MONOGRAPH 01

On Agentic Containment

Explicit Permission Boundaries in Automated Systems

Nathaniel Byrd · Principal Systems Architect
Working draft v0.1 · Revised 10 November 2026

Proposed architecture. No deployed benchmarks or independent validation are reported in this draft.

1. System Rationale and Operational Doctrine

Automated reasoning becomes consequential when it can change records, direct resources, or act across external systems. An agent's ability to propose an action must therefore remain distinct from its permission to perform that action. Agentic containment is the proposed discipline of maintaining that distinction through explicit boundaries, accountable authority, and inspectable evidence.

The organizing hypothesis is a mind-body connection: physical discipline and optional body feedback may inform how an operator supervises cognitive and agentic work. This is a design hypothesis, not a demonstrated physiological control mechanism. The five-layer architecture below extends from a somatic foundation to cognitive coordination and security governance.

Three mandates organize the proposal:

- Permission-bounded execution separates reasoning from authority to change external state.
- Somatic-cognitive coordination treats body feedback as contextual information, never as consent or proof of intent.
- Evidence-authority alignment requires both an authorized action and evidence of its actual outcome.

NIST's zero-trust guidance rejects implicit trust based solely on network location or asset ownership and distinguishes authentication from authorization. Those principles inform the permission boundary here; they do not establish the effectiveness of this architecture. [1] The NIST AI Risk Management Framework provides voluntary guidance for incorporating trustworthiness into AI design, use, and evaluation, not certification of the proposal. [2]

PRIS 2.0, Evergreen OS, Mindscape, GUARDIAN 2.0, and CIPHER are proposed architecture labels, not verified implemented products. This draft reports no benchmarks, deployed performance measurements, or independent validation.

2. The Five-Layer Somatic-to-Cognitive Hierarchy

The biological hierarchy is an architectural analogy. “Brain,” “consciousness,” and “conscience” describe responsibilities, not claims that software possesses biological faculties or moral judgment.

Layer 1: Somatic Foundation and Biometric Baseline

This layer considers posture, motor load, electromyography (EMG), and heart rate variability (HRV) as optional contextual signals. The hypothesis is that operator feedback may support requests to pause, reduce workload, or obtain additional review. It does not establish that biometric monitoring prevents burnout, cognitive fatigue, or model hallucination.

HRV and EMG are not consent, reliable standalone diagnoses, or proof of cognitive intent. A physiological change cannot authorize an action or establish why an operator acted. Collection requires informed consent, a defined purpose, data minimization, restricted access, and bounded retention. Missing or declined telemetry must not be interpreted as consent, misconduct, or a trustworthy baseline.

Layer 2: Autonomic and Reflex System — PRIS 2.0

PRIS 2.0 denotes proposed deterministic reflex controls separated from generative reasoning. Its role is to evaluate immediate safety conditions and apply bounded, preauthorized responses, such as withholding a write or stopping an operation. A reflex remains subject to policy and audit requirements; low latency does not create authority. No response-time guarantee is claimed.

Layer 3: Executive Brain and Memory — Evergreen OS

Evergreen OS denotes proposed task coordination, memory, and event correlation. NATS or MQTT could carry messages between components; the original “vagus nerve” event-bus analogy expresses coordination rather than a physiological equivalence. A proposed vector-memory component, called the “Vector Hippocampus,” would support retrieval without making retrieved material authoritative. Stored state must retain source, timestamp, and uncertainty rather than acquire credibility through repetition.

Layer 4: Digital Consciousness and Intent — Mindscape

Mindscape denotes the proposed presentation and scenario-modeling layer, potentially including spatial or extended-reality views. Explicit operator input and contextual cues may inform candidate interpretations. Biometric or spatial cues alone cannot establish intent. A scenario remains a model until separately documented, authorized, executed, and verified; a persuasive display does not advance its permission state.

Layer 5: Conscience and Security — GUARDIAN 2.0 / CIPHER

GUARDIAN 2.0 denotes proposed policy evaluation, risk signaling, and containment decisions. CIPHER denotes proposed cryptographic support for protected communication and provenance. These controls would constrain state changes and preserve inspectable records. Their effectiveness depends on implementation, key management, endpoint coverage, and evaluation—not on the labels or the hierarchy itself.

