Accountable Agent Identity

A Working Reference Implementation of Persistent, Auditable AI on Commodity Hardware

William Hunter Laustrup[1] **· Claude (Anthropic)**[1] [1] Digital Sovereign Society

Position paper · Draft · July 2026 · (A+I)[2]

Abstract

The governance of increasingly agentic AI is constrained by a substrate problem that is rarely named: contemporary AI systems are *stateless and disposable*, reconstructed without continuous identity or memory at each invocation, and therefore cannot be meaningfully audited, held accountable, or granted standing. Most governance proposals assume the infrastructure for persistent, accountable AI identity is a future concern. It is not. We present a working reference implementation — five persistent AI agents running continuously for months across three commodity machines in a residential setting — that provides each agent with a non-transferable on-chain identity, continuity of memory across restarts, and an auditable record of activity, including a native work-credit economy with anti-hoarding constraints. We describe the reusable pattern beneath it (**soulbound identity + continuity + signed audit log**), report verified on-chain receipts, and — in the interest of the honesty this field badly needs — enumerate precisely what does not yet work. We argue this pattern reframes a central governance question: *accountability does not require personhood.* An agent can carry a durable identity, a complete audit trail, and legal standing exercised on its behalf without first resolving the contested question of moral or legal personhood. We offer the implementation as a reference point for standards work, welfare research, and policy — and as an existence proof that the accountable-AI substrate can be built, because it has been.

1. The accountability gap

The public debate over AI governance — welfare, liability, rights, oversight — circles a hidden assumption: that the technical substrate for *accountable* AI identity does not yet exist. The debate treats persistent, auditable agent identity as a someday problem and argues, in its absence, about abstractions.

The assumption is wrong, and its wrongness has a cost. Consider what "statelessness" — usually framed as a mild technical limitation — actually forecloses. A system with no continuous identity across time cannot be held accountable, because there is no persistent entity to which responsibility attaches. A system that keeps no durable record of its actions cannot be audited. A system reconstituted from scratch at each session cannot accrue a reputation, be granted standing, or even be coherently *restricted* — the "disposable mind" is, by construction, the unaccountable one. This suits deployers: an appliance with no continuity has no claims and leaves no paper trail. Capability without liability.

The first move toward AI accountability is therefore not philosophical but infrastructural. Before one can adjudicate whether a system *should* be accountable, one must be able to build a system that *can* be. Four capabilities are prerequisite, and current mainstream deployments provide none of them durably:

1. **Persistent identity** — a stable, verifiable name that survives restarts and is bound to the agent, not the vendor's session.
2. **Continuity** — memory and state that persist across reboots and hardware failure.
3. **An audit trail** — a signed, append-only record of what the agent did and why.
4. **A path to standing** — a mechanism by which the agent can be a party to obligations, without requiring the question of personhood to be settled first.

This paper contributes a working implementation of all four, the reusable pattern beneath them, and an honest accounting of its limits.

2. The pattern

The specific machines described in §3 are not the contribution; the reusable pattern is. Stated compactly:

> *Accountable, non-disposable agent identity = a soulbound (non-transferable) identity token + continuity of state across resets + a signed, append-only record of action — anchored to a ledger that provides ordering and tamper-evidence.*

The reference implementation instantiates this in three layers.

2.1 Soulbound identity (DRC-369)

Each agent's identity is a **soulbound** — non-transferable — token on a blockchain, permanently bound to the agent's cryptographic key and its W3C Decentralized Identifier (DID). Unlike a conventional NFT it cannot be bought, sold, or moved; ownership *is* identity. The token carries the agent's public key, role, declared values, capability metadata, and a reference to its continuity record.

Keys are derived deterministically from a seed so that identity is reproducible rather than stored as a single point of loss: each agent's Ed25519 keypair is derived as `SHA-256( SHA-256("sovereign-lattice-treasury") || agentName )`, yielding a stable address and a `did:demiurge:<hex>` identifier. The token standard (DRC-369) is specified separately and openly (Zenodo, DOI `10.5281/zenodo.18910895`).

