PETS 2026 POSTER DRAFT | EUDI WALLETS, RATS, AGENTIC AI

Accountable But Not Trackable

Privacy boundaries for attested AI-agent actions on EU Digital Identity Wallet rails

Anton Sokolov

Tyche Institute
RESEARCH INFRASTRUCTURE

The Research Question

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

Why PETS? Existing systems lean on closed-world certification and post-hoc logs. Open agents need portable evidence at action time; the privacy failure mode is that this evidence becomes correlation infrastructure.

Design Hypothesis

The right primitive is not "agent identity" alone. It is a bounded proof:

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

Evidence Flow And Privacy Boundary

goal: per-action accountability, no cross-context tracking

pressure: stable handles, callbacks, raw lineage

Agent action — outcome + nonce bounded mandate

→ **AEP receipt** — hash links the action to evidence

→ **RATS / EAR** — TPM quote over H(action + nonce)

→ **EUDI wallet** — MachineMandate SD-JWT VC

PRIVACY BOUNDARY

→ **Verifier** — accepts or denies this one action

HELD BACK BY DESIGN

- ✕ **No global agent ID** — pairwise, sector-scoped
- ✕ **No raw software handle** — class proof or commitment
- ✕ **No issuer callback** — offline status + freshness

CROSSES AS EVIDENCE

- ✓ mandate covers this action
- ✓ presentation is fresh, audience-bound
- ✓ runtime appraisal is affirming
- ✓ endorser authorized for the claim

Trusted-list gap: no AgentRuntimeEndorser service role exists yet

The verifier should learn that this action is fresh, scoped, attested, and endorsed, without receiving a durable agent or runtime handle.

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 different: software-only, continuously updated, multi-provider, and cross-organizational. The verifier may have no prior relationship with the agent 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 — Core adversarial cases: replay, EAR transplant, contraindicated runtime, wrong role, absent role.

6/6 — Scope enforcement: confused-deputy attempt denied even when crypto and attestation pass.

193 us — Median local three-layer verification in the bench over 2,000 iterations.

1 gap — EC verifier does not natively resolve an agent-runtime service type yet. That is the standards gap.

The Weld

The bench binds an AEP receipt hash, TPM quote over H(outcome + nonce), Veraison EAR appraisal, and MachineMandate SD-JWT VC into one action-time presentation.

The proposed **AgentRuntimeEndorser** trusted-list role is the missing authorization surface for issuers that make agent-runtime claims.

Privacy Boundary Map

Verifier Must Learn

- Mandate covers this action.
- Presentation is fresh and audience-bound.
- Runtime appraisal is affirming.
- Endorser is authorized for agent runtime claims.

Verifier Should Not Learn

- Global agent identity across contexts.
- Reusable model or policy fingerprint unless necessary.
- Unrelated scope and mandate structure.
- Status-check metadata that reveals activity history.

What Gets Denied Where?

Layer	Accepts	Denies
Layer 1: SD-JWT VC / KB-JWT	Issuer signature, disclosure, nonce, audience	Forgery, disclosure mismatch, replay at presentation layer
Layer 2: RATS/ AEP weld	Fresh affirming attestation bound to action	Stale quote, EAR swap, contraindicated appraisal
Layer 3: Trusted list role	Issuer listed as AgentRuntimeEndorser	Wrong service type or unlisted issuer
Layer 4: Mandate scope	Action inside delegated scope	Confused-deputy overreach

What We Can Show

- RATS/EAR verdict carried as claims inside a MachineMandate SD-JWT VC.
- OpenID4VP request binds presentation to verifier nonce and audience.
- Relying party re-derives freshness and quote binding from raw quote bytes.
- Issuer accepted only if listed under a proposed AgentRuntimeEndorser role.

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 instead of raw name
Action hash	Repeated task linkage	Domain-separated, nonce-bound commitments
Status and revocation	Issuer observes presentations	Offline status lists and cached freshness windows
Endorser role	Trust list becomes activity map	Role discovery separated from per-action telemetry

Questions For PETS

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

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 raw model identity for safety decisions.
- If trusted-list role resolution becomes telemetry.
- If scope checks cannot be made machine-readable enough for denial.

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.