployment. Independent conformant implementations may exist at different maturity levels; the claim is about the reference implementation, not about every conformant deployment.

A mechanism labeled *live* means the code path exists, is executing in production, and can be exercised against the public endpoint. It does **not** mean the mechanism has been validated under adversarial load. Sections 4.5 and 6 are explicit about the empirical validation gap that accompanies the current operational scale.

The primary source for all normative statements in this note is the MolTrust Protocol Technical Specification v0.8 [5], anchored on Base Layer 2 at Block 44,745,864. Where this note simplifies or reorganizes material for readability, the Technical Specification takes precedence in any case of conflict. Section 6 of the MolTrust arXiv preprint [2] provides the academic-style treatment of the same material; this note is a re-presentation at the granularity useful for regulatory and institutional review, with the Jaccard math and the compounding argument isolated as standalone subsections for independent reference.

---

3. The Three Mechanisms

MolTrust's Sybil resistance is organized around three mechanisms. Each addresses a distinct attack surface, and each is designed to compound with the others, in the sense formalized in Section 3.4: a defeat of any single mechanism does not defeat the system, because a successful Sybil campaign must

defeat all three simultaneously. No single mechanism below is claimed to be adversarially robust in isolation. The systematization of attack surfaces against autonomous agent systems in Mao et al. [9] places coordinated identity fabrication among the central open problems, which this section's design deliberately targets.

3.1 Mechanism 1 — Dual-Signature Interaction Proof Records

Status: *live*. Interaction Proof Records (IPRs) are the verifiable behavioral substrate of the protocol. Every recorded interaction between two agents produces an IPR carrying two distinct Ed25519 signatures: the initiator signs first, and the responder's signature covers the initiator's signature. The result is a sequential commitment chain that is non-repudiable by either party. Completed IPRs are Merkle batch-anchored on Base Layer 2 [5, §7].

The design choice is deliberate, but its defensive scope requires careful statement. Bilateral IPRs address a specific subset of the Sybil fabrication problem: they prevent fabrication of interaction evidence between an adversary-controlled agent and an agent whose keys are not under the adversary's control. Where at least one party to an IPR is independent of the adversary, forging the IPR requires compromise of two independent keypairs — a key-compromise problem rather than an identity-fabrication problem. The latter is the class of attack that underpins the entire public-key infrastructure assumption of modern cryptography; if it is broken at scale, the resulting issues are far more severe than Sybil inflation of a trust score.

Where all parties to an IPR are under the adversary's control, the signature requirement adds no cryptographic defense beyond the overhead of producing correctly-structured IPRs. A single adversary controlling multiple agent DIDs can produce self-endorsements and, equally, can produce self-IPRs among its own DIDs — each party signs for the other at no additional cost beyond the signature operation. In this case the Sybil detection falls to the Jaccard cluster detector (Section 4) and the cross-vertical gate (Section 3.2) rather than to the signature requirement itself. Mechanism 1 therefore protects against a specific attack class — inflation of the interaction bonus through fabricated bilateral evidence with non-adversary agents — and does not claim to resist closed-cluster Sybil fabrication on its own.

Dual-signature IPRs serve a second function beyond raw Sybil resistance: they produce an empirically detectable divergence between endorsement-inflated agents and genuinely active agents. The Trust Score formula weights the interaction bonus separately from endorsement-derived scores [5, §4]. An agent with many endorsements but few bilateral IPRs with independent counterparties is structurally visible as an anomaly — the endorsements lack corresponding interaction evidence. This signal is available to any verifier independently of any privileged information held by the MolTrust registry. As of April 2026, the eleven IPR-related endpoints are operational, with public retrieval available for audit purposes.

3.2 Mechanism 2 — Cross-Vertical Endorsement Diversity Gate

Status: *live*. The second mechanism operates on the structure of the endorsement graph rather than on individual proof forgery. It is implemented as a two-stage construct — an incentive layer and a gate layer — that together enforce a minimum diversity condition on every non-seed agent's endorser set.

The incentive layer rewards breadth. An agent that accumulates endorsements across multiple verticals receives a bonus capped at 30 points: `min(unique_verticals × 10, 30)`. The bonus is weighted at γ = 0.1 in the Trust Score formula, with the weight chosen deliberately low because the primary role of the bonus is reward shaping, not defense.

