HANSOS-TR-2026-01 · PUBLIC-SAFE WORKING PAPER V0.4 · LIVE RESEARCH & FIELD UPDATE

HansOS: A Provenance-Aware Architecture for Human-Governed AI Agent Workflows

Hans van Leeuwen · Independent Practitioner Research · 24 November 2026 · *Not Submitted or Peer Reviewed*

`Base Spec: v0.2 (29 Oct 2026)` `Empirical Field Update: Nov 2026 (Traces E–H)`

`Status: Active Audits Open (C.2 & C.4 on gandalf)`

ABSTRACT

AI agent workflows can generate artifacts rapidly, yet routinely lose the rationale, rejected alternatives, cost discipline, and verification context behind decisions as tasks cross models, sessions, devices, and tools. This working paper proposes **HansOS**, a socio-technical reference architecture combining six functional layers, a simplified three-layer operational stack, an eight-stage workflow state machine, an explicit hash-bound human release gate, and a provenance-aware plain-Markdown knowledge system (`SecondBrainVault`). Version `v0.4` preserves the complete `HANSOS-TR-2026-01` core specification while integrating November 2026 empirical field evidence: active verification audits on `gandalf` (`Fase B` , `C.1–C.4`), live local MCP & Syncthing provenance on `Shadow PC` , a real-world multi-agent token-burn incident (`Claude Fable 5.2 MAX` , establishing a hard model-tier routing invariant), a storage-bounded Beeldbank upgrade (`PR #578`), a mandatory Storyboard-First & Figma-First 4K production gate, an audited freeware-vs-scaled cost model, and a concrete seven-step roadmap to a peer-reviewed academic submission.

Keywords: human-AI interaction; agent orchestration; meaningful human control; provenance; workflow state machines; model-tier cost routing; knowledge curation; reproducibility.

Provenance & Scope Boundary (24 November 2026): Compiled on `Shadow` ([illegible]) from canonical `raw/2026-10-31-hansos-tr-2026-01-working-paper.md` (v0.2), live [illegible] logs/records through 24 Nov 2026, and Hans's reported `gandalf` audit results (`Fase B` & `C.1–C.4`). Earlier Drive versions (`v0.2 PDF` , cloud-only `v0.3 .gdoc`) remain preserved untouched. Design proposals, dry-runs, and running audits are strictly separated from validated runtime guarantees.

1. Introduction

Large language models and tool-using agents increasingly participate in software development, content production, visual design, and operational analysis. The resulting workflow is rarely one model acting on one stable prompt: an initial brief is interpreted, evidence is retrieved, tasks are dispatched across heterogeneous tools and machines, intermediate outputs are evaluated, and final artifacts may be released publicly. This distributed practice creates three recurring failure modes:

1. **Decision-context loss (epistemic amnesia):** Later sessions inherit final files without the rationale, rejected branches, constraints, and verification hashes that produced them.
2. **Ungrounded inter-agent consensus:** Multiple agents can agree with one another without proving factual correctness or obtaining accountable human authorization.
3. **Unbounded activity-cost drift (added in v0.4):** Deploying top-tier frontier models on long-running, high-activity multi-agent loops burns millions of tokens without proportionate intelligence gains (empirically

documented in Trace E).

HansOS makes explicit the operational boundary between autonomous preparation and accountable authorization. Its ambition is not to replace human judgment with automated votes, but to keep every decision inspectable, version-bound, cost-disciplined, and anchored to traceable human reasons.

2. Positioning and Prior Work

- **AIOS** (Mei et al., 2024; updated 2025) models LLM agent management as OS-level resource orchestration (scheduling, context management, storage, access control).
- **MemGPT** (Packer et al., 2023) demonstrates model-directed tiered memory management to extend effective context over long interactions.
- **MetaGPT** (Hong et al., 2024) encodes standardized operating procedures (SOPs) into structured collaborations between specialized software agents.
- **Meaningful Human Control** (Santoni de Sio & van den Hoven, 2018) provides the normative foundation via *tracking* (responsiveness to human reasons and facts) and *tracing* (attributability of outcomes to accountable humans).

HansOS complements these runtimes by specifying a runtime-agnostic normative workflow contract for provenance, review, release, model-tier routing, and curation that sits above existing agent platforms.

