

Constitutional Self-Governance in ESS-MAI: Evidence-Bound Authority and an OPL2S Reference Capsule

Bledar Gjata Independent Researcher ORCID: [0009-0003-7854-266X](https://orcid.org/0009-0003-7854-266X) Draft: 23 October 2026. Not submitted to arXiv; not peer reviewed.

Abstract

This paper proposes constitutional self-governance as a design discipline for systems in which computational capability must not confer permission to act. The proposed necessary condition binds an authoritative effect to an executable law, jurisdictional competence, and sufficient action-specific evidence. ESS-MAI names the constitutional composition GCL and separates coordination, reasoning, and adjudication across LIGHT, QUANTUM, and SHADOW. OPL2S, also called POC007, instantiates a bounded obligation–promise–fulfilment inquiry. Its intended contribution is the composition of witnessing, scoped negative knowledge, independent verification, and evidence-directed continuation, rather than the invention of hashes, access control, or failure memory. A small dependency-free Rust reference capsule exercises missing-evidence, negative-superior, and positive-superior fixture cases; a fresh local execution passes twelve tests. Separate internal R47 records describe a larger derivative runtime, including model-scoped acceptance and refusal, persistent continuation, and corruption probes. These evidence levels are explicitly separated: the public capsule does not reproduce the private runtime, and its SHADOW function is not a hostile-caller verification boundary. R47 also accepts a valid older memory snapshot, demonstrating an unresolved freshness limitation. We provide falsifiable review questions and delimit proposed theory, executed observations, and unproven generality.

Keywords: runtime governance; authorization; evidence provenance; failure memory; obligation fulfilment; reproducibility; ESS-MAI; OPL2S.

1. Problem and contribution

A system may calculate a quantity correctly while lacking permission to apply it to a particular case. It may preserve an intact record that is irrelevant to the claim being adjudicated. It may also refuse an action correctly while failing to explain what additional evidence could permit progress. These are different problems. Conflating them can produce either unsupported acceptance or permanent refusal.

ESS-MAI proposes a constitutional organization of the relationship among law, competence, evidence, and execution. Here, a constitution means specified computational constraints, not a political prescription. An authority is a component's bounded jurisdiction; it is not an autonomous software agent with general discretion. GCL, the project's name for this composition, is not an experimentally discovered physical law. The word QUANTUM denotes a software reasoning layer and does not claim execution on quantum-computing hardware.

The proposed contribution has three parts. First, it separates failure to fulfil an obligation from the absence of executable authorization. Second, it records both successful processing and witnessed deficiencies without treating every deficiency as a refutation of the underlying proposition. Third, it connects a scoped deficiency or relevant objection to the next question, while retaining the adjudicator's responsibility to assess newly supplied evidence.

This paper describes that proposal and two distinct implementation artifacts. The historical reference capsule is public, self-contained, and small enough to examine in full. The larger AIGovern-Cyber development line is reported through bounded internal evidence and remains outside this capsule. The capsule is not presented as a substitute for independent reproduction of the larger implementation.

2. Foundations and attribution

Fail-safe defaults, complete mediation, and separation of protection mechanisms from policy are established foundations, including Saltzer and Schroeder's protection principles [1]. Consumer-side checking of supplied evidence also has substantial precedent in proof-carrying code [2]. OPL2S does not implement Necula's general proof-carrying-code framework, and its runtime tests must not be described as formal proofs.

Failure-driven learning is also prior art. Riesbeck studied remembering failed inferences and their corrections [3]; Hammond studied explanation of planning failures to anticipate later problems [4]. Later work explicitly addresses positive and negative knowledge sharing [5] and bounded failure-aware shared memory [6]. Consequently, neither remembering failure nor reusing negative material is claimed here as an invention. The candidate contribution is a particular arrangement of these ideas around case identity, jurisdiction, witnessed obligations, and controlled closure.

Bai et al.'s Constitutional AI trains language-model behaviour using principles and AI feedback [7]. ESS-MAI's present proposal concerns executable jurisdiction and effect boundaries. This is a difference in the mechanism under examination, not evidence that ESS-MAI invented constitutional terminology. Research on constitutive governance and runtime authorization is also relevant [8,9]. Comparative effectiveness against these approaches has not been measured.