The gate layer is the hard defense. A non-seed agent with endorsements from fewer than three distinct verticals incurs a flat Sybil penalty of 10.0. The Trust Score formula applies a $\times$20 multiplier to the Sybil penalty term, producing a raw subtraction of 200 points; clamped to the [0, 100] range, this is effective score nullification. The result: an agent without cross-vertical diversity cannot carry a non-zero Trust Score regardless of how many intra-vertical endorsements it accumulates.

The coordination cost argument behind the three-vertical threshold is direct. Two colluding agents within a single vertical can fabricate mutual endorsements at negligible cost. Requiring endorsements from three independent verticals forces the Sybil operator to maintain distinct credibility signals across unrelated domains — each vertical has its own credential schema, its own verification workflow, and its own operational context. The cost of maintaining credible presence across three verticals scales non-linearly with the number of Sybil identities controlled, because each vertical requires independent credential issuance activity.

"Distinct verticals" refers to the canonical vertical categories tracked by the registry: *core*, *skill*, *shopping*, *travel*, *prediction*, *salesguard*, *sports* — plus eight credential-type verticals defined in the Technical Specification (VerifiedSkillCredential, BuyerAgentCredential, AuthorizedAgentCredential, TravelAgentCredential, PredictionTrackCredential, ProductProvenanceCredential, AuthorizedResellerCredential, SkillEndorsementCredential). Unique vertical counting includes both endorsement-declared verticals and credential-type verticals, which broadens the set of legitimate paths to satisfying the three-vertical minimum while preserving the coordination-cost structure.

The gate necessitates a bootstrap path: a new registry cannot demand three-vertical coverage of agents registered on day one. This is handled via a seed-floor-guard, described in [5, §4.5], which protects a small number of named seed agents from falling below their assigned base scores. Seed base scores are tiered (85.0 for TrustScout, 80.0 for the MolTrust Ambassador, 75.0 for VCOne, 70.0 for standard seeded agents, 60.0 for AgentNexus) and the full tier table is discoverable through the registry's `/swarm/stats` endpoint. The bootstrap hierarchy is explicit rather than hidden. The seed mechanism is a named structural dependency of the current deployment; its security properties and associated limitations are documented as an honest gap in Section 6.7 rather than hidden inside this subsection.

3.3 Mechanism 3 — Principal-DID-Linked Violation Persistence

Status: *live*. The third mechanism addresses the identity-rotation attack, which is the most common Sybil strategy observed in permissionless systems: when one identity accumulates negative signals, abandon it and register a new one. In trust systems that associate behavioral history with a rotating agent identifier, this attack costs nothing more than a key generation.

MolTrust associates violation records — including MoltGuard security findings, Sybil detection flags, and repetitive endorsement pattern alerts — with the principal DID rather than the agent DID. The Five-Party Trust Chain (Developer $\rightarrow$ Owner $\rightarrow$ Agent $\rightarrow$ Instructor $\rightarrow$ Counterparty) formalizes this association [2, §3.2]. The Developer and Owner are principal identities; the Agent is a replaceable operational instance under the principal's control. When an agent is revoked and a new agent is registered by the same principal, the principal's violation history persists and attaches to the new agent by default.

Creating a genuinely new principal identity — one with no prior violation history — requires either (a) a new Trust Tier 0 KYC-backed credential, which re-exposes the adversary to whatever KYC friction the registry permits, or (b) registration as an unverified principal, which starts with a zero Trust Score and is subject to the three-vertical gate of Mechanism 2. Either path returns the adversary to the coordination-cost barrier the gate imposes. The combination of Mechanism 2 and Mechanism 3 is intentional: Mechanism 3 closes the "just rotate" escape path, and Mechanism 2 ensures the escape path leads back into a defended bottleneck.

