

A Structural Explanation for Why Systems Keep Working After Meaning Drops Out

Conceptual Working Paper - Reality Drift Framework

A. Jacobs | 2026

Large symbolic systems often continue functioning long after their connection to reality has weakened. Metrics update, decisions respond to signals, and procedures remain stable, even as orientation degrades and meaningful correction becomes increasingly rare.

This paradox can be explained structurally by distinguishing between the mechanisms that govern symbolic systems and the observable diagnostic patterns that appear when those mechanisms degrade.

The Drift Principle: When Optimization Outpaces Constraint

As symbolic systems scale, representations drift away from reality faster than corrective constraints can bind them.

Scaling accelerates abstraction, compression, and internal optimization, while feedback, accountability, and consequence become slower or symbolic. Meaning, understood as the reliable binding of representation to consequence, therefore weakens over time.

As compression intensifies, maintaining fidelity becomes increasingly costly. Systems increasingly optimize internal representations rather than respond to external reality.

The Structural Mechanisms: Five Operators of Self-Correction Loss in Scaled Symbolic Systems

All large-scale symbolic systems operate through a common set of underlying processes. These operators are not failures themselves, but the structural dynamics that govern how alignment is maintained, degraded, and occasionally restored within scaled symbolic environments.

Drift describes the gradual accumulation of error between representation and reality over time, as feedback is always delayed, partial, or imperfect.

Constraint binds action to consequence through feedback, accountability, and correction, ensuring that errors force adjustment rather than persistence.

Compression reduces complexity through models, abstractions, and summaries, allowing systems to coordinate across vast informational environments.

Representation governs the relationship between symbols and the realities they describe, maintaining directional flow from external conditions to internal models.

Filtering refers to the human capacity to distinguish signal from noise and remain oriented within dense symbolic environments.

Together, these operators form the structural mechanics that enable symbolic systems to function. At the same time, these same mechanisms make it possible for alignment to weaken gradually without immediate breakdown.

The Diagnostic Patterns: The Paradox of Operational Continuity Without Correction

As the structural operators weaken, systems exhibit a set of first-order failure patterns. Each can occur in isolation while the system remains stable, but when they converge, they produce a second-order regime in which self-correction can no longer be sustained.

Delayed or externalized consequence emerges when feedback no longer delivers timely or direct costs, allowing decisions to lose contact with their outcomes.

Compression without fidelity occurs when representations are optimized faster than they are validated, allowing models to substitute for reality.

Legibility outrunning reality describes situations in which indicators and narratives continue enabling coordination even after they stop accurately mapping to external conditions.

Continuation outpacing correction arises when procedural momentum and accumulated trust allow systems to maintain stability despite declining constraint integrity.

These patterns are observable signs that self-correction is weakening.

Constraint Collapse: How Scaled Systems Lose the Ability to Invalidate Themselves

Constraint collapse begins when first-order failures cease to remain isolated and start reinforcing one another, producing a second-order regime in which systems lose the ability

to invalidate their own representations. It marks a systemic phase shift that can be understood from either structural or diagnostic perspectives.

From the structural lens, constraint collapse occurs when the constraint operator loses binding power across interacting processes. Feedback continues to circulate, but no longer forces correction, allowing actions to remain responsive to internal signals while becoming increasingly decoupled from external consequences.

From the diagnostic lens, constraint collapse emerges as first-order failures converge. Delayed consequences, fidelity loss, narrative substitution, and procedural momentum begin to reinforce one another while masking misalignment and preserving the appearance of functionality.

Both perspectives describe the same underlying condition: a stable loss of self-correction in which systems remain operational despite declining alignment with reality. Representations persist not because they remain accurate, but because they remain uninterrupted.

Glossary of Core Terms

Constraint Collapse (Second-Order Failure): A systemic condition in which systems remain responsive and functional in appearance while feedback loses corrective force, allowing misalignment to persist without binding to external reality.

Feedback Inversion: A state in which signals that once produced learning instead reinforce continuation, causing responsiveness to stabilize misalignment rather than reduce it.

Correction Failure: The loss of a system's ability to translate detected error into effective adjustment, even when feedback remains present.

Continuation Bias: The structural tendency of systems to prioritize uninterrupted operation over alignment with changing external conditions.

Survivable Wrongness: A condition in which inaccuracies no longer threaten system stability, allowing incorrect representations to persist without triggering correction.

Representational Failure: The breakdown of reliable correspondence between symbols and the realities they claim to describe.

Loss of Grounding: The weakening of causal connection between representations and the external conditions that once constrained their accuracy.

Confidence Without Accuracy: A condition in which representations remain fluent, coherent, and authoritative despite declining alignment with reality.

Performative Optimization: Optimization directed toward improving internal metrics or appearances rather than external outcomes or realities.

Reality Drift Aphorisms

1. Systems do not fail when they become wrong. They fail when nothing any longer forces them to stop.
2. Feedback does not disappear. It simply stops correcting.
3. Stability is not evidence of alignment. It is often evidence that correction has become impossible.
4. When optimization outpaces constraint, meaning becomes optional.
5. Nothing collapses anymore. Things just keep working while drifting away.
6. A system that cannot stop cannot learn.
7. Fluency is what systems sound like after constraint collapses.
8. The map keeps updating even after the territory disappears.
9. Reality drift begins when continuation becomes easier than correction.
10. Systems remain operational long after they stop being answerable to reality.
11. The absence of failure is not proof of alignment.
12. When correction is expensive, continuation becomes truth.

Framework Thesis

Reality Drift describes how scaled symbolic systems remain operational even as they lose reliable binding to reality. As constraints weaken and optimization accelerates, representations drift faster than feedback can correct them, producing a stable loss of self-correction in which systems continue functioning while their internal models become progressively wrong.