Witnessed Systems and Bounded Agency

A research dossier. The question, the program, the evidence behind it, and the parts it does not establish.

Zain Dana Harper | Seattle, Washington | zaindharper@gmail.com |
ORCID 0009-0001-7175-5393 (https://orcid.org/0009-0001-7175-5393)

The question

A made system can be given operational standing. Its claims and its actions should stay bounded by evidence another party can re-derive, by evaluation independence that was recorded rather than assumed, and by human authorization that stays explicit where the consequences are real.

The word doing the most damage right now is "verified." A model produces an output. A harness transforms it. A tool acts on it. An evaluator scores it. A report summarizes the score. A lab or a regulator then treats the report as evidence about a deployed system. At every one of those transitions, identity, version, provenance, independence, and scope can be lost, and the claim keeps travelling. It stays accurate at one narrow layer while acquiring much broader meaning further down the chain.

So the questions I work on are the boring ones that decide whether any of it holds. Verified by whom. Against which criterion. Using which model, dataset, prompt stack, tool permissions, and wrapper. Did the verifier grade its own work. Can another party re-derive the claim. Does a result about integrity quietly become a claim about safety.

Seven open questions

1. How should claims made by frontier agents be scoped to criteria that can be independently witnessed and reproduced?
2. How should an evaluator record whether it graded its own work, inherited the claimant's assumptions, or depended on a shared failure surface?
3. How can a proof packet preserve provenance and semantics across adapters, model providers, organizations, and time?
4. How do we tell belief failure, perception failure, memory failure, value conflict, and action-selection failure apart in an agent?
5. Which properties of human cognition can inform experimental design without falsely equating people and models?
6. Where must human authorization stay irreducible once AI systems enter robotics, laboratories, and infrastructure?
7. How does verification infrastructure stay open, inspectable, and useful to people outside well-funded laboratories?

Five workstreams

1. Evaluation provenance and independent witness

Extend Witnessed Independence into a practical evaluation-provenance schema. It records evaluator identity, model and dataset versions, shared dependencies, negative controls, and whether the verifier inherited claims from the system it assessed. The receipt assigns no trust score. It exposes the structure a reader needs to judge how much independence a result actually had, and it gets tested across several evaluation frameworks rather than one.

2. Live-state verification and drift

Generalize EMET's narrow discipline into live-state monitoring, where an agent, model, policy, or robotic controller may drift between evaluation and deployment. Identity, integrity, behavior, and authorization are four separate questions and get four separate answers. MATCH, DRIFT, and UNVERIFIABLE stay real outcomes, so missing evidence is never laundered into confidence.

3. Proof-carrying agent actions

Develop Proof Packets into a minimal envelope for a consequential action: named inputs, model and tool versions, capability requests, policy checks, external observations, uncertainty, human authorization, and replay instructions. The test that matters is whether a packet survives a tool adapter and a multi-agent handoff without changing meaning.

4. Epistemic and cognitive dynamics

Study the gap between what an agent represents, what it says, and what it does. Behavioral experiments and causal analysis, with carefully bounded concepts from psychology and neuroscience, to separate uncertainty from confabulation, deception, memory failure, and policy conflict. Human parallels here are analogies and experimental inspiration. They are not claims of equivalence and the writing will not let them drift into one.

5. Embodied and robotic authorization

Move from language-only systems to physical action. What evidence is required before a robot may cross a safety boundary, how sensor uncertainty should propagate into action permissions, and how an independent witness stays advisory instead of quietly becoming a second controller. Early settings are controlled manipulation, environmental sensing, and field-robotics simulation.

What already exists

This is not a proposal for work that has not started. Each workstream extends something public with source, tests, and a permanent record. None of it is peer reviewed and none of it is a finished platform.