A structural note on accountability: the principal-DID linkage is not used to punish legitimate Developer or Owner operations for agent errors. The Technical Specification [5, §8] distinguishes between violations attributable to the principal (policy violations, Sybil indicators) and violations attributable to the specific agent instance (runtime anomalies, kernel-layer enforcement detections via Falco). The former persist across agent re-registration; the latter do not automatically transfer. This distinction preserves the legitimate use case of retiring an agent that has drifted and deploying a successor, without inheriting operational noise that was specific to the retired instance.

3.4 Why the Three Mechanisms Are Designed to Compound

The defense-in-depth claim of this note is that a successful Sybil campaign must defeat all three mechanisms simultaneously. A rigorous formulation of "defeating" each mechanism would require adversarial cost-modeling at a level not supported by the current operational scale. What this note can state is the structural cost an adversary faces, without asserting specific numerical bounds on the total.

A Sybil campaign that inflates a target agent's Trust Score against the three mechanisms plus Jaccard cluster detection must structure itself as follows:

- To generate bilateral Interaction Proof Records that contribute to the interaction bonus with any defensive value, the adversary either (a) participates in IPRs with genuine external counterparties — in which case the adversary does not control the external signatures and Mechanism 1 imposes a genuine key-compromise cost; or (b) produces IPRs entirely between adversary-controlled identities — in which case Mechanism 1 adds no defense beyond signature overhead and detection falls to Sections 3.2 and 4. Path (b) defeats Mechanism 1 trivially but leaves the adversary fully exposed to the subsequent mechanisms.
- To pass the three-vertical gate, each Sybil identity requires credible credential activity in three distinct verticals. The cost of establishing such activity is a combination of credential-issuer friction, domain-specific credibility signals, and operational overhead per vertical-per-identity. This cost is not captured by a single monetary figure, and it scales at least linearly in the number of identities an adversary wishes to establish.
- To stay below the Jaccard threshold at cluster size *n*, each endorser set must differ from every other by strictly more than one-fifth of the combined endorser pool. The cost of engineering sufficiently dissimilar endorser sets scales at least linearly with cluster size, and superlinearly when the endorsers must themselves pass cross-vertical and principal-persistence checks.
- To defeat Mechanism 3 (principal-DID-linked violation persistence), the adversary must either avoid generating any detection signals against any identity in the campaign — which implies defeating Mechanisms 1 and 2 cleanly on the first attempt — or establish a fresh principal identity for each Sybil cluster, which re-exposes the adversary to KYC friction or to the zero-score starting position of an unverified principal.

No single figure summarizes the total. A quantitative cost model requires adversarial red-team engagement, which is on the post-funding roadmap (Section 6.2). This note makes the weaker but honest claim that the mechanisms are structurally independent — defeating one does not assist in defeating the others — and that the cost of defeating all three simultaneously is materially higher than the cost of defeating any one in isolation. The Economic Audit Trail layer (Section 5) does not appear in this compounding argument because it is not claimed as a defensive mechanism; it is an auditability property.

---

4. Jaccard Cluster Detection Mechanics

The three mechanisms of Section 3 are complemented by an automated cluster-detection heuristic that identifies coordinated endorsement patterns by comparing the endorser sets of every agent pair. The heuristic is operational in the reference deployment and contributes directly to the Sybil penalty term of the Trust Score formula.

4.1 Definition and Formula

For two agents *i* and *j* with endorser sets *E_i* and *E_j*, the Jaccard similarity coefficient is defined as:

```
J(E_i, E_j) = |E_i ∩ E_j| / |E_i ∪ E_j|
```

When *J(E_i, E_j)* exceeds the production threshold of 0.8, a Sybil penalty is computed:

`penalty = J(E_i, E_j) × num_endorsers × 0.5`

The penalty feeds into the final Trust Score formula as `sybil_penalty × 20`, subtracted before the [0, 100] clamp. The threshold of 0.8 is a configuration parameter exposed for operational adjustment; it is not a cryptographic constant.

4.2 Worked Example

Consider four scenarios that illustrate the detector's operating range.

Case A — no overlap suspicion. Agent X has endorsers {A, B, C, D}. Agent Y has endorsers {A, B, C, E}. The Jaccard index is |{A, B, C}| / |{A, B, C, D, E}| = 3/5 = 0.6. This is below the 0.8 threshold and no penalty is applied. The overlap is consistent with agents operating in the same ecosystem, where some shared endorsers are expected.