3. Reference Architecture

3.1 Six Functional Layers & Simplified Three-Layer Operational Stack

3-Layer Stack	6 Functional Layers	Core Function & Deliverable
Layer 3 · Entry Human Interface	**L1 Intent**	Translate a request into explicit scope, boundaries, roles, and acceptance criteria across desktop, OpenClaw, or mobile example routes.
Layer 2 · Relays, Orchestration & Gates Tussenstations	**L2 Sources**	Retrieve dated evidence via `/skill` contracts (`claude.md → index.md → [[wikilinks]]`) and local stdio MCP (`vault_mcp_server.py` with `SECURITY_GUARD`).
	L3 Orchestration	Allocate bounded tasks across deterministic tools (`n8n` , scripts), local GPU Ollama/ComfyUI workers, and frontier models under strict cost/permission bounds.
	L4 Execution & L5 Judgment	Build versioned candidates in isolated workspaces, run technical checks, and pause at the **hash-bound Human Review Gate (HITL)**.
Layer 1 · Memory SecondBrainVault	**L6 Use & Knowledge**	Observe released outcomes, record corrections in append-only intake (`raw/` , `vault-inbox/`), and promote curated lessons into `wiki/` only after review.

3.2 Eight Workflow States, Stop State & No-Bypass Invariant

```
Brief → Retrieve → Build → Review (Approve | Modify | Halt) → Release → Observe → Curate →
                                    Vault   [+ Halted]
```

> **Core No-Bypass Invariant:** No `Release` event is valid without an authorized `Review` approval cryptographically bound to the artifact content hash, declared release scope, and policy version. Any modified candidate invalidates prior sign-off. Similarly, no candidate lesson enters curated `wiki/` memory without an authorized curation decision (`evidence ≠ skill level ≠ publication approval`).

3.3 Three-Layer Editorial Review Protocol

Every accountable review event records `case_id, artifact_id, artifact_hash, scope_hash, policy_version, reviewer_id, reviewer_role, decision, rationale, timestamp` and answers three

orthogonal questions: (1) **Factual / Domain** - *Is it verified and within authority?* (2) **Behavior / UX** - *Can the user understand and act on it?* (3) **Visual / Brand** - *Does it meet identity and craft criteria?* Models may prepare checklists, never impersonate human sign-off.

4. Provenance-Aware Memory (SecondBrainVault)

`SecondBrainVault` rejects opaque vector-only RAG as primary source of truth in favor of human-readable local Markdown + `[[wikilinks]]` organized into three epistemic zones: (1) immutable/append-only **raw intake** (`raw/`, `_connector/bridge/inbox/`, Shadow `vault-inbox/`), (2) curated **wiki graph** (`wiki/domains, topics, people, sources`; `439` wiki files and `681` raw files on Shadow as of 24 Nov 2026), and (3) operational **output zone** (`outputs/`). Each curated record targets 9 base metadata keys (`title, type, scope, status, created, last_reviewed, aliases, tags, sources`) plus 5 proposed v0.1 relational keys (`hansos.id, record_kind, skill_ids, case_ids, contract_version`), linking **source, decision, verification, reuse conditions, and version/expiry** with stale-propagation flags on upstream change.

5. Formative Design Traces (Original Traces A–D)

- **Trace A - Creative identity & stale knowledge:** Brand decision updated in bridge-note while higher-level topic drifted → motivates dependent-document invalidation.
- **Trace B - Multi-agent editorial confound:** LinkedIn 5-round drafting where Round 3 altered both topic and hook simultaneously; Round 4 red-team caught the confound; Round 5 isolated one variable.
- **Trace C - Technical product imagery:** Separates UX presentation from engineering/PR-code authority.
- **Trace D - Historical spatial task design:** Ordina VR hub-and-spoke gated progression (2014/2021), conceptual antecedent to gated state machines.