ARTIFACT	WHAT IT DOES	STATUS
EMET	Re-derives byte-integrity facts and refuses to emit trust, safety, or permission. Six boundaries separating an is-question from an ought-question.	Systems paper, four implementations, frozen 1.0.0 specification. DOI (https://doi.org/10.5281/zenodo.21230267)
Witnessed Independence	Records whether a verifier graded its own work.	Published preprint. DOI (https://doi.org/10.5281/zenodo.21232206)
Proof Packets	A derive-don't-trust envelope around an agent action.	Published preprint. DOI (https://doi.org/10.5281/zenodo.21231406)
BuildLang	Typed capability effects, so authority is explicit in the program rather than ambient.	Systems paper and a public compiler. DOI (https://doi.org/10.5281/zenodo.21231253)
Flywheel	The engine the rest plug into. Default-deny capability gate, sealed hash-chained receipts, transitive witness graph, oracle registry.	Public flagship. First preregistered confirmatory run completed 2026-08-04. Open
Conferred Existence	The philosophical ground: made minds, perception, memory, external anchors, legitimate standing.	Thesis, archived research corpus. DOI (https://doi.org/10.5281/zenodo.20773724)

The lineage runs in one direction. Conferred Existence asks what a made mind could legitimately be owed and be held to. The Conservation of Faithfulness asks what survives communication between differently perceiving bounded agents. Witness and Verification Under Bounded Rationality supplies the operating rule: a verdict reaches exactly as far as its criterion is independently witnessed and re-derivable, and past that boundary the system holds an unsupported bid rather than a fact. EMET turns that rule into software. The full set is on the publications page and the arguments are on the research index.

The program

Public Verification Commons

A small, open, technically rigorous layer that helps researchers, maintainers, auditors, and public-interest organizations tell what has actually been witnessed from what has merely been asserted. Twelve months, deliberately narrow, four outputs.

- **An evaluation-provenance receipt.** Claimant, evaluator, model and dataset versions, shared dependencies, negative controls, and whether the evaluator generated, selected, or transformed the evidence it scored. No trust score.
- **A proof-packet reference implementation** for tool-using agents, designed so adapters and multi-agent handoffs cannot silently change what a claim means.
- **A conformance corpus** covering ordinary cases, negative controls, stale versions, self-evaluation, shared-dependency failure, wrapper drift, incomplete evidence, and claims that are genuinely unverifiable. It rewards refusal to overclaim, not only successful verification.
- **A teaching and adoption layer**: a ten-minute reviewer route, example integrations, short governance briefs, and workshops for maintainers, independent researchers, and public-interest technologists. Permissively licensed throughout.

Twelve months

MONTHS	WORK
1 to 2	Scope, threat model, research review, partner interviews, benchmark design.
3 to 4	Evaluation-provenance schema and reference implementation.
5 to 6	Conformance vectors, negative controls, adapter tests.
7 to 8	Proof-packet demonstrator for multi-agent and tool-using systems.
9 to 10	Embodied-action preflight in simulation, with explicit non-claims about real-world safety.
11 to 12	Independent replication, public documentation, policy brief, teaching module, workshop.

A workstream stops or narrows if the people it is meant to serve cannot name a decision it improves. That rule is part of the plan, not a caveat attached to it.

What it costs

SCOPE	AMOUNT	WHAT IT BUYS
Minimum	$35,000	Nine months. Provenance schema, conformance corpus, one reference implementation, public report.
Recommended	$85,000	Twelve months. Research time, compute and hosted testing, independent technical review, two implementations, workshops, governance brief.
Expanded	$150,000	Adds a second engineer or research contractor, formal-methods consultation, an embodied-simulation track, broader integrations and accessibility work.

At the recommended tier the money goes to research time, compute and hosted testing, independent technical review, documentation and workshops and accessibility, and the ordinary administrative costs of running a project, with a contingency line. What a grant buys here is continuity and external review: the time to turn a scattered public body of work into infrastructure other people can test, and the budget to pay someone outside the project to try to break it.

What this does not claim

The most useful part of a dossier is the part that says where the evidence stops. Every row below is the language I use, and I use it consistently rather than upgrading it when the audience changes.