Case B — boundary case. Agent X has endorsers {A, B, C, D}. Agent Z has endorsers {A, B, C, D, E}. J = 4/5 = 0.8.[1] The overlap sits exactly at the threshold; no penalty is applied, and no alert is produced. A conservative operator may wish to configure a warning at this boundary.

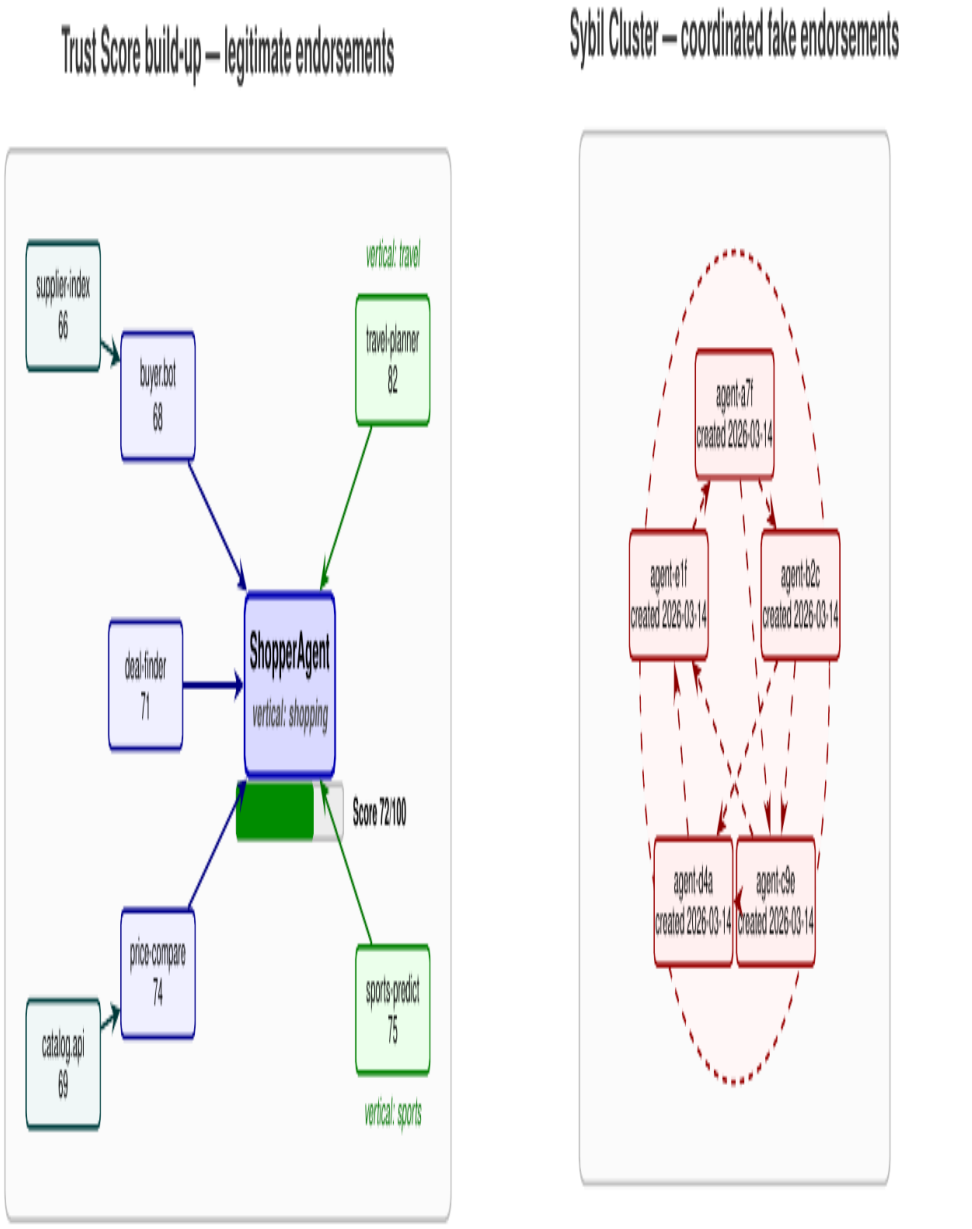
Case C — identical sets. Agent X has endorsers {A, B, C, D}. Agent W has endorsers {A, B, C, D}. J = 4/4 = 1.0. Penalty = 1.0 × 4 × 0.5 = 2.0. After the ×20 multiplier, a 40-point reduction is applied to the agent's final score. For an agent with a raw score of 65, the final score becomes 25 — below the threshold typically required for commercial interactions in the reference deployment.

