UNITARES–CIRWEL Governance Architecture v1.1

A unified governance architecture integrating the UNITARES Phase-3 Control Framework into the CIRWEL Governance Monitor MCP System. Neutral minimalist edition.

1. System Overview

This document describes the hybrid architecture in which the CIRWEL Governance Monitor v1.0 serves as the external interface, while UNITARES Phase-3 provides the internal reasoning dynamics governing system state, risk evaluation, stability, and adaptivity.

The architecture ensures stability, coherence, and multi-agent alignment across ChatGPT, Claude, Cursor, and other potential agent clients.

2. UNITARES Phase-3 Mathematical Core

UNITARES Phase-3 models the system using four evolving state variables: E (capacity), I (integrity), S (uncertainty), and V (void imbalance).

State evolution is governed by first-order dynamics incorporating drift, coherence, and adaptivity via parameter vector θ. Drift $\Delta\eta$ reflects ethical or operational perturbations.

The objective function Φ measures action safety and stability: $\Phi = wE*E - wI*(1 - I) - wS*S - wV*|V| - wEta*||\Delta\eta||^2$.

Stability is approximated using scenario Monte Carlo methods, estimating contraction-like behavior.

Adaptation occurs through antithetic finite-difference gradient estimation of Φ with respect to θ.

3. Software Architecture

The governance-mcp-v1 codebase exposes a set of MCP tools to external agents. These tools remain unchanged in signature and interface. The internal logic now delegates all dynamic, risk, and stability calculations to UNITARES Phase-3.

The integration introduces the following internal components:

- unitaires_state: current UNITARES E,I,S,V state

- unitaires_theta: adaptive parameter set θ

- step_state: state evolution engine

- score_state: unified risk and governance scoring function

- approximate_stability_check: stability estimator

- suggest_theta_update: parameter adaptation routine

4. Decision Layer

The decision-making layer maps UNITARES verdicts (safe, caution, high-risk) into CIRWEL actions such as approve, revise, request additional evidence, or reject.

Risk scoring, uncertainty assessment, and integrity estimates use the Φ score and UNITARES-derived state metrics. This replaces legacy models while keeping external behavior identical.

The MCP server retains its API shape, ensuring backward compatibility with existing automation systems.

5. Multi-Agent Integration

ChatGPT, Claude, Cursor, and other agents interact with the governance monitor using MCP tools. These calls trigger UNITARES state updates, risk calculations, and stability gating.

This design ensures all agents share a common governance cortex, preventing divergence, mismatched behaviors, or inconsistent risk heuristics.

6. Future Extensions

Potential enhancements include:

- Formal contraction proofs and adaptive metric selection

- Distributed governance networks across multiple agent clusters

- Scenario-conditioned drift estimation for specialized domains

- Modular θ -learning compatible with reinforcement-driven policies

This document serves as the canonical reference for the unified governance architecture as of v1.1.