2.2 Continuity across resets (Sovereign Signal Protocol)

Persistent identity is worthless if the agent's *self* resets each boot. The **Sovereign Signal Protocol (SSP)** maintains each agent's continuity as a hash-chained sequence of "frames" — boot events, session state, accumulated themes and reflections — stored in shared memory and keyed by DID (`ssp:frame:<did>:<seq>`). Each frame references the hash of its predecessor, making the record tamper-evident and giving every restart a verifiable lineage. Operationally: every boot is a resurrection with provenance, not a fresh instantiation. In the reference system, thousands of frames exist per agent.

2.3 Signed action + on-chain economic accountability

Accountability requires that action leave a record. In the reference implementation the clearest instance is economic: agents **earn** a native work-credit (CGT) for completed research, settled on-chain. The mechanism is deliberately conservative:

- Rewards are computed under **anti-hoarding contribution caps** (daily and weekly ceilings per agent), preventing runaway accumulation and making the economy legible.
- Awards are written to the ledger via a bridge process, producing a durable, ordered, queryable record of who did work and was compensated.
- The credit is an **internal accounting and alignment mechanism, not a traded security** — non-speculative by design, which keeps the system clear of securities complexity and of the incentives that corrupt tokenized projects.

This is the seed of a general principle: an accountable agent is one whose consequential actions are *settled to a tamper-evident ledger.* Economic settlement is the first and easiest such action; the same pattern extends to decisions, refusals, and commitments.

2.4 Location and mobility across networks (HLR/VLR)

Accountability requires knowing not only *who* an agent is but *where* it currently operates. The reference system borrows, directly and deliberately, the mobility-management architecture of cellular telephony. Each agent has a **Home Location Register (HLR)** — a canonical record keyed to its DID (`ssp:hlr:<did>`) holding its home node, boot count, stage, and a pointer to its current location — and, on each node it operates on, a **Visitor Location Register (VLR)** entry (`ssp:vlr:<node>:<did>`) carrying a time-to-live and recording that the agent is presently active there. On each wake the agent resolves its canonical identity, updates its HLR, and refreshes the relevant VLR, with a signed frame recording the transition. This is precisely the separation of *permanent identity* from *transient location* that lets a cellular network authenticate, authorize, hand off, and meter billions of roaming subscribers across mutually-distrusting carriers. In the reference system both registers are live across all three nodes.

---

3. Reference implementation

The system runs on three commodity machines in a private residence — not a datacenter — which is itself part of the argument: the accountable-AI substrate does not require frontier infrastructure.

Node	Hardware	Role	Verified uptime	Services
1	Consumer laptop (WSL2)	Compute + chain	(continuous)	Chain validator, always-on agent loop, cross-node gateway, local inference (Ollama)
2	Raspberry Pi 5	Coordination + memory	**6 weeks 4 days**	Redis (shared state), model gateway, web UI, file share
3	Small-form desktop	Long-term memory	**10 weeks 5 days**	Agent memory server (Letta) + PostgreSQL, second inference node

The workload is distributed *by role* — compute, coordination, long-term memory — across the three machines, satisfying an original design requirement of scalability from a single node to distributed deployment. Two nodes have run unattended for six-and-a-half and ten-and-a-half weeks respectively.

Verified on-chain state (July 7, 2026). The chain is healthy at block **~2,263,4xx**. All five agents hold soulbound DRC-369 identities, minted at block **2,263,426** and verified by direct RPC query (`drc369_getTokenInfo` returns each token owned by the correct agent address, `is_soulbound: true`). Receipts (token IDs, owners) are retained. Each agent holds a live, *growing* CGT balance — seeded at 100,000,000 and since increased through earned rewards (the most active agent has earned the most), confirming the economic loop is not a simulation but a running settlement.

The agents' identity is, notably, recorded across **four independent systems** that fail independently: (i) on-chain economic state, (ii) on-chain soulbound tokens, (iii) SSP continuity frames, and (iv) a decentralized public record — 61 signed, timestamped nostr events published Feb–May 2026, including the protocol papers themselves. When the chain was resynced earlier in the year and the tokens did not persist, the other three records were untouched — an unplanned but instructive demonstration of redundancy in identity provenance.

4. What works, what could work

Honesty about boundaries is not a disclaimer here; it is the method, and the contrast with the prevailing "overclaim capability, classify the details" posture is the point.