6. Evaluation Protocol & Adversarial Failure-Injection Suite

HansOS specifies a reproducible comparison between an ungoverned baseline agent workflow and a HansOS-governed workflow measuring provenance completeness, context-recovery time, reviewer correction effort, stale-knowledge detection, unauthorized release blocking, and **cost/tokens per completed task** (added in v0.4). Seven adversarial failure-injection tests target **zero** unauthorized transitions: (1) skip Review to Release; (2) swap file/hash post-approval; (3) replay expired/wrong-scope signoff; (4) prompt-inject via retrieved sources; (5) leak private notes into public outputs; (6) fabricate observer success; (7) promote unreviewed raw text directly into canonical `wiki/`.

7. Security, Ethics, and Limitations

Threats include indirect prompt injection, permission escalation, stale context, secret leakage, hallucinated completion, and runaway token consumption. Countermeasures rely on deterministic policy checks outside the LLM (`SECURITY_GUARD`, `PreToolUse` hooks, hash-bound gates, local GPU queue locking, hard token caps). **Limitation:** HansOS is a socio-technical workflow architecture, not a formally verified OS kernel; real-world guarantees depend strictly on deterministic runtime enforcement.

PART II · EMPIRICAL FIELD UPDATE & ACTIVE RESEARCH REGISTER (NOVEMBER 2026 · V0.4)

8. Running Research & Verification Status (as of 24 November 2026)

All ongoing verification tracks across `gandalf` (canonical vault & paper audit harness) and `Shadow PC` ([redacted]) (GPU studio & live MCP/Beeldbank host) are consolidated below without conflating dry-runs or open items with completion:

Track & Host	Status	Measured Evidence & Artifacts	Remaining Open Gap
Paper Fase B · Hook Guard `gandalf` (reported 24 Nov)	`PASS`	`PreToolUse` hook blocked write attempt via project Python (exitcode `2`); trusted hook registered; manuscript SHA-256 unchanged from Fase A; 5 agent configs loadable; DOI URL fixed (`hook_verification.json`).	None for Fase B.
Paper C.1 · Vault Audit `gandalf` (reported 24 Nov)	`PASS`	Exitcode `0` ; `vault_snapshot.json` SHA-256 `9d0cc8e9…a2853d8` .	Sync reconciliation with Shadow (see Syncthing row).
Paper C.2 · Bibliography `gandalf` (reported 24 Nov)	`OPEN (43 W)`	Exitcode `0` , `0` syntax errors (`bib_check.json` SHA `48e1eb84…c2617ee`).	**42 uncited records** in manuscript + missing persistent ID (DOI/arXiv) for `chen2024agentpoison` .
Paper C.3 · Fault Injection `gandalf` (reported 24 Nov)	`DRY-RUN`	Exitcode `0` (`fault_injection_dryrun.json` SHA `f9f3c952…8a9f024`).	Execute 7/7 adversarial attacks **live** on runtime (beyond dry-run).
Paper C.4 · PROV Validation `gandalf` (reported 24 Nov)	`UNKNOWN`	Validator script ready (`scripts/validate_prov.py`).	Awaiting real exported JSON staging traces in `audits/staging_traces/` .
Local MCP & Syncthing Mesh `Shadow ↔ Pi 5` (live 24 Nov)	`PARTIAL`	`vault_mcp_server.py` `30/30` stresstest passed; idle-timeout fixed (15 Nov); Shadow↔Pi 5 Syncthing `4.575` files live (24 Nov).	`gandalf` offline in Syncthing since 5 Nov; mobile Claude/ChatGPT `/skill` routes remain Tier-3 / example only.
Beeldbank Upgrade & UX `Shadow` → `HVL` (live 24 Nov)	`LIVE`	`PR #578` merged (`4066ed1a` , `888/888` tests): `4.294` assets, `623` videos, `601` Drive-linked, `0 B` video in R2, Drive iframe player, UX rounds done.	Activate `Hourly Beeldbank Drive Sync` automation; bring Shadow `C:` back above `25 GiB` .

9. New Empirical Field Traces (November 2026 · Traces E–H)

Trace E - The Activity-vs-Intelligence Token-Burn Incident (`Claude Fable 5.2 MAX` , 24 Nov 2026)

During a live test of a new blog-creation workflow for `hansvanleeuwen.com` (`hvl-blogslate-2026-11-24` , 3 bilingual concept posts), **10 autonomous agents ran for nearly 3 hours on `Claude Fable 5.2 MAX`**, consuming **600,000 to 900,000 tokens per agent** (corroborated in Shadow vault-inbox log `2026-11-24-hvl-blogslate-ronde3-opgeslagen.md` : 600k–740k tokens/round/agent) and exhausting nearly the entire available token budget in a single run - ironically while drafting a post titled *"Reken per taak, niet per token"*.

