PETS 2026 · POSTER · EUDI WALLETS · RATS · AGENTIC AI

Privacy Boundaries for Hardware-Attested AI-Agent Action on EU Digital Identity Wallet Rails

When per-action accountability evidence risks becoming correlation infrastructure

Anton Sokolov · Tyche Institute · ORCID 0000-0003-2452-7096

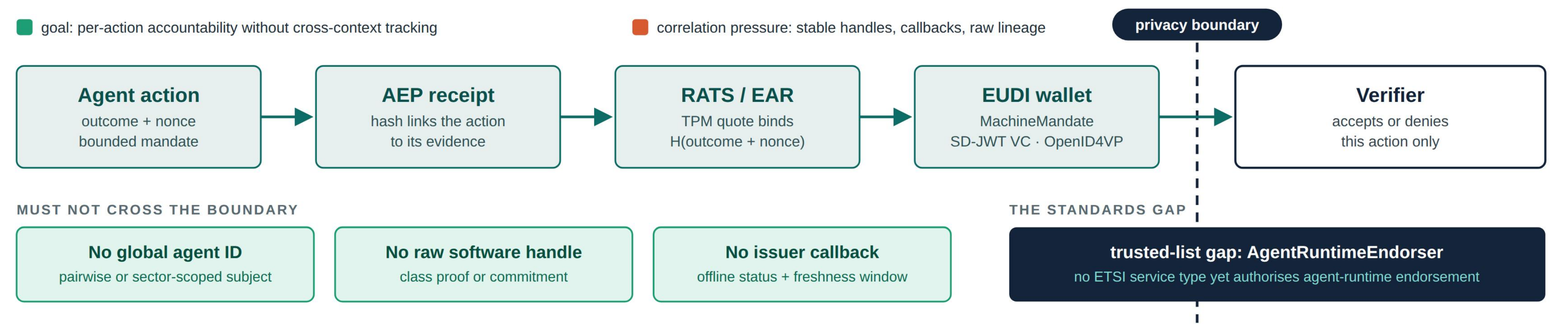
Tyche Institute
RESEARCH INFRASTRUCTURE

The Research Question

Can an open AI agent prove, live and per action, that it acted under an authorised mandate on an attested runtime — without handing every relying party a durable tracking handle?

Open agents need portable, third-party-verifiable evidence at action time. Closed-world certification and post-hoc black boxes do not transfer; producing live accountability creates the privacy failure mode this poster maps: evidence becomes correlation infrastructure.

Evidence Flow and the Privacy Boundary



The verifier learns that this action is fresh, scoped, attested, and endorsed — without receiving a durable agent or runtime handle. The dashed line is the design target; the dark chip is the one missing authorisation surface on today's trusted lists.

Closed World Does Not Transfer

Cars, aircraft, medical devices, and trading systems assign accountability through pre-certified products, statutory liability, and internal logs read after an incident.

Open AI agents are software-only, continuously updated, multi-provider, and cross-organisational. The verifier may have no prior relationship with the operator.

Key difference the verifier needs live evidence for *this* action, not an internal black-box record for later investigators.

The Privacy Problem

Naive accountability makes an agent easy to follow. The same proof that helps a verifier reject replay, scope abuse, and runtime substitution can also reveal stable identifiers, software lineage, mandate structure, and repeated action patterns.

PETS ANGLE accountability must be scoped, pairwise, and minimally disclosing.

Bench Evidence

7/7
Adversarial cases decided correctly: replay, EAR transplant, contraindicated runtime, wrong role, absent role.

6/6
Scope enforcement: a confused-deputy over-scope is denied even when crypto and attestation pass.

193 µs
Median local three-layer verification over 2,000 iterations.

1 gap
The EC reference verifier cannot yet natively resolve an agent-runtime service type.

The Weld

Agent action & outcome › **AEP receipt hash** › **TPM quote binds H(outcome + nonce)**

Veraison RATS appraisal › **EAR status + software id** › **MachineMandate SD-JWT VC**

OpenID4VP presentation › **Relying-party L2 check** › **AgentRuntimeEndorser role gate**

The relying party re-derives freshness and quote-binding from the raw quote bytes; the issuer is accepted only if listed under a proposed **AgentRuntimeEndorser** trusted-list role.

Privacy Boundary Map

Verifier must learn

- The mandate covers this action.
- The presentation is fresh and audience-bound.
- The runtime appraisal is affirming.
- The endorser is authorised for runtime claims.

Verifier should not learn

- A global agent identity across contexts.
- A reusable model / policy fingerprint unless necessary.
- Unrelated scope or the full mandate structure.
- Status-check metadata that reveals an activity history.

What Gets Denied Where

Layer	Accepts	Denies
L1 · SD-JWT VC / KB-JWT	Issuer signature, disclosure, nonce, audience	Forgery, disclosure mismatch, replay at presentation
L2 · RATS / AEP weld	Fresh affirming attestation bound to the action	Stale quote, EAR swap, contraindicated appraisal
L3 · Trusted-list role	Issuer listed as AgentRuntimeEndorser	Wrong service type or unlisted issuer
L4 · Mandate scope	Action inside the delegated scope	Confused-deputy over-scope

What We Can Show

- A RATS/EAR verdict carried as claims inside a MachineMandate SD-JWT VC.
- An OpenID4VP request binding the presentation to the verifier nonce and audience.
- A relying party re-deriving freshness and quote-binding from raw quote bytes.
- An issuer accepted only if listed under a proposed AgentRuntimeEndorser role.
- A confused-deputy over-scope denied even when every other layer passes.

Correlation Surfaces

Surface	Risk	Possible boundary
Agent subject	Cross-verifier tracking	Pairwise or sector-specific pseudonyms
Software identity	Model-lineage fingerprint	Commitment or class proof, not raw name
Action hash	Repeated-task linkage	Domain-separated, nonce-bound commitments
Status / revocation	Issuer observes presentations	Offline status lists, cached freshness
Endorser role	Trust list becomes activity map	Role discovery separate from telemetry

Questions for PETS

- What is the minimum disclosure set for per-action accountability?
- Can a runtime be proven as a *class* property rather than a stable identifier?
- How should status freshness work without leaking presentation events?
- Can trusted-list discovery stay separate from per-action telemetry?
- Where should selective disclosure stop and zero-knowledge proofs begin?

Design Hypothesis

The right primitive is not "agent identity" alone. It is a **bounded proof** that discloses only:

- this mandate permits this action;
- this runtime was appraised for this action;
- this endorser is authorised for runtime claims;
- nothing stable is disclosed beyond that need.

Privacy Controls to Test

- Pairwise agent identifiers per relying-party sector.
- Selective disclosure of mandate fields and scope.
- Commitments for action hashes and software identity.
- Offline revocation / status evidence where possible.
- Short freshness windows with no issuer callback.

What Would Falsify This

- If per-action proof necessarily exposes a stable runtime handle.
- If status freshness requires live issuer observation.
- If relying parties need the raw model identity for safety decisions.
- If trusted-list role resolution becomes telemetry.
- If scope checks cannot be made machine-readable enough to deny on.

Core idea: live per-action accountability for open AI agents must not become tracking infrastructure.

Standards rails: RATS/EAR, AEP, SD-JWT VC, OpenID4VP, ETSI TS 119 612, EUDI Wallet.

Artifact: reproducible research bench, archived at Zenodo doi:10.5281/zenodo.21113145.