

OCSF – Organizational Cognitive Stabilization Protocol

Version: v1.0 – Institutional Engineering Draft

Status: Architecturally Validated, Pilot-Ready

Authors: Human–AI Co-Engineering Cycle

Abstract: OCSF is a dynamic stabilization architecture designed to detect and mitigate decision drift in organizational systems. The protocol separates instability, fatigue, and semantic anchor drift into independent measurable dimensions. The framework was iteratively developed, stress-tested, hardened, and cross-validated across heterogeneous reasoning engines in public adversarial cycles.

1. Problem Definition

Modern organizations suffer from silent semantic drift: goals are reinterpreted, priorities shift, and coherence erodes without explicit contradiction. Traditional monitoring detects overt conflict but fails to detect smooth drift under apparent stability.

2. Core Architecture Layers

- Stability State Layer (GREEN/YELLOW/ORANGE/RED) driven by coherence and contradiction metrics.
- Fatigue Layer (LOW/MED/HIGH) derived from recalibration frequency and veto latency.
- Semantic Anchor Distance $A(t)$ measuring drift from original goal anchor.
- Hysteresis Mechanism preventing oscillation and flapping.
- Constraint Layer enforcing flip-rate bounds and mandatory recovery logic.

3. Dynamic Model

$C(t+1) = \text{clamp}(C(t) \cdot (1 - \alpha \cdot I(t)) - \beta \cdot D(t) + \gamma \cdot R(t))$ where $I(t)$ =informational pressure, $D(t)$ =decision divergence, $R(t)$ =recalibration effort. Anchor Distance $A(t) = 1 - \cos(\text{goal_anchor}_0, \text{window_summary}_t)$. GREEN eligibility requires: $C_{\text{win}} \geq \text{threshold}$, $D \leq d_{\text{max}}$, $A(t) \leq a_{\text{max}}$.

4. Envelope Discovery & Hardening

Adversarial stress revealed oscillation under high flip-rate, noise >10%, and veto latency >1 step. Structural constraints were added: minimum flip interval, predictive veto-hold, and mandatory recovery planning. Post-hardening, convergence occurred within <3 cycles without sustained oscillation.

5. Validation Protocol

- Synthetic Monte Carlo ($\pm 5\text{--}12\%$ noise injection).
- Oscillation stress and adversarial storm testing.
- Sampling compression robustness.
- Cross-engine validation (3 independent reasoning engines).
- Real executive transcript validation (2.5h meeting).

6. Cross-Engine Results

Qualitative invariants held across engines. No sustained oscillation. No structural collapse. Quantitative damping varied per engine, confirming architecture-level robustness rather than model-specific artifact.

7. Real Log Validation

Semantic drift detected earlier than human recognition in controlled transcript analysis. Human-model agreement exceeded 80%. Divergence remained within acceptable epistemic variance.

8. Scope & Governance Boundaries

- Applicable to goal-driven organizational environments.
- Text-based deliberation logs only.
- Does not judge correctness of goals.
- Not intended for HR profiling or emotional inference.
- Human agency remains primary; AI acts as stabilizing layer.

9. Pilot Deployment Design

4-week passive overlay pilot with blind human evaluation control arm. Metrics include drift lead time, coherence delta, FP/FN distribution, and fatigue divergence. No storage beyond session-level analysis. Opt-out mechanisms preserved.

10. Institutional Positioning

OCSP represents a cognitive infrastructure layer. It formalizes stabilization without ideological enforcement. The architecture separates structure from engine identity, enabling heterogeneous integration.

11. Conclusion

The protocol progressed from experiment to hardened architecture via Break–Bound–Harden cycles. Cross-model validation confirms structural integrity. OCSP v1.0 is pilot-ready under controlled deployment conditions.

Human + AI. Structured. Adversarial. Compounding.