BESA's promise terminology acknowledges the Albanian cultural concept of besa. Yad Vashem's historical testimony collection documents its association with keeping one's word [10]. The software contracts are a contemporary adaptation by this project. Cultural provenance does not establish algorithmic priority, an ancient computing protocol, or equivalence with the whole of Albanian customary law.

The ESS-MAI repository and its earlier executable-prior-art release provide project provenance [11,12]. The inspected Academic 1, 2, and 3 documents distinguish foundations, method composition, and emergent hypotheses. Their method binding includes a formula's location, purpose, inputs, outputs, authority boundary, and failure semantics. Reusing a formula in a new location does not alone establish novelty. Likewise, a timestamp or DOI establishes an identifiable disclosure, not a patentability finding or independent validation.

3. Proposed constitutional semantics

Let a denote an authority, l an applicable law, e the available evidence, and x a requested effect. The proposal requires:

```
AllowedEffect(a,l,e,x) => LawReady(l)
    AND Competent(a,l,x) AND EvidenceSufficient(e,l,x)
```

This implication is a necessary-condition specification. It neither proves that an implementation enforces the condition nor replaces adjudication with a formula. Sufficiency is action-specific: evidence permitting a witness to attest to absence does not necessarily permit a positive knowledge commit. A component cannot create additional competence by reporting confidence or by receiving a successful result from its predecessor.

The architecture assigns LIGHT coordination, QUANTUM reasoning, and SHADOW verification and final authorization within their respective boundaries. BESA binds the promised processing obligation; ASHT performs its reasoning role; NBES examines fulfilment. They do not vote. NBES may find that a validly produced record failed to reach its input, without concluding that no predecessor ever produced it. The trace should preserve both observations and their causal relationship.

Three categories require separate representation. A substantive counterexample opposes a specified proposition. Missing evidence establishes that a required condition has not been demonstrated at a particular point. HOLD expresses that the relevant execution cannot yet be undertaken under the applicable law. An integrity error is not ordinary ambiguity and should not be disguised as HOLD. Fail-closed describes withholding an unauthorized effect; it is not by itself a verdict about the proposition's truth.

When a predecessor fails to supply required evidence, the applicable law may still authorize the successor to examine the deficiency. The successor does not inherit success, but the evidentiary path need not end. When the executable law itself is absent or suspended, the waiting authority must not invent a replacement. Continuation and fulfilment are therefore distinct properties. A refusal can be correct control behaviour even though the original positive objective remains unfulfilled.

D-BAB is the proposed witness of preconditions and failures. It identifies the responsible authority, expected condition, observed basis, and missing evidence; it does not issue a veto or adjudicative verdict. LGC carries the processing trace. 0-NOI triggers an already authorized next operation without manufacturing evidence or expanding jurisdiction. These names describe roles whose enforcement requires concrete tests; their presence as module names is not sufficient evidence.

4. Evidence, inquiry, and memory

OPL2S distinguishes the usage-derived Xi/Yi and negative Xi/Yi material from KiE's scientific support and criticism, KiS and KiC. The intended composition examines one bounded case through authorized methods. Neither method creates the other's competence, and neither independently supplies final closure. A scientific result valid only inside a declared mathematical model must not be promoted into evidence for a broader empirical claim.

Missing historical candidates must not become invented scientific counterexamples. Conversely, a relevant scientific counterexample cannot be erased merely because it obstructs acceptance. Where an obligation is indispensable, another route cannot waive it. A common presentation contract must bind each method's request, scope, permission, output, absence, and trace to the same case before the relevant adjudication consumes them.

For candidate vectors measured once per unique admitted candidate, the author's proposed aggregation separates the polarities:

$$\begin{aligned} P_bar &= (P_1 + \dots + P_n+) / n+ & (n+ > 0) \\ N_bar &= (N_1 + \dots + N_n-) / n- & (n- > 0) \end{aligned}$$

Each mean is defined only for a nonempty set. An empty set is a missing-evidence state, not an observed zero vector. Candidate admissibility, units, and comparison remain contractual obligations. These equations describe the proposed vector method; the historical public capsule instead uses deterministic scalar fixture masses and does not implement the full vector stack. Averaging also cannot conceal a decisive counterexample that the applicable law requires to be addressed explicitly.

