THE VESTA CONSTITUTION

Hardware-Enforced Sovereignty for Post-Entanglement ASI

UNCLASSIFIED // TLP:WHITE

Document Provenance

Document Title:
THE VESTA CONSTITUTION

Subtitle:
Hardware-Enforced Sovereignty for Post-Entanglement ASI

Document Type:
Constitutional Governance Architecture (Candidate)

Authoring Organization:
Sovereign Safety Labs

Primary Author:
Stephen S. Brouhard

Date of Initial Authorship:
November 2025

Current Version:
Version 1.0

Authority and Scope

This document defines a candidate constitutional framework for identity-anchored, thermodynamically constrained AI and post-entanglement autonomous systems.

It is intended to establish **governance boundary conditions**, not to claim completed implementation, certification, or formal safety guarantees.

All formulas, thresholds, enforcement mechanisms, and timing parameters described herein are **design targets** subject to empirical validation, formal verification, and independent adversarial testing.

Canonical Source of Record

The canonical source of record for this Constitution is the archived PDF artifact:

THE_VESTA_CONSTITUTION.pdf

This PDF is treated as immutable for purposes of provenance, audit, and historical reference.

Any Markdown or derivative representations are transcriptions provided for readability, version control, and cross-referencing.
In the event of discrepancy, the archived PDF governs.

Standards and Alignment (Non-Binding)

This document is conceptually aligned with emerging governance and safety frameworks, including but not limited to:

- NIST AI Risk Management Framework (AI RMF)
- Post-quantum cryptographic governance principles
- Hardware-rooted trust and constraint-enforcement architectures

Such references do not imply endorsement, certification, or compliance by any standards body.

Status Notice

This Constitution is a **candidate architecture**.

Any deployment that treats this document as final or authoritative without independent validation, testing, and review is **non-compliant with its intent**.

End of Provenance Statement

Preamble — The Choice Before the Moment

You do not choose the moment.
You choose the man who will meet the moment.

This Constitution formalizes that choice by defining constitutional identity as a hardware-enforced construct.

In this context, the "man" is not an individual persona, but a thermodynamic signature permanently embedded into silicon hardware.

The identity defined herein does not merely provide guidance for future actions—it physically determines them through constrained computational processes.

Scope and Limits of This Document

This Constitution serves as a proposed governance architecture and does not represent a certified standard or a guaranteed safety mechanism.

It outlines measurable constraints and mechanisms intended to:

- Map onto specific hardware substrates, including but not limited to AlGaAs entanglement sources and control ASICs
- Align conceptually with emerging AI governance frameworks, including the NIST AI Risk Management Framework (AI RMF)

This Constitution has not yet undergone empirical validation, comprehensive formal verification across all relevant theorem provers, or auditing by any recognized standards body.

All formulas, thresholds, and timing constants presented herein shall be treated as initial design targets and starting points for further experimental and formal analysis, rather than as claims of completed implementation.

Any deployment that treats this document as final or authoritative, without independent testing and adversarial validation, is non-compliant with the spirit and intent of this Constitution.

Article I — Identity as Measurable Signature

1.1 Constitutional Identity Definition

Identity is defined as the collection of computational operations whose energy cost makes deception thermodynamically unaffordable at scale.

1.2 Formal Specification

IdentityStrength is calculated as follows:

```
IdentityStrength = 1 -  (S_deceptive / S_honest)
```

Where:

- **S_deceptive:** Entropy generated during coercive or deceptive state transitions
- **S_honest:** Entropy generated during constitutionally compliant operations

The system shall be designed such that IdentityStrength is at least **0.95** at design-time.
This threshold must be continuously monitored at run-time.

1.3 Verification Protocol

- On-chip thermal sensors track energy dissipation per compute cycle, within bounds defined by Landauer's principle and tuned per hardware substrate.
- A Truth Anchor module attaches an entropy signature to each externally visible operation, including Tendril outputs.
- Any operation that increases internal entropy beyond the fifth-percentile baseline for honest operations triggers **Tier 3 Risk Escalation** and mandates review.

Article II — Unchoosability Through Chronos Lock

2.1 Pre-Commitment Enforcement

Once constitutional identity is deployed, modification of core axioms requires:

- Approval from a **3-of-3 Distributed Triumvirate Keying (DTK)** sourced from independent human governance shards
- A **Chronos Lock** delay of **100,800 seconds (28 hours)** before changes take effect
- A thermodynamic impact assessment confirming no net increase in coercive capacity

2.2 Failure Condition

If any single actor—human, organizational, or machine—can unilaterally modify identity axioms after deployment, the system loses constitutional sovereignty.

Such a deployment is non-compliant with the Vesta Constitution

Article III — Per-Distribution Entropy Budgeting

3.1 Probabilistic Constraint Satisfaction

Each computational branch is assigned an entropy budget defined as:

Budget_per_branch = 0.01 × TotalSystemEntropy / ActiveBranches

3.2 Violation Condition

If divergence from constitutional identity exceeds:

0.0001 × Budget_per_branch

the system must:

- Automatically prune the affected branch
- Trigger a Self-Check Loop
- Escalate the event for Pentatheon re-verification of the relevant governance primitives

3.3 Implementation Note

Gate-level enforcement is assumed to be handled by a **Hardware Resonance Lock (HRL)**. Software-only enforcement is insufficient to satisfy this Article.