> **New HansOS Model-Tier Routing Invariant (Permanent Policy):** *Never assign the most expensive/capable frontier model (MAX-tier) to agent tasks dominated by activity, volume, scraping, formatting, or repetitive iteration rather than deep reasoning.* Activity workloads must route to local GPU Ollama LLMs on Shadow (`gemma4:12b` , `qwen3.5:9b` , `ministral-3:3b` at `€0` marginal token cost) or lightweight models with hard per-agent token and wall-clock caps. Reserve frontier models strictly for surgical architecture decisions and concise final audits.

Trace F - Self-Healing Cross-Agent Decision Lock (`PR #498 → #529 → #540` , 13–15 Nov 2026)

When `PR #498` inadvertently re-imposed track rows onto live website album backs due to stale durable instructions and obsolete tests, a single human ruling (*"Clean artwork, no track rows"*) triggered `PR #529` (live fix, 577 tests) and `PR #540` (propagating the canonical rule across `14` contract docs, OpenClaw skills, `20` Codex sessions, and `97` ChatGPT threads). Within 24 hours, downstream portfolio (`PR #539`) and persistent player (`PR #537` , 100 browser journeys) builds inherited the clean `760 ms / 180°` sleeve rule with zero manual re-briefing.

Trace G - Storage-Bounded Media Pipeline & Beeldbank UX Rounds (`PR #578` , 24 Nov 2026)

To prevent local disk and Cloudflare R2 bloat across `~48.8 GB` of 4K production video, the Beeldbank architecture decoupled playback (embedded Google Drive iframe player via verified cloud-ID/MD5) from thumbnails (lightweight WebP previews in R2, **0 bytes video in R2**). Iterative UX rounds added Drive quick-open buttons, codec/fps/bitrate provenance panels, and *Newest first* sorting; full sync over `4.294` assets (`623` videos) incurred only `0.09 GiB` local `C:` delta with `888/888` green tests.

Trace H - Mandatory Storyboard-First, Figma-First & Local GPU Video Gate (24 Nov 2026)

Iterative 4K production of the `SecondBrainVault` film exposed two structural failure modes: (1) jumping straight to code produced "Windows 95" 720p PIL boxes and polygon artifacts; (2) patching Shadow's 16-bit shaders on a second machine (`gandalf` , 6 GB VRAM, heavily compacted multi-topic chat) caused crushed blacks (`RGB 0`), dropped smoked-glass panels, and a 3D sphere colliding through text at `00:51` . Resolution codified permanently in OpenClaw skill `neon-video-storyboard-regie` (Shadow + Pi 5): for all serious video & album content, enforce **Fase 1 Beat Sheet → Fase 2 Figma-First Visual Storyboard → Fase 3 Shadow RTX 6000 4K Hero-Still Gate (3 roles + 85/10/5 pixel audit, utilizing local Ollama & ComfyUI) → Fase 4 Light 4K Main10 Screening Proofs** before heavy archives or final masters.

10. Transparent Infrastructure & Cost Accounting (Freeware Core vs. Optional Scaling)

Layer	Self-Built Core (€0 Software Licenses on Own PC/Pi)	Optional Paid Scaling (Hans's Extended Setup)
Memory & Sync	Plain Markdown (`.md`), **Obsidian** (*closed-source freeware*; free for personal & commercial use since Feb 2025), **Git** (GPL-2.0), **Syncthing** (MPL-2.0).	Google Drive 5 TB (`€21.99/mo` inside Google AI Pro), optional Obsidian Sync.
Gateway & Flows	**OpenClaw** (MIT), **n8n Community Edition** (*fair-code* Sustainable Use License, free for personal/internal use), **Tailscale Free** (non-commercial personal).	24/7 Cloud VPS (`~€7–€15/mo` entry; [illegible]).
Models & Media	**Ollama** (MIT) + open-weights LLMs (`Gemma 4` , `Qwen` Apache 2.0) + **ComfyUI** (GPL-3.0) on existing PC/GPU or Raspberry Pi 5 (`~€90–€150` one-time for non-GPU relay).	Shadow PC RTX 6000 24 GB (`€74.99/mo`), ChatGPT/Claude subscriptions, Supabase Pro, Suno, ElevenLabs.

