7 Questions for Detecting Self-Correction Failure in Complex Systems

A diagnostic for systems that continue functioning after correction fails

Use: Answer each question based on actual system behavior, not documented policy. Ambiguous or procedural answers indicate elevated drift risk.

1. **What would make this system stop or pause—immediately?**

   Indicator: If nothing can, drift is already present.

2. **Where does error produce direct, unavoidable consequence?**

   Indicator: If consequences are deferred, symbolic, or reputational only, feedback is weak.

3. **Who has veto power and when was it last successfully used?**

   Indicator: Veto authority is defined but not operational.

4. **If this decision is wrong, how reversible is it in practice?**

   Indicator: Reversibility on paper ≠ reversibility in reality.

5. **What metric or proxy is being optimized instead of the actual outcome?**

   Indicator: Evidence of optimization drift.

6. **Where is accountability documented, but responsibility diffused?**

   Indicator: Accountability is recorded in artifacts (logs, reports, sign-offs) rather than held by someone who can act.

7. **What does everyone privately recognize that the system cannot formally acknowledge?**

   Indicator: Narrative lock-in indicator.

Interpretation: If multiple answers rely on documentation, narrative justification, or deferred consequences, the system is likely continuing beyond its self-correction threshold.

Constraint Collapse and Feedback Inversion

How scaled systems lose the ability to invalidate themselves while maintaining operational continuity