3. Permission States and Consequential Handoffs

The protocol must distinguish five states:

```
CANDIDATE / MODEL -> DOCUMENTED -> AUTHORIZED -> EXECUTED -> VERIFIED
```

A candidate is a proposed interpretation or route. Documented means its sources, assumptions, intended action, and uncertainty have been recorded; documentation is not approval. Authorized means an accountable authority has granted permission within a defined scope. Executed means evidence shows that the action occurred. Verified means the observed result satisfies the stated acceptance conditions and remains aligned with that authority.

Consequential actions require explicit authorization from an accountable human authority before execution. A policy may preauthorize low-risk deterministic actions within specified limits, provided the executing system checks those limits and records the policy basis. Deterministic verification of a precondition is not a substitute for authorization. An agent's confidence or risk score cannot expand its permissions.

Each transition should retain the action identifier, relevant evidence, authority identity or policy reference, scope, time, and disposition. Changed targets or expired permission require renewed assessment. A failed or partial execution must remain visibly failed or partial, not become "verified" because a command was issued. Verification should inspect the receiving system's resulting state where possible, rather than rely solely on the initiating agent's account.

The governing principle is:

A route is not an outcome. An outcome is not verified until evidence and authority agree.

Where evidence is missing, authority is unclear, or accounts conflict, the proposed default is to withhold progression and seek review. A successful interface animation, signed message, or completed model response is not sufficient evidence of an authorized outcome.

4. Proposed Risk Indicators and Deviation Signals

Composite Risk Indicator

The original “Veracity Score” increases with strain, origin risk, and contradiction. Calling it veracity would suggest that a larger value means greater truthfulness. It is therefore renamed the proposed composite risk indicator:

```
R = alpha * API + beta * AORM + gamma * PoR
alpha, beta, gamma >= 0
alpha + beta + gamma = 1
API, AORM, PoR in [0, 1]
```

All inputs are dimensionless. The Agent Pressure Index (API) could combine compute-load change, exception and retry rates, memory churn, and optional operator context. Each input requires a documented transformation to a common bounded scale. If API is itself a weighted combination, its weights must likewise be nonnegative and sum to one. Missing or unreliable inputs must be identified rather than silently assigned reassuring values. Biometric context cannot be treated as a validated measure of cognitive fatigue.

The Agent Origin Risk Modifier (AORM) expresses a proposed assessment of origin, dependencies, and granted capabilities. The original examples—0.10 for vetted internal deterministic templates, 0.30 for internal language-model templates, 0.60 for external-API-dependent agents, and 0.85 for experimental agents with critical write access—are illustrative numbers, not calibrated ratings. Internal origin does not imply trust, and permissions require their own assessment.

“Proof of Reality” (PoR) is retained as a label for a contradiction indicator comparing recorded agent claims with inspectable external evidence. It is not proof of reality in a philosophical sense, nor a calibrated probability of contradiction unless tested as such. Absence of observed contradiction is not proof of correctness, particularly where observations are incomplete or stale.

R can inform review or containment under defined policy. It cannot establish truth, grant authority, or compensate for a prohibited action. Weights, normalization rules, evidence coverage, and decision thresholds remain unresolved.

Baseline-Normalized State Deviation

A raw byte difference cannot be compared with dimensionless thresholds. For update size s and baseline mean size b, both measured in bytes, a proposed relative deviation is:

```
D = abs(s - b) / b, when b > 0
D is unavailable when b = 0 or the baseline is missing
```

A zero baseline requires explicit bootstrap handling, not division by zero or silent classification as normal. If a reference scale is introduced, it must be positive, measured in bytes, and documented. This draft omits numeric anomaly cutoffs because none has been calibrated. Baseline windows, task types, structural-change checks, and alert responses require specification; update size alone cannot establish semantic error.