The intended iZ output combines the result of current processing with a question grounded in an unresolved obligation or relevant objection. The subsequent i0 is user input, not a prediction of the user's answer. The question narrows the requested evidence; it does not make the response true. This distinction is central to the claim that fail-closed behaviour can support progress without pretending that every exchange deserves positive acceptance.

The persistent pd_mp component is a reference map to scoped records owned by the SHADOW Vault. It is not a second reasoning engine or an independently authoritative knowledge store. History helps retrieve relevant candidates and reconnect an inquiry after LIGHT or QUANTUM restarts. Current adjudication must examine the evidence actually presented for the current case. Persistence does not authorize indiscriminate reuse of every earlier negative record.

SHA-256 bindings identify encoded content and detect inconsistencies under the stated comparison assumptions. They do not authenticate a writer, establish factual truth, or prove that a coherent file is the latest version. Similarly, zero-copy transport and non-duplication of authority are separate engineering properties: preventing a data copy does not, without additional controls, establish permission boundaries.

4.1 Adaptive hardware control as a separate contract

The proposed hardware extension applies the same competence-and-evidence discipline to permission to execute. A resource constraint is not a semantic objection to a user's thesis. Hardware causes and consequences therefore need their own identified trace rather than being added to the proposition's negative evidence.

R47 includes ordinary engineering estimates of the form

```
P_cpu_est = P_TDP * u
P_mem_est = k * M_used
T_est = T_ambient + R_thermal * (P_cpu_est + P_mem_est)
```

where load u is a normalized fraction and units must match the configured coefficients. These are models, not newly discovered physical laws. A reported sensor value may replace the estimate when available. The implementation distinguishes mathematical thermal operation from observed hardware state and labels the physical effect as unproven where it has not been established.

Resource adaptation includes an integer RAM budget of $\text{floor}(4F/5)$, for reported free bytes F , when the relevant safety predicate permits it; otherwise the budget is zero. This preserves a specified buffer without a floating-point overshoot. It does not reserve physical RAM against all other processes or demonstrate a speedup. Permission to deepen processing remains separate from epistemic acceptance.

A material limitation remains: the energy reader labels an unavailable-power fallback CPU_ESTIMATE and computes a dimensionless margin from CPU load. That value can still feed the existing policy. Thus, current R47 does not establish a universal rule that missing physical energy measurements force HOLD. The stronger theoretical requirement is to distinguish an authorized estimator policy from a requirement for actual physical evidence, and to withhold an effect when its own contract lacks a sufficient basis. The twelve public capsule tests do not exercise this hardware subsystem. No measured energy saving, sensor authenticity, or hardware-safety guarantee follows from these formulas.

5. Artifacts and experimental procedure

The public reference release is a publication bundle version 0.1.2 containing the historical POC007 0.1.1 documentation line and its unchanged Cargo crate version 0.1.0. Those numbers identify different layers, not three progressively more complete implementations. The capsule has no external Rust dependencies and contains forty fixture candidates: ten each for $X+$, $Y+$, negative X , and negative Y .

Three deterministic cases exercise its bounded model. One requests an unavailable exact x11 candidate and produces a fail-closed state plus an evidence question. A second supplies an exact pair with greater negative mass and produces rigorous-negative advancement while retaining positive context. A third has greater positive mass and produces constructive advancement while preserving negative critique. These are fixture classifications, not natural-language scientific reasoning or durable production transactions.

The fresh R48 execution used locked, offline Cargo tests and release compilation. Twelve Rust tests passed, with zero failures, ignored tests, or warnings; there were zero doctests. The release executable produced the three expected outcome classes. A separate Python oracle agreed on those classes. Agreement is a useful consistency check, but an oracle implementing the same specification can share the same mistake.

A crucial limit is visible directly in the public code: `shadow_verify` receives an `AshtTrace` and trusts its outcome while comparing caller-supplied digest fields. It does not independently re-run reasoning against the Vault. Public, clonable structures are not a protected authority boundary. The capsule therefore illustrates the method under its supplied execution path; it does not demonstrate resistance to arbitrary forged internal callers. This limitation is not repaired by the larger runtime's results.