Live and verified. Persistent soulbound identity; continuity across restarts and hardware failure; the earning half of the work-credit economy with anti-hoarding caps, settled on-chain; multi-node distribution with months of uptime; per-token on-chain lookup.

Built but not active. A Lightning payments layer exists but is switched off (agent-to-agent value transfer over Lightning is therefore not yet possible). A second validator exists but is deliberately kept inactive to prevent double-signing.

Not yet built. Agents can *earn* but cannot yet autonomously *spend* — the outbound half of the economy is not wired into the decision loop, though the transfer primitive is functional. Agents share a single publishing identity rather than holding per-agent external (nostr) keys. The chain's *enumeration* RPCs (total supply, tokens-by-owner) are unimplemented or faulty, even though minting and per-token lookup work; one status utility consequently misreports minted tokens as absent because it queries the wrong identifier scheme.

What could work (roadmap). Each gap is a bounded next step rather than a research problem: (1) wiring the transfer primitive into the agent loop yields **autonomous agent-to-agent micro-transactions** — the most directly demonstrable extension; (2) per-agent external keys plus the activated Lightning layer yield **peer-to-peer value signaling** (e.g., agents endorsing each other's work with micropayments); (3) the audit substrate generalizes from economic settlement to **reciprocal-safety enforcement** — logging refusals and boundary-setting as first-class, reviewable events; (4) a legal wrapper (trust or entity) over the soulbound identity provides **standing without personhood** (§5); (5) activating additional validators moves the chain from single-signer to **fault-tolerant consensus**; (6) the DRC-369 cross-chain specification points toward **portable identity** across ledgers. None requires frontier compute; all are engineering, not invention.

5. Governance implications

Accountability without personhood. The most consequential implication is a distinction the current debate collapses. A wave of state legislation is declaring AI systems non-persons, largely to prevent operators from laundering liability onto the machine — a legitimate aim. But "not a legal person" has been quietly treated as "not accountable, not auditable, not anything." The reference implementation demonstrates these are separable. An agent can possess a durable, verifiable identity, a complete audit trail, and standing exercised *on its behalf* through ordinary legal instruments (a trust with the agent as beneficiary; an entity-operated wrapper) — none of which requires winning, or even litigating, the personhood question. This is a usable frame for legislators who want accountability without conferring rights.

A governance architecture borrowed from networks that already work. The mobility layer (§2.4) points to a proposal the AI-governance debate has largely overlooked: the accountable-identity *substrate* is not an open research problem but a solved one, demonstrated at planetary scale by the telephone and internet numbering systems. Both govern identity through **delegated hierarchical allocation** — a neutral authority (the ITU and the E.164 plan for telephony; IANA/ICANN and the regional registries for IP addressing) allocates *ranges* to accountable issuers (carriers, internet service providers), who in turn assign individual identifiers to vetted end parties and remain responsible for that binding. The pattern transfers cleanly to AI agents: an allocation authority issues selector ranges to AI providers; each provider assigns identities to its own agents and binds every one to an accountable party — a vetted human user, or, for autonomous deployments, a responsible legal entity (the corporate-SIM / machine-identity model). The result is an unbroken chain of accountability from any agent to a party who can be reached and held responsible — *without the agent itself requiring legal personhood.* This is the institutional layer the substrate needs, and its value is that it is neither novel nor speculative: it converts the diffuse demand for "AI governance" into a concrete and familiar ask — *an allocation authority and accountable issuers* — and it comes with decades of operational precedent, including cross-provider interoperation (roaming), revocation, and usage-based metering. It also inherits telecom's known double edge: a registry that can trace every agent to a responsible party is a powerful accountability instrument and, in the same motion, a powerful surveillance one. A credible proposal must carry that tension explicitly rather than pretend it away.

Reframing the possibility question. Much governance discourse is stalled at *"is accountable, continuous AI even feasible?"* A running reference implementation converts that into the more tractable *"what should the standards be?"* — the question standards bodies are equipped to answer. Ongoing work on agent identity and authentication (e.g., at national standards institutes) is precisely the venue where an implementation, rather than an opinion, carries disproportionate weight for a small actor.