CLAIM	STATUS	HOW IT GETS SAID
The research record	Public DOIs, not peer reviewed	Systems paper, published preprint, research note, or archived research corpus. Never "peer reviewed" and never "published in" a venue that has not accepted it.
Researcher standing	Independent, evidence-supported	Independent researcher or systems researcher. Not professor, not academic researcher, not affiliated with a lab I am not in.
The upstream engineering record	Public GitHub record	Counted from the API, with the declined pull requests published beside the accepted ones. Open is not accepted, and a merged link-list row is not a contribution.
Field practice	Strong self-report	Eleven years of applied arboriculture and operational responsibility.
Anatomy, physiology, nutrition, botany	Self-directed and applied	Substantial self-directed study. Never licensed expertise, never clinical training, never advice.
Volunteer service	Self-reported, letters requested	Stated as self-reported until an organization confirms it in writing.
The confirmatory run	One preregistered study	Zero verdict disagreements across 2,646 certificate bodies, held-out selection uplift on the solvable family, and a null on the harder family that is kept and labeled uninformative. No capability uplift is claimed for any model.

CLAIM	STATUS	HOW IT GETS SAID
Infrastructure controls	Partly stubbed	The kill switch is off by default and cloud credential revocation is a stub pending live IAM wiring. Cross-layer correlation is heuristic today. Both are stated in the repository, not only here.

Standing and participation

Work only counts as public when somebody outside the project can act on it. These are the places where that has happened.

23 merged upstream | **20** repositories | **8** records with DOIs | **14** public engines

- **Coalition for Secure AI.** Individual contributor to Workstream 4, Secure Design for Agentic Systems, contributing secure-design patterns and a reference implementation.
- **Puget Sound Programming Python.** Invited talk on building verifiable AI workflows in Python, scheduled 2026-08-19.
- **Upstream open source.** Defect repair, missing validation, and test coverage merged into Datasette, tomlkit, pydantic-ai, DeepEval, pydash, grimp, and a set of smaller agent and evaluation projects. The full ledger, with the declined work, is on the portfolio.
- **PlantAmnesty.** Volunteer time each year since 2015 with a Seattle community network of horticulturists and arborists. Self-reported, letter requested.

Where this is going

The route here was not academic. High school finished in 2013 and everything since has been self-directed, learned from source code, primary sources, field work, and the consequences of shipping. That is a real gap in research training and I would rather name it than route around it.

So the plan runs on two tracks at once. The work continues either way, because it has for three years without funding. Alongside it I am returning to formal study, starting small and local rather than pretending a full load is compatible with the rest of the year, with the goal of the methods training and the supervision that independent work cannot supply on its own. Research fellowships, funded programs, and eventually funded graduate study are the direction. None of that changes what is on this page today, which is the point of writing the evidence down before the credentials exist.

What is held back

This dossier is drawn from a larger private package covering education planning, funding applications, health and disability access, benefits eligibility, and family circumstances. Those parts stay off a public page. They are relevant to a specific application reviewer under that application's own terms, and to nobody else.

What is held back: transcripts and grade estimates, medical and disability detail, benefits and eligibility screening, household finances beyond the project budget above, family circumstances, and the names and private details of the people I have volunteered with. Also held back are program-by-program application drafts and deadlines, which go stale fast and are mine to disclose when I submit them. If you are reviewing an application from me and need any of it, ask and you will get it directly.

Reach me

If you fund this kind of work, review it, maintain something it should integrate with, or want to argue with the premise, write to zaindharper@gmail.com. Arguing with the premise is genuinely welcome. A conformance corpus is only worth building if people who disagree with me try to break it.

Companion documents: full CV, resume, portfolio and upstream ledger, publications, research index.

Download this page: Dossier PDF. Printed from this page, so the two cannot drift apart.

Updated 2026-08-13. Program and funding sources were checked against their official pages on 2026-08-06; any deadline or amount should be re-checked at the source before it is relied on. Upstream counts are read from the GitHub API and gated by a test.