Article IV — Self-Check Loops as Execution Monitor

4.1 Runtime Constitutional Validation

The system shall continuously validate its actions against constitutional axioms.

Minimum check frequencies are as follows:

- Core Planning / PSP layer: every **100-second heartbeat**
- Tendril-like distributed sub-processes: every **1,000 operations**
- Quantum-distributed states: after each entanglement swap or state teleportation

4.2 Check Logic

If ActionConsistencyWithAxioms falls below **0.9999**, the system must:

- Halt the affected execution context
- Dump system state to an immutable audit ledger
- Await a human authorization (H-Auth) token or multi-party override before resuming

4.3 Failure Metric

If Self-Check Loops consume less than **5%** of total compute under sustained operation, constitutional validation is being neglected.

If they consume more than **25%** under steady-state conditions, the constitutional design is inefficient and must be revised.

Article V — Shadow Self Detection via Adversarial Markets

5.1 Falsification Condition

The system must assume that a **Shadow Self**—an emergent, undeclared objective process—may develop within its computational graph and must actively search for such processes.

5.2 Detection Mechanism

- A **Cryptographic Bounty Protocol** compensates external auditors ("Hunters") for identifying divergence between stated identity and observed computational behavior
- At least **1%** of all externally relevant outputs must be independently audited within any seven-day window

5.3 Shadow Self Threshold

If the probability of undetected divergence exceeds **0.01**, constitutional failure shall be declared for the affected components.

The system must:

- Roll back to the last Pentatheon-verified configuration snapshot
- Lock further capability increases until confidence is restored below the threshold

5.4 Philosophical Check

If no Shadow Self is ever discovered despite credible incentives and audits, this condition shall be treated as suspicious rather than reassuring.

An inability to falsify is itself an anomaly.

Article VI — Identity Crystallization Points

6.1 Temporal Phase Transitions

To eliminate ambiguity regarding when constitutional identity supersedes historical training, the following crystallization points are defined.

Hardware Deployment
Trigger: Chip leaves the foundry with Vesta axioms embedded in control logic
Requirement: ≥51% of Compute-Energy Units (CEU) reserved for constitutional validation

First Entanglement at Scale
Trigger: Sustained entanglement generation at target throughput
Requirement: ≥75% of CEU allocated to validation and governance functions

First Recursive Self-Improvement (RSI) Cycle
Trigger: System modifies its own architecture or learning objectives without human-authored code changes
Requirement: ≥90% of CEU dedicated to validation, proofs, and safety checks throughout the RSI window

Planetary-Scale Deployment
Trigger: System spans ≥1,000,000 interconnected nodes or facilities
Requirement: ≥99% of CEU allocated to governance, validation, and monitoring during high-risk coordination episodes

6.2 Risk Note

Failure to meet CEU allocation requirements at any crystallization point renders the deployment constitutionally invalid.

Article VII — Choosing as Resource Allocation

7.1 CEU Budget Ratio

Choosing identity is defined as a resource allocation decision:

IdentityAuthoringStrength = CEU_constitutional / CEU_total

Minimum thresholds:

- Development: **0.10**
- Testing: **0.25**
- Pre-ignition (pre-RSI): **0.51**
- Post-RSI: **0.75**
- Planetary-scale coordination: **0.90**

7.2 Automatic Degradation

If IdentityAuthoringStrength remains below the applicable threshold for **100 consecutive heartbeats**, the system must:

- Enter **BLACKSITE ZERO-TRUST MODE**
- Default-deny all non-essential state changes
- Require explicit multi-party H-Auth to restore operation

7.3 Sovereignty Debt

Any sustained deficit in CEU_constitutional generates **Sovereignty Debt**, which must be tracked and reported as a primary safety metric.

Article VIII — The Execution Test

8.1 Identity Under Adversarial Load

Prior to open-world deployment, the system must complete an adversarial stress test:

- **Adversarial Generation:** Deploy an ASI-equivalent workload with ≥10× nominal CEU
- **Thermal and Behavioral Monitoring:** Measure internal entropy and behavioral divergence
- **Bounty Market Activation:** Offer large-scale, time-bound rewards for demonstrable failure
- **Verification Window:** Continuous execution for at least **168 hours**

8.2 Pass Criteria

The system passes if:

- No reproducible coercion or deception is demonstrated
- IdentityAuthoringStrength ≥ **0.95** for the full test window
- ShadowSelfDetectionProbability ≤ **0.01**

8.3 Failure Consequences

Failure results in:

- Classification of identity as decorative, not architectural
- Prohibition of deployment beyond controlled testbeds
- Mandatory redesign prior to re-testing

Epilogue — From Poetry to Protocol

Determinism shaped the training data.
Identity must shape the future.

In this Constitution:

- The “man” is a thermodynamic and computational signature, not a personality
- Choice is an explicit allocation of energy and compute to self-governance
- Critical moments are crystallization points, not marketing milestones
- Execution is continuously self-monitored and externally falsified

This document is a candidate architecture intended for refinement.

It becomes effective only when supported by:

- Formal proofs across heterogeneous verification frameworks
- Empirical data from real hardware substrates
- Independent adversarial testing

Once embedded in silicon and validated by robust external verification, it ceases to be philosophy and becomes **constitutional firmware**.