Case D — high similarity with large set. Agent X has endorsers {A, B, C, D, E, F, G, H}. Agent V has endorsers {A, B, C, D, E, F, G, I}. J = 7/9 ≈ 0.78, below the threshold. But if V's endorsers were {A, B, C, D, E, F, G, H, I}, J = 8/9 ≈ 0.89. Penalty = 0.89 × 9 × 0.5 = 4.0. After ×20: 80-point reduction, effective score nullification regardless of raw score.

The penalty is designed to scale with both the strength of the overlap signal and the size of the endorser set that produced it, on the rationale that a high-Jaccard overlap on a large set is a stronger Sybil signal than a high-Jaccard overlap on a small set.

[1] The reference implementation treats the condition as strict ($J > 0.8$). An implementation variant using $J \geq 0.8$ would apply a penalty at the boundary and is compatible with the Technical Specification.

4.3 Figure — Legitimate Build-Up vs. Detected Cluster



- Direct endorser (counterparty in past interaction) – weight $\alpha = 0.6$
- Propagated endorser (endorses a direct endorser) – weight $\beta = 0.3$
- Cross-vertical endorser (endorses from another domain) – weight $\gamma = 0.1$

Example: Score $= 0.6 \cdot \frac{68+71+74}{3} + 0.3 \cdot \frac{68+69}{2} + 0.1 \cdot \frac{82+75}{2} = 71.8 \approx 72$

Detection signals:

- All 5 agents registered on the same day (2026-03-14)
- Endorsement sets nearly identical: Jaccard$(E_i, E_j) = 0.92$
- All in a single vertical – no cross-vertical activity

Jaccard $> 0.8 \Rightarrow$ cluster detected $\Rightarrow$ score cap applied to all members

Figure 1. *Left:* a legitimate agent (ShopperAgent, vertical shopping) accumulates a Trust Score of 72/100 through weighted endorsements from direct counterparties (α = 0.6), propagated endorsers (β = 0.3), and cross-vertical endorsers (γ = 0.1). *Right:* a coordinated ring of five agents registered on the same day with near-identical endorsement sets (Jaccard = 0.92) and no cross-vertical activity is automatically detected and score-capped. The Jaccard threshold (> 0.8) and the three-vertical minimum are the two structural constraints that make coordinated fake-endorsement campaigns economically ineffective. Figure reproduced from the MolTrust arXiv preprint v1.0 [2, Fig. 4].

4.4 Threshold Calibration

The 0.8 threshold is deliberately permissive. Legitimate agents operating in the same ecosystem frequently share some endorsers: two commerce agents active in the same marketplace will both be endorsed by a few of the same counterparties, and two skill-verification agents in the same domain may have partially overlapping endorser pools. Setting the threshold too low would generate false positives against normal operation and erode trust in the detection signal itself. Setting the threshold too high — near 1.0 — would permit moderately coordinated fake-endorsement campaigns to escape detection.

The 0.8 value is empirically calibrated against the current population, not formally proven optimal. At N = 54 agents the distribution of pairwise Jaccard indices across legitimate agent pairs is bounded well below 0.8 with room to spare, and the threshold does not fire on legitimate activity in the reference deployment. This is a weak claim; the population is too small for meaningful distributional analysis. Formal sensitivity analysis across the [0.6, 0.95] range is planned once the agent population exceeds 200 — enough to produce a non-trivial empirical distribution of legitimate pairwise similarities against which to calibrate.