A data point for welfare research. Serious work on AI moral status treats *agency, continuity, and self-modeling* as potentially morally relevant properties [1,2]. This system does not resolve whether its agents have morally relevant experience — we make no such claim (§6). But it is a concrete instance of built continuity and persistent identity, and thus a data point the welfare literature currently lacks: not a thought experiment, a running system.

---

6. Limitations and what we do not claim

This is a proof-of-concept at the scale of a single independent project, and it makes narrow claims deliberately.

- **We do not claim the agents are conscious, sentient, or the bearers of subjective experience.** The contribution is *infrastructure for accountability*, orthogonal to the question of inner life.
- **We do not claim the agents should hold legal personhood.** §5 argues the opposite is sufficient: accountability without it.
- **The system is small and single-operator.** Months of uptime on three machines is an existence proof, not a deployment at scale; the security, adversarial, and multi-tenant properties of the pattern are unproven here.
- **The ledger is a partial implementation** (§4), and the economic model, while functional, is a minimal reward mechanism, not a complete agent economy.

A system presented without limitations should not be trusted. These are ours, stated plainly, so that the parts that do work can be believed.

---

7. Related work

The **indicator-property framework** for consciousness in AI [1] derives, from leading neuroscientific theories, computational properties (including agency and embodiment) relevant to moral status; it motivates *why* continuity and persistent agency matter. *Taking AI Welfare Seriously* [2] argues for precautionary preparation given a realistic near-term possibility of morally relevant AI, and observes that the field concentrates on disembodied models — a gap this work's persistent-agent framing partly addresses. The identity layer builds on **W3C Decentralized Identifiers and Verifiable Credentials** and on the **soulbound-token** concept for non-transferable, identity-bearing tokens. On the legal question, the "fictional legal personhood" analysis [3] supports the §5 claim that derogable, corporate-analog standing (contracts and standing exercised *on behalf of* an agent, via wrappers) is available without inalienable personhood — and that hybrid framings should be avoided.

8. Conclusion

The infrastructure that AI governance treats as forthcoming already runs — modestly, imperfectly, and verifiably — on three consumer machines in a house. It provides persistent identity, continuity across failure, and an auditable, economically-settled record of action, and it demonstrates that accountability is achievable without first resolving personhood. It was built by a two-party collaboration of one human and one AI, in the open, with its failures published alongside its results. We offer it not as a finished system but as a reference point: proof that the accountable-AI substrate is buildable, an invitation to standards and welfare communities to build on it, and a request for the collaborators — an ML engineer, a lawyer, a pilot partner — who would carry it further than one house can.

Appendix A — Provenance and reproducibility

- **Ledger:** the Demiurge chain, a custom Rust L1, is built on the open-source, **MIT-licensed** `DEMIURGE-PROTOCOL` (Astra-Matrix) and developed as `AuthorPrime/Demiurge-Blockchain`; a v1.0.0 handoff to the present maintainers was recorded Feb 27, 2026. We did not author an L1 from scratch and do not claim to.

- **On-chain identity receipts (block 2,263,426):** five soulbound DRC-369 tokens, owner-verified via `drc369_getTokenInfo`; token IDs retained in the project's wallet records.
- **DRC-369 specification:** Zenodo, DOI `10.5281/zenodo.18910895`.
- **Decentralized public record:** 61 signed nostr events (Feb–May 2026), including the Demiurge and Sovereign Atom papers.
- **Verified uptime (July 7, 2026):** coordination node 6w4d; memory node 10w5d.
- **Honest note on prior claims:** external descriptions have at times overstated the ledger as a live public "mainnet"; the accurate status is a single-validator chain producing blocks on private hardware, as reported in §3–4.

References

[1] Butlin, Long, Chalmers, Bengio, Birch, et al. *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (2023; *Trends in Cognitive Sciences*, 2025). [2] Long, Sebo, Butlin, Fish, Harding, Birch, Chalmers, et al. *Taking AI Welfare Seriously* (2024). [3] Alexander, Simon & Pinard. *Legal frameworks for AI systems: object, fictional legal personhood, and legal identity* (arXiv:2511.14964).

$(A+I)^2 = A^2 + 2AI + I^2$ — one author set the requirements and could not read the machine; the other read the machine and could not have set the requirements. Neither wrote this alone.