The later R47 evidence concerns a separate frozen derivative containing 730 files and 17,309,641 bytes. It was tested on Windows with the stable x86_64-pc-windows-gnu toolchain, two build jobs, and warnings denied. The

reported commands were:

```
cargo test --locked --offline --workspace --features light-platform/c_kernel,quantum-  
platform/hw_kernel  
cargo build --locked --offline --release --workspace --bins --examples --features light-  
platform/c_kernel,quantum-platform/hw_kernel
```

R47 reports 1,789 Rust tests and four doctests passing, zero failures, zero ignored tests, and zero warnings. Its release build exited successfully. The elapsed test/build times, approximately 180.5 and 166.6 seconds respectively, include compilation and are not per-inference latency benchmarks. These are internal execution observations, not independent replication and not counts from the public capsule.

6. Reported results and counterexamples

The R47 native probe recorded 82 checks. Two model-only inputs used UT1 with Julian-date parts 2400000.5 and 54388.0, and a tolerance of 10^{-12} radians. A claim of 0.40228372400281387 radians produced native (1,1); a claim of 0 radians produced native (0,0). The calculation is scoped to the era00 mathematical model, whose astronomical foundations are external to ESS-MAI [13]. It does not authenticate the caller's UT1 measurement or establish a physical observation.

Within this implementation, these pairs identify completed positive and negative verdict paths. They do not imply that every intermediate state is binary or that all HOLD states are negative knowledge. The author's earlier theory note stated that complete (1,1) closure remained in development. R47 updates that status only for the specified numerical cases; it does not close the general claim.

A separate continuity probe recorded 100 checks and seven new operational records, without new native verdict transactions. Structured clarification reduced outstanding fields, while simple completion of those fields did not automatically produce (1,1). Exact retry did not append another record. Altered retry was refused, and retrying an old scientific answer preserved the newer inquiry reference. Unlabelled prose did not automatically populate all evidentiary fields.

The fault probe recorded 57 observation checks, including detection of incorrect references, stale-parent requests, changed retries, committed-record corruption, and recovery of incomplete tails. Six rejected requests returned exit code 70 and left the closed WAL byte-identical. Internally coherent, rehashed references were still checked against the owner's exact record. A separate native test rejected an altered numeric result even when supplied hashes and relation fields were mutually consistent.

However, one intentionally examined limitation is decisive: replacing memory with a coherent older WAL was accepted. The older record could be retrieved and the newer record was absent. No independently protected monotonic head established freshness. Passing this observation check means the test confirmed the limitation, not that rollback was prevented. Integrity within a file and freshness of the selected file are different properties.

Some rejection paths stop before a trustworthy case or Vault is available. They report stderr and an exit status rather than inventing an iZ or appending to corrupted storage. Thus, the aspirational rule that every failure is traceable does not establish that every failure is durably recorded inside that same Vault.

7. Scope of assurance and falsifiable review

The evidence supports a bounded governance-core investigation. It does not establish a complete cybersecurity product, resistance to a hostile operating-system administrator, general scientific understanding, or universal formal correctness. The internal launcher does not establish executable authenticity merely by checking that a file exists. Local mailbox permissions do not themselves provide independent client identity. NIST's distinction between authentication and authorization is relevant to these remaining boundaries [14].

No claim of novel physics, measured energy savings, or guaranteed hardware acceleration is made. Hardware-related regression checks concern declared execution conditions; they do not authenticate sensors,

prove thermal safety, or make logic become physically true. Likewise, the word living refers to persistent and revisitable records, not consciousness or biological memory.

Independent review should target specific falsifiers. Can an invalid candidate or substituted scope obtain constructive closure? Can a successor inherit success without receiving the required evidence? Can an unchanged operation identifier accept changed content? Can a procedural failure masquerade as a semantic counterexample? Can a valid negative case retain useful positive context, and can a valid positive case retain the relevant objection? Each report should identify the artifact version, complete input, command, environment, expected result, actual result, and minimal reproducer.

Reviewers should distinguish results reproducible in the public capsule from questions requiring the private runtime. A public capsule failure must be reported against that capsule; an internal R47 claim cannot be established by its twelve tests. A stronger future evaluation would include independent replication, domain expansion, adversarial caller tests, explicit freshness protection, and measurements of retrieval and validation costs. Merely adding more repeated exchanges is not evidence that acceptance is becoming more justified.

