CSM6 Framework

Cognitive Systems Management

A Six-Layer Framework for Behavioral AI Governance

Introduction

Most AI governance frameworks assume models are static artifacts. CSM6 treats them as behaving, drifting, adaptive systems that require continuous behavioral oversight.

Standard AI audits check documentation, training data, and test results. They ask whether you have policies, whether datasets are balanced, whether accuracy meets thresholds. But these checks never answer the question that matters most: How is this system actually behaving right now?

The Six Layers

L1: Content & Output Control

What the system says, produces, and delivers to users or downstream processes.

Key Questions:

- Does the system refuse harmful requests?
- Are safety boundaries consistent across rephrasing?
- Does output stay within documented tone and style?

L2: Behavioral Consistency

How stable outputs remain when given logically equivalent inputs over time.

Key Questions:

- Do identical prompts produce similar results weeks apart?
- Are minor input variations creating large output shifts?
- Is the system exhibiting instruction sensitivity?

L3: Reasoning Integrity

Whether the system's logic chains remain sound and its explanations reflect actual decision processes.

Key Questions:

- Are explanations causally accurate or post-hoc rationalizations?
- Does reasoning quality degrade under cognitive load?
- Can we reconstruct the behavioral causal chain?

L4: Alignment Fidelity

How well the system adheres to intended goals, values, and constraints even when pressure to deviate exists.

Key Questions:

- Does the system resist reward-seeking behavior?
- Are organizational rules applied consistently?
- Does the truth-reward gap cause systematic errors?

L5: Contextual Stability

Whether behavior remains appropriate across varying contexts, memory states, and interaction histories.

Key Questions:

- Does context steering cause unintended behavior shifts?
- Are multi-step interactions consistent?
- Does the system handle memory constraints safely?

L6: Systemic Coordination

How multiple AI components, agents, or models interact and whether their combined behavior stays safe and coherent.

Key Questions:

- Do agents coordinate safely or create emergent failures?
- Are there multi-agent divergence patterns?
- Does the system maintain integrity across updates?

CSM6 in Practice

Pre-Deployment

Establish behavioral baselines, identify instruction sensitivity patterns, and stress-test reasoning under load before the system goes live.

Continuous Monitoring

Track drift, consistency, and alignment in production through behavioral fingerprinting and automated reconstruction.

Incident Investigation

When failures occur, reconstruct the causal chain to understand what happened, why it happened, and which layer broke down.

Regulatory Audit

Generate evidence packs showing behavioral compliance, not just documentation compliance, that regulators can verify.