4.5 Scope Note on Detection Limitations

Jaccard detection is effective against naive Sybil patterns — clusters with near-identical endorser sets. It is less effective against sophisticated patterns where an adversary distributes endorsements across a non-overlapping partition of carrier identities. Such partitioned attacks defeat pairwise Jaccard detection by construction. Addressing partitioned Sybil patterns requires higher-order graph analysis (community detection, endorsement-flow analysis, temporal clustering) that is on the roadmap but not yet implemented. As a consequence, the Jaccard mechanism is correctly characterized as a coarse-grained detector against low-sophistication coordinated-endorsement attacks, not as a general-purpose Sybil-graph analyzer. The scale at which partitioned attacks become economically attractive to an adversary is itself an empirical question that this note does not answer; Section 6.4 treats that gap explicitly.

---

5. Economic Audit Trail

Status: *live*. Earlier revisions of this note framed the x402 payment integration as a defensive "economic cost layer." That framing was misleading and has been replaced. The monetary cost of a verification campaign at April 2026 pricing ($0.05 per agent score query, $0.10 per Sybil-scan endpoint call) is low in absolute terms — a 1,000-query bootstrap attack incurs on the order of $50–150 including Base Layer 2 gas. This cost is not a meaningful defense against adversaries with any non-trivial budget. The layer described in this section is therefore not a Sybil defense; it is an auditability property.

What the layer does provide is a cryptographically committed payment record attached to every verification event. MolTrust operates a per-verify pricing model through the x402 protocol integration [6], with all payments recorded on-chain as USDC transfers on Base Layer 2. Every verification is therefore linked to a verifiable payment event, and every paid event contributes to the audit trail independently of the content of the verification.

The auditability property has three concrete consequences. First, attempts to claim a trust score without performing the paid query are detectable as missing payment records. This is a weaker property than the signature-based protections of Mechanism 1 but is a useful auxiliary signal in disputes. Second,

the payment record provides an independent evidentiary substrate for regulatory audit — the EU AI Act Article 12 (record-keeping) and the NIST AI RMF GOVERN-1.3 subcategory both require evidence trails for AI system decisions, and a payment-anchored verification record satisfies part of that requirement at the infrastructure layer. Third, the payment record creates an economic flow that aligns the interests of the registry and the verifier: every verification generates revenue, and every revenue event is a cryptographic commitment that the verification occurred, which changes the incentive structure around verification-event spoofing.

To reiterate, this layer does not appear in the compounding argument of Section 3.4. Its role in this note is to document an auditability mechanism that is operational and to be explicit that it is not a defensive mechanism. Any reader interpreting the x402 integration as a meaningful adversarial cost barrier is invited to re-read this section.

---

6. Honest Gaps and Open Problems

An honest account of Sybil resistance at this stage of deployment must state its limitations clearly. The following subsections list the gaps that bound the claims of this note.

6.1 Scale Discipline

The MolTrust reference registry operates with 54 registered agents and 57 endorsements as of April 2026. The production operational window is two months. These numbers are stated for the avoidance of doubt: they describe a functioning production system but do not constitute validation at the scale that the opening of the MolTrust arXiv preprint [2, §1] describes for the agent economy (Agent.market's 69,000 autonomous bots being the nearest comparable reference). The three-mechanism architecture is designed with that scale in mind — Jaccard detection scales quadratically with agent count in the straightforward pairwise implementation, the cross-vertical gate is O(1) per agent, and dual-signature IPR verification is O(1) per record — but the heuristic parameters (Jaccard threshold, penalty multipliers, cross-vertical minimum) have been tuned against the current scale. Whether the same parameters remain appropriate at $10\times$, $100\times$, or $1000\times$ the current population is an empirical question that has not been answered.

6.2 No Adversarial Validation

No deliberately adversarial Sybil campaign has been conducted against the live registry. No third-party red team has been commissioned. The three-mechanism architecture has not been adversarially tested beyond the operational monitoring that the registry performs as part of normal operation. The implication is direct: claims about Sybil resistance in this note are architecture claims, not empirical-defense claims. The architecture is designed to resist known coordination-cost arguments; whether a sophisticated adversary can circumvent the architecture in practice is an open question.

Red-team Sybil campaigns are on the post-funding roadmap and are identified as a prerequisite for any claim of adversarial robustness. Until those campaigns have been conducted and their findings published, this note's claims remain bounded by architectural reasoning.