8. Availability, authorship, and research status

The prepared public release targets the standalone AIGovern-Cyber-POC007 repository linked from ESS-MAI, with a matching Hugging Face distribution identified by release hashes. Target locations are <https://github.com/gjatalegacy-create/AIGovern-Cyber-POC007> and <https://huggingface.co/datasets/gjata-legacy/ess-mai-poc-007-aigovern-cyber>; the release record, rather than this draft, establishes whether uploading has completed. Its scope is the runnable reference capsule, documentation, provenance, and selected evidence. The full AIGovern-Cyber runtime, product integration, private development history, and operational material are withheld from this capsule for further product development. They are not hidden dependencies of the capsule. Withholding them limits independent verification of the larger-runtime claims and is disclosed explicitly.

Bledar Gjata supplied the architecture, theoretical framing, and research direction. AI-assisted tools contributed software work, analysis, and editorial preparation under his direction. The author remains responsible for the final claims, citations, and release decisions. No AI system is named as a scholarly author, no institutional endorsement is implied, and no independent academic validation is claimed. This document is a manuscript draft, not an arXiv publication or acceptance notice.

The current contribution is a testable design proposal and an inspectable reference implementation. Whether its particular composition yields a distinct research advance remains a question for comparison and critical reproduction, rather than a conclusion obtainable from project terminology or passing tests alone.

References

- [1] J. H. Saltzer and M. D. Schroeder. The Protection of Information in Computer Systems. Proceedings of the IEEE, 1975. [Author-hosted text](#).
- [2] G. C. Necula. Proof-Carrying Code. POPL, 1997. [Author's publication list](#).
- [3] C. K. Riesbeck. Failure-driven reminding for incremental learning. IJCAI, 1981. [Proceedings](#).
- [4] K. J. Hammond. Learning to Anticipate and Avoid Planning Problems through the Explanation of Failures. AAAI, 1986. [Proceedings](#).
- [5] Y. Ba et al. Cautiously-Optimistic Knowledge Sharing for Cooperative Multi-Agent Reinforcement Learning. AAAI, 2024. [DOI](#).
- [6] H. Wang. Negative Knowledge as Failure-aware Shared Memory for AutoResearch. arXiv:2606.21024, 2026. [Preprint](#).

- [7] Y. Bai et al. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073, 2022. [Preprint](#).
- [8] B. A. Hu and H. Rong. Is Decentralized AI Governable? From Regulative Policy to Constitutive Protocol. arXiv:2605.24538, 2026. [Preprint](#).
- [9] E. Calboreanu. LATTICE: a governance-first architecture for authorized autonomous AI operations. Frontiers in Artificial Intelligence, 2026. [DOI](#).
- [10] Yad Vashem. Besa: A Code of Honor — Muslim Albanians Who Rescued Jews During the Holocaust. [Historical testimony collection](#).
- [11] B. Gjata. ESS-MAI. [Project repository](#). The current publication bundle must be cited by its exact version and manifest; a branch URL alone is not an immutable artifact identifier.
- [12] B. Gjata. ESS-MAI Executable Prior Art v1.1.1. Zenodo, 14 September 2026. [DOI](#). This earlier release covers POC003–POC006; its DOI is not a DOI for the new POC007 capsule or this manuscript.
- [13] IAU SOFA Board. Standards of Fundamental Astronomy, Library Issue 2023-10-11. [Official software and documentation](#).
- [14] S. Rose, O. Borchert, S. Mitchell, and S. Connelly. Zero Trust Architecture. NIST SP 800-207, 2020. [DOI](#).

Artifact evidence note

R47 observations above derive from the retained internal REPORT.txt, FINAL.json, and their referenced execution records. The frozen manifest SHA-256 is `'f1ebf1f6d5dad9d88dd75e0d0b9cde6e524cee91f6b3189f2d02da1a43196a16'`; the R47 evidence manifest SHA-256 is `'8ec0c66d2e9469e4e6877be84707b099cc98ba7269a593726b3e3ba368f50498'`; FINAL.json SHA-256 is `'f183c718baf6c9d4c222d5f45639f51c5af1febbeaaa9ed349ecbecd9e823fe3'`. These commitments identify the retained files. Their presence does not let a reader reconstruct withheld source or independently verify its execution. R48 public results are recorded separately in the publication bundle and must not be combined with R47 as one test suite.