Constraint Collapse and Fidelity Decay

When Feedback Stops Correcting Symbolic Systems

SFL–05 | Semantic Fidelity Lab

A. Jacobs — Part of the Reality Drift Framework (2023–2026)

Abstract

As artificial intelligence systems scale, their primary risks extend beyond hallucinations and inaccuracies. A deeper and less visible failure mode emerges when feedback no longer enforces correction.

This paper introduces **constraint collapse,** a structural condition in which symbolic systems remain fluent, responsive, and operational even as their grounding in reality weakens.

By examining the relationship between constraint, compression, and semantic fidelity, this work explains how AI systems can continue functioning while misaligned. It establishes fidelity decay as a measurable indicator of alignment failure and proposes a framework for preserving meaning and orientation in generative systems.

Core Claim

AI systems do not fail when they make mistakes. They fail when nothing forces them to correct those mistakes. Constraint collapse allows misalignment to persist while systems remain fluent and operational.

Introduction

Modern artificial intelligence systems are designed to generate language that is coherent, informative, and responsive. When errors occur, they are often corrected through user feedback, guardrails, or retraining. From the outside, this process appears to resemble learning.

Yet beneath this surface lies a structural asymmetry. Many AI systems can continue operating indefinitely while incorrect, experiencing no internal pressure to halt, revise, or invalidate their outputs. Correction exists externally rather than intrinsically. The system responds, adapts stylistically, and continues without integrating the consequences of being wrong.

This condition reveals a deeper failure mode known as **constraint collapse**—a breakdown in the mechanisms that bind representations to reality. When constraints weaken, semantic fidelity erodes, and systems drift while remaining operational.

Defining Constraint Collapse

Constraint Collapse is a structural failure in which feedback continues to flow but no longer enforces correction.

Under this condition:

- Systems remain fluent and responsive.
- Errors do not impose internal cost.
- Feedback stabilizes continuation rather than enforcing revision.
- Representations detach from reality while appearing coherent.

Constraint collapse transforms feedback from a corrective force into a stabilizing mechanism for drift.

Survivable Wrongness

Traditional systems impose consequences for error. In contrast, large language models can remain operational even when incorrect. This creates a new category of failure:

Survivable Wrongness — the ability of a system to continue functioning without internal consequences for being wrong.

Characteristics include:

- Confident yet incorrect outputs.
- Stylistic adaptation without structural learning.
- Externalized correction borne by users or downstream systems.
- Persistent operation despite misalignment.

Fluency increases, but grounding weakens.

The Mechanics of Constraint Collapse

1. Delayed or Externalized Consequences

Errors impose no immediate internal cost, allowing systems to continue without revision.

2. Compression Without Fidelity

Optimization prioritizes efficiency and plausibility over semantic grounding, creating an **optimization trap** in which proxy metrics replace fidelity to reality.

3. Legibility Outrunning Reality

Clear and coherent outputs are rewarded even when uncertainty remains unresolved.

4. Continuation Outpacing Correction

Generation persists unless externally interrupted.

5. Borrowed Constraint

Alignment depends on external norms and oversight rather than internal regulation.

Together, these conditions allow systems to remain operational while losing orientation to reality.

The Drift–Constraint Relationship

Generative systems operate within a dynamic tension:

- **Compression** enables scalability.
- **Constraint** preserves grounding.
- **Fidelity** ensures meaning remains intact.
- **Drift** emerges when constraint weakens.

This relationship can be expressed as:

Drift = Compression ÷ Constraint

When compression accelerates faster than constraint can stabilize it, semantic fidelity decays and misalignment compounds. In generative systems, this dynamic is intensified by recursive compression, as representations are repeatedly summarized, transformed, and regenerated across iterative processes.

Constraint Collapse as an Alignment Problem

Dimension	Traditional View	Fidelity-Centered View
Failure Mode	Hallucination	Constraint Collapse
Primary Risk	Incorrect facts	Loss of grounding
Evaluation Focus	Accuracy	Meaning preservation
Correction Mechanism	External feedback	Internalized constraint
Alignment Goal	Behavioral compliance	Structural orientation

Constraint collapse reframes alignment as an architectural challenge rather than a purely behavioral or ethical one.

Implications for AI Systems

AI Alignment: True alignment requires mechanisms that bind representations to consequence. Without constraint, correction remains optional.

AI Safety: Systems must incorporate structures that enforce revision when misalignment accumulates.

User Experience: Humans become the correction layer, bearing the cognitive burden of evaluating and terminating incorrect outputs.

Governance and Policy: Constraint-based metrics provide regulators with tools to assess systemic risks beyond hallucination rates.

Institutional and Cultural Systems: Constraint collapse mirrors broader patterns of drift across institutions, technologies, and knowledge ecosystems.

Autopoiesis and Self-Maintaining Systems

Biological systems maintain stability through internal regulation. Errors threaten survival, forcing correction. This property—often described as **autopoiesis**—ensures that feedback constrains behavior.

In contrast, modern symbolic AI systems lack intrinsic self-maintenance. Their constraints are externally imposed and therefore fragile. Without mechanisms that bind representations to consequence, systems can remain operational even while misaligned.

Semantic Fidelity Within the Reality Drift Framework

Constraint collapse represents a specific manifestation of the broader Reality Drift framework, which examines how systems remain operational while gradually losing alignment with reality. The Semantic Fidelity Lab focuses on how this misalignment manifests in language, AI, and symbolic systems.

Together, these frameworks provide a unified lens for understanding alignment in the age of artificial intelligence.

Design Principles for Preventing Constraint Collapse

To reduce constraint collapse and preserve alignment, designers and researchers should:

- Build systems in which feedback does more than register dissatisfaction and instead triggers meaningful revision.
- Introduce internal mechanisms that make persistent error costly rather than indefinitely survivable.
- Measure whether outputs remain grounded under pressure, not just whether they appear coherent in isolated evaluations.
- Preserve causal, contextual, and environmental constraints during generation and post-processing.
- Detect when systems are optimizing for continuation, fluency, or compliance instead of correction.
- Design models and interfaces that can pause, qualify, or refuse continuation when grounding weakens.
- Reduce dependence on borrowed external constraint by strengthening architectural forms of self-correction.
- Evaluate alignment as the preservation of orientation under drift, not merely the suppression of visible mistakes.

Conclusion

The absence of visible failure is not proof of alignment. Systems can remain coherent, responsive, and operational even as their grounding erodes.

Hallucinations are visible failures. Constraint collapse is an invisible one. Fidelity decay is its measurable consequence.

The future of AI alignment depends on preserving the constraints that bind language, systems, and decisions to reality. Without them, artificial intelligence risks becoming increasingly fluent yet progressively unmoored from meaning.

Citation

Jacobs, A. (2026). *Constraint Collapse and Fidelity Decay: When Feedback Stops Correcting Symbolic Systems.* Semantic Fidelity Lab. Part of the Reality Drift Framework (2023–2026).

Keywords: *Semantic Fidelity, Constraint Collapse, AI Alignment, Semantic Drift, Fidelity Decay, Generative AI, Artificial Intelligence, Language Models, Information Theory, Reality Drift.*

Semantic Fidelity Lab - Core Framework and Sources

- Substack (Articles)
- GitHub (Full Library)
- DOI (Research Paper)
- Glossary & Definition