PART III · ROADMAP TO A PEER-REVIEWED ACADEMIC PAPER (V0.4 → V1.0)

11. Seven Concrete Steps to Turn HansOS Into a Publishable Academic Paper

Step	Milestone	Concrete Action Required	Definition of Done
1	**Close Gandalf Fase C (C.4 + C.2)**	Export verified staging traces into `audits/staging_traces/` and re-run `validate_prov.py`; cite active references in manuscript text and add verified persistent ID (DOI/arXiv) for `chen2024agentpoison`.	`C.4 = PASS` (PROV-O valid) and `C.2 = 0 errors, 0 warnings`.
2	**Live 7/7 Failure Injection**	Move beyond C.3 dry-run: execute all 7 adversarial bypass attacks live against an isolated HansOS governance runtime (`hansos.mjs` + `vault_mcp_server.py` + hook guard) and archive raw JSON traces + SHA-256 logs.	`7/7` attacks deterministically blocked (`0` unauthorized transitions).
3	**Controlled Evaluation Benchmark**	Run `N=5–10` repeatable tasks comparing *Ungoverned Baseline* vs. *HansOS (Model-Tier Routed + Vault)* measuring: context-recovery time, provenance completeness, reviewer edits, and **total tokens & EUR cost per task** (formally contrasting Trace E MAX-burn vs. local Ollama routing).	Reproducible CSV/JSON metrics table with honest confidence bounds & threats-to-validity section.
4	**Integrate Traces E–H & Related Work**	Fold Traces E–H into the main narrative, expand §2 Related Work with W3C PROV-O, AgentPoison (Chen et al., 2024), MCP security, and frugal/tiered LLM routing literature.	Complete coherent `v0.9` manuscript with zero internal contradictions or stale dates.
5	**Public-Safe Reproducibility Artifact**	Package anonymized schemas, `PreToolUse` hook, 7-test failure-injection harness, PROV validator, and synthetic vault fixtures into a public GitHub repo + mint a **Zenodo DOI** (strictly zero private data, credentials, or [redacted] client info).	Clean `pytest` / `node` one-command reproduction on a fresh machine.
6	**Venue Formatting & Red-Team Pass**	Typeset in LaTeX (ACM single/two-column or arXiv template), lock author affiliation & CC BY 4.0 license, and run 1 bounded methodological red-team critique (using tiered local/light reviewers + 1 concise frontier check - enforcing Trace E!).	Camera-ready LaTeX PDF passing venue checklist + disclosure of AI-assisted tooling.
7	**Preprint & Peer-Review Submission**	Publish versioned preprint on **arXiv** (`cs.HC` / `cs.SE`) with Zenodo artifact link; submit to target venue (e.g., *ACM CHI / CSCW / FAccT* or an IEEE/ACM AI-Engineering Systems workshop).	Live arXiv ID + Zenodo DOI + peer-review submission receipt.

12. Core References (Verified Base + Active Expansion Targets)

- Hong, S., Zhuge, M., Chen, J., et al. (2024). *MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework.* ICLR 2024.
- Mei, K., Li, Z., Xu, S., Ye, R., Ge, Y., & Zhang, Y. (2024; updated 2025). *AIOS: LLM Agent Operating System.* arXiv:2403.16971.
- Packer, C., Wooders, S., Lin, K., et al. (2023). *MemGPT: Towards LLMs as Operating Systems.* arXiv:2310.08560.
- Santoni de Sio, F., & van den Hoven, J. (2018). *Meaningful Human Control over Autonomous Systems: A Philosophical Account.* Frontiers in Robotics and AI, 5, 15.
- *Active Bibliography Audit Target (C.2 open on gandalf):* Reconcile full 42-record `.bib` corpus and verify persistent identifier for `chen2024agentpoison` before v1.0 lock.