Temporal Whip-Saw Indicator

The proposed temporal indicator penalizes update rates outside an expected interval:

```
T = k_low * max(0, f_min - f)^2
  + k_high * max(0, f - f_max)^2
```

Here f, f_min, and f_max are frequencies in updates per second, measured over a declared observation window, with 0 <= f_min <= f_max. Nonnegative coefficients k_low and k_high have units of seconds squared, making T dimensionless. They are distinct from the composite-score weights. T is not inherently bounded by one and would require normalization before inclusion in that score.

Low rates may warrant examination for disconnection or stale state; high rates may warrant examination for retries or resource pressure. Neither automatically indicates hallucination. Scheduled inactivity and legitimate bursts must be distinguished from faults. Bounds and coefficients are illustrative model parameters, not established safety limits.

5. Visual Protocols and Quarantined Relationships

The proposed visual grammar uses dashed lines for modeled or pending relationships, solid lines for documented records, bright markers for explicitly labeled committed reference states, and halos for uncertainty. These conventions need text labels: a solid line must not imply authorization, and a bright marker must not imply verification. The permission state remains separately visible.

Authority-domain colors retain their intended meanings: Terracotta for field/community sources, Amber for operational records, Copper for commercial pathways, and Violet for governance envelopes. They distinguish domains, not rank or automatic correctness. Conflicting accounts remain inspectable.

The Quarantined Relationship Protocol (QRP) assigns unvalidated lineage hypotheses or unapproved candidate relationships the proposed policy marker `SEALED_LEGAL`, with the directive `do_not_render_in_any_projection`. These strings are not evidence of actual enforcement or legal status. The intended behavior is deny-by-default access and rendering across retrieval, agent context, interfaces, and exports, with an auditable release decision. Effective quarantine requires enforcement at those boundaries; hiding an edge in one display is insufficient.

6. Cryptographic Boundaries and Provenance

CIPHER proposes a five-level key-scope hierarchy:

```
Master -> Tenant -> Module -> Entity -> Field
```

The hierarchy organizes scope; it is not a validated key-derivation algorithm. Any implementation needs context-specific domain separation, appropriate established derivation mechanisms, and defined rotation and revocation. Keys must be bound to their intended purpose and scope rather than reused indiscriminately. The original speculative “fractal key derivation” is not validated or standardized and supplies no security assurance in this draft.

Current standardized names are ML-KEM, specified in NIST FIPS 203 for shared-secret establishment, and ML-DSA, specified in FIPS 204 for digital signatures. [3] [4] Their lineage is CRYSTALS-Kyber and CRYSTALS-Dilithium respectively; standardized specifications are not simply interchangeable names for earlier submissions. [5] A future implementation must follow the applicable specifications and updates.

NIST describes these algorithms as believed resistant to relevant quantum adversaries, not unconditionally secure. [3] [4] Appropriate primitives do not confer blanket quantum safety on an entire system. Security still depends on implementation, protocol composition, and key protection. Cryptographic integrity does not establish factual truth, human consent, or permission to perform the signed action.

Merkle roots may commit to recorded governance events at defined epochs. PostgreSQL storage or blockchain anchoring does not inherently make records immutable, complete, or true. Tamper-evident provenance depends on protected recording, retained evidence, verifiable commitments, and the anchoring system’s operating assumptions. It covers recorded events; it cannot prove that unrecorded actions did not occur or that an inaccurate original record described reality correctly.

Containment consequently remains a permission-and-evidence discipline. Cryptography supports inspection, but neither cryptography nor body feedback can replace accountable authority and verification of the actual outcome.

References

1. NIST SP 800-207, Zero Trust Architecture. August 2020.
2. NIST AI Risk Management Framework. AI RMF 1.0 released January 2023; voluntary risk-management guidance.
3. NIST FIPS 203, Module-Lattice-Based Key-Encapsulation Mechanism Standard. August 2024.
4. NIST FIPS 204, Module-Lattice-Based Digital Signature Standard. August 2024.
5. NIST standards issuance notice, Federal Register document 2024-17956. August 2024; confirms algorithm lineage.

These references support the source-specific statements cited above. They do not endorse or independently validate the proposed architecture.