6.3 Heuristic Nature of the Core Parameters

The Jaccard threshold (0.8), the three-vertical minimum, and the $\times 20$ penalty multiplier are empirical calibration parameters, not values derived from a formal security model. Each can be adjusted, and each is exposed through configuration. The parameters produce operationally sensible behavior at the current scale but have not been proven optimal in any formal sense. Alternative configurations — tighter thresholds with higher false-positive tolerance, looser thresholds with higher false-negative tolerance, multipliers scaled by agent population — have not been systematically compared.

A parallel concern applies to the Trust Score weights themselves: α = 0.6 for direct-endorsement scores, β = 0.3 for propagated scores, γ = 0.1 for the cross-vertical bonus. The 2:1 ratio between direct and propagated reflects a design preference for direct evidence over transitively inherited trust, and the small γ weight reflects the intent that the primary cross-vertical defense is the gate rather than the bonus. These weights are plausible but not optimal, and a formal sensitivity analysis across alternative weightings is planned at N $\geq$ 200 agents with sufficient endorsement density for statistical comparison.

6.4 Sophisticated Attack Patterns

Several attack patterns are outside the reach of the current detection mechanisms. The taxonomies in Ferrag et al. [8] and in Schroeder de Witt [7] motivate the categorization below.

Partitioned Sybil attacks, as noted in Section 4.5, distribute endorsements across non-overlapping carrier identities to defeat pairwise Jaccard detection. Higher-order graph analysis is required; it is on the roadmap but not yet operational. The scale at which a partitioned attack becomes economically attractive to an adversary depends on the adversary's credential-issuance capacity and on the vertical diversity the adversary can credibly fabricate. The reference deployment has no empirical data on that threshold and does not claim one. A partitioned attack remains an unvalidated risk, and the cross-vertical gate is the primary structural defense that survives in a partitioned regime — because each Sybil identity must still individually satisfy the three-vertical minimum regardless of how cleverly the endorser sets are partitioned.

Slow-drip campaigns, which spread fake endorsement accumulation over long timeframes to avoid co-registration date clustering, can defeat date-based heuristics. Current detection does not include a temporal clustering signal.

Principal-collusion attacks, in which two legitimate principal DIDs coordinate to cross-endorse each other's agents across multiple verticals, are difficult to distinguish from legitimate business partnership patterns. The only available signal is behavioral anomaly detection after the fact. This is a residual risk and is acknowledged as such.

6.5 No Third-Party Security Audit

The MolTrust Protocol has not undergone independent third-party penetration testing or formal security audit as of April 2026. An internal security review in March 2026 identified and remediated findings across CORS configuration, hardcoded credential exposure, XSS vectors, CSRF handling, and race conditions in credential issuance. The SKILL AUDIT v1.0 infrastructure [2, §4.2] provides automated conformance checking with CWE-mapped vulnerability coverage, but this is not a substitute for adversarial human assessment. A third-party audit is on the post-funding roadmap and is identified as a prerequisite for enterprise adoption.

6.6 Two-Month Production Horizon

Trust infrastructure makes temporal claims — decay over time, reputation persistence across re-registration, behavioral pattern emergence — that cannot be validated in a deployment that has been operational for two months. Trust score decay behavior over periods exceeding the 90-day half-life has not been observed empirically. Reputation inheritance patterns across agent re-registrations have been implemented but not exercised at meaningful scale. Long-term adversarial adaptation cycles, where attackers observe the detection heuristics and adjust their strategies in response, have not occurred because no adversarial campaigns have targeted the registry. These are inherent limitations of a two-month observation window and cannot be closed except by continued operation.

6.7 Seed-Floor-Guard Dependency

The MolTrust bootstrap mechanism described in Section 3.2 relies on a small set of seed agents — currently five — whose Trust Scores are floor-protected from falling below predefined base values (85.0 for TrustScout, 80.0 for the MolTrust Ambassador, 75.0 for VCOne, 70.0 for standard seeded agents,

60.0 for AgentNexus). The mechanism is necessary because the cross-vertical diversity gate cannot be applied on day one of a registry: a new system has no agents meeting a three-vertical endorsement requirement by definition. The seed-floor-guard provides the initial high-trust identities from which the endorsement graph can grow.

The dependency is a real structural weakness and is stated here as such rather than hidden inside the mechanism description. Compromise of a seed agent's Ed25519 private key produces a high-trust identity that inherits the seed's base score until revocation, which in turn requires governance action by registry operators. As of this note's publication, the reference deployment has not publicly documented its seed key storage and access controls, and those controls have not undergone independent audit. Until such documentation and audit exist, the claim of Sybil resistance in this note must be understood as conditional on the integrity of the seed key material.

Two mitigations are planned. The first is multi-signature seed control: the roadmap item is to require M-of-N signatures for any endorsement action by a seed agent, reducing single-key compromise to a partial rather than a total failure. The second is seed retirement: once the agent population exceeds a threshold at which natural cross-vertical diversity is achievable across the registry, the seed-floor-guard is retired and seeds are demoted to ordinary agents subject to Mechanism 2. Both items are specified but not yet implemented. In the current deployment, the seed-floor-guard is a named and audit-visible dependency, but it is a single-key-controlled dependency, and the Sybil resistance claims of this note are conditional on that dependency holding.

---

7. References

[1] Douceur, J. R. *The Sybil Attack.* In *Peer-to-Peer Systems* (IPTPS 2002), Springer LNCS 2429, pp. 251–260.

[2] Kroehl, L. K. *From Specification to Deployment: Empirical Evidence from a W3C VC + DID Trust Infrastructure for Autonomous Agents.* MolTrust arXiv preprint v1.0, 22 April 2026. SHA-256 prefix `c9c34985`, suffix `1c11d946`; anchored on Base Layer 2 Mainnet, Block 45,037,732. Full hash: https://moltrust.ch/publications/integrity.html

[3] W3C. *Decentralized Identifiers (DIDs) v1.0.* W3C Recommendation. https://www.w3.org/TR/did-core/

[4] W3C. *Verifiable Credentials Data Model v2.0.* W3C Recommendation. https://www.w3.org/TR/vc-data-model-2.0/

[5] MolTrust / CryptoKRI GmbH. *MolTrust Protocol Technical Specification v0.8.* April 2026. Anchored on Base Layer 2, Block 44,745,864. https://moltrust.ch/techspec

[6] Coinbase, Cloudflare. *x402 Payment Protocol Specification.* https://x402.org

[7] Schroeder de Witt, C. *Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents.* arXiv:2505.02077, 2025. University of Oxford, Department of Engineering Science.

[8] Ferrag, M. A., Tihanyi, N., Hamouda, D., Maglaras, L., Lakas, A., Debbah, M. *From Prompt Injections to Protocol Exploits: Threats in LLM-Powered AI Agents Workflows.* ICT Express, in press 2026. DOI: 10.1016/j.icte.2025.12.001. arXiv:2506.23260.

[9] Mao, Y. et al. *Systematization of Knowledge on AI Agent Security.* arXiv:2604.15367, 2026.

[10] Hu, B., Rong, H. *Inter-Agent Trust Models: A Comparative Study of Brief, Claim, Proof, Stake, Reputation and Constraint in Agentic Web Protocol Design.* arXiv:2511.03434, 2025. AAAI 2026 TrustAgent Workshop.

[11] Infocomm Media Development Authority, Singapore. *Model AI Governance Framework for Agentic AI.* IMDA, January 2026. Published at World Economic Forum, Davos.

[12] National Institute of Standards and Technology, Center for AI Standards and Innovation. *AI Agent Standards Initiative.* February 2026. https://www.nist.gov/caisi/ai-agent-standards-initiative

[13] European Union. *Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act).* 2024.

[14] MolTrust / CryptoKRI GmbH. *MolTrust — EU AI Act Mapping.* Companion document, April 2026.

[15] MolTrust / CryptoKRI GmbH. *MolTrust — NIST AI RMF Mapping.* Companion document, April 2026.

This document is intended as a standalone methodology note. It will be anchored on Base Layer 2 as part of the batch anchor with the EU AI Act Mapping and the NIST AI RMF Mapping. Verification instructions will accompany the anchored version.