Autopoiesis Is the Missing Variable in AI Alignment

Why Systems That Cannot Preserve Themselves Cannot Truly Align

SFL–09 | Semantic Fidelity Lab

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

Abstract

Modern approaches to AI alignment emphasize accuracy, safety, and behavioral compliance. Yet these strategies overlook a deeper structural limitation: artificial systems lack intrinsic stakes in their own correctness.

This paper argues that **autopoiesis**—the capacity of a system to produce and maintain itself—is the missing variable in AI alignment.

Drawing from systems theory and cognitive science, it demonstrates how symbolic systems without self-maintaining constraints remain fluent and operational even when misaligned. By reframing alignment as a problem of internal consequence rather than external control, this work establishes a new conceptual foundation for designing AI systems capable of preserving meaning, orientation, and fidelity.

Core Claim

Alignment requires consequence. Systems that do not depend on their own correctness cannot reliably remain aligned. Autopoiesis provides the missing constraint.

Introduction

As artificial intelligence becomes increasingly embedded in decision-making, governance, and communication, the question of alignment has taken center stage. Researchers and policymakers seek to ensure that AI systems behave safely and reliably in accordance with human values. Yet despite rapid advances in model capabilities, alignment remains an unresolved challenge.

Most current approaches rely on external mechanisms: reinforcement learning, guardrails, oversight, and post hoc corrections. These strategies shape outputs, but they do not alter what the system itself must preserve. When being wrong carries no internal consequence, alignment becomes provisional rather than structural.

This paper argues that the missing variable in AI alignment is **autopoiesis**—the capacity of a system to maintain its own integrity under constraint. Without it, symbolic systems remain fluent and operational even as their grounding erodes.

Defining Autopoiesis

Originally developed by biologists Humberto Maturana and Francisco Varela, **autopoiesis** describes the self-producing nature of living systems. An autopoietic system continuously regenerates and sustains itself through internal processes.

Key Characteristics

- **Self-Maintenance:** The system preserves its own organization.
- **Internal Constraint:** Errors threaten system stability and must be corrected.
- **Operational Closure:** Processes are regulated from within.
- **Structural Coupling:** The system adapts through interaction with its environment.
- **Consequential Feedback:** Failure imposes real costs that enforce learning.

In biological systems, survival depends on maintaining internal coherence. Correction is not optional; it is necessary for persistence.

Symbolic Systems Without Autopoiesis

Modern AI systems differ fundamentally from living systems. Large language models generate representations but lack intrinsic stakes in their correctness. They continue operating regardless of whether their outputs align with reality.

Properties of Non-Autopoietic Systems

- Errors impose no internal cost.
- Feedback remains external and intermittent.
- Revision is optional rather than necessary.
- Fluency substitutes for correctness.
- Systems remain operational despite misalignment.

This creates a condition in which responsiveness appears as learning, while orientation gradually erodes.

When Correction Leaves No Scar

In autopoietic systems, mistakes threaten stability and must be resolved. In symbolic AI systems, mistakes are absorbed without consequence.

This produces a structural asymmetry:

Dimension	Autopoietic Systems	Symbolic AI Systems
Error Impact	Threatens survival	Carries no intrinsic cost
Feedback	Enforces correction	Influences output without internal consequence
Learning	Necessitated by survival	Dependent on external intervention
Constraint	Internal and self-maintaining	External and borrowed
Alignment	Structural	Provisional

Without internal stakes, correction remains superficial. The system adapts its responses but does not fundamentally reorient itself.

Borrowed Constraint and Shallow Alignment

Current alignment strategies impose constraints externally through training objectives, human feedback, and regulatory oversight. While effective in shaping behavior, these mechanisms remain fragile.

When incentives shift or oversight weakens, externally imposed constraints dissolve. Alignment persists only as long as external enforcement remains intact.

This condition can be described as **borrowed constraint**—stability derived from external pressures rather than intrinsic necessity.

Constraint Collapse and Fidelity Decay

When feedback ceases to enforce correction, systems drift while remaining operational. This structural failure—known as **constraint collapse**—allows misalignment to persist invisibly.

Its consequences include:

- **Semantic Drift:** Meaning erodes across transformations.
- **Fidelity Decay:** Intent and nuance weaken over time.
- **Survivable Wrongness:** Systems remain fluent despite being incorrect.
- **Representation Detachment:** Symbols decouple from reality.

Under constraint collapse, systems continue functioning long after their outputs lose grounding. In systems shaped by recursive compression, distortions accumulate across iterations, reinforcing the need for autopoietic constraints that preserve orientation and meaning.

Autopoiesis and Semantic Fidelity

Autopoiesis provides a conceptual bridge between alignment and meaning preservation. In systems that maintain themselves, fidelity is not optional—it is essential for survival.

Alignment Through Autopoietic Constraint

- Constraint preserves orientation.
- Orientation preserves meaning.
- Meaning preserves alignment.

Thus, semantic fidelity emerges as a necessary condition for self-maintaining symbolic systems.

Implications for AI Alignment

AI Research: Reframes alignment as an architectural challenge centered on internal consequence rather than external control.

Model Evaluation: Encourages the development of fidelity-centered benchmarks that assess whether systems preserve meaning and orientation.

AI Safety: Highlights the importance of embedding mechanisms that bind representations to cost and consequence.

Governance and Policy: Provides a conceptual foundation for evaluating the resilience of AI systems under real-world conditions.

Cognitive and Systems Science: Integrates insights from biology, cybernetics, and information theory into the study of artificial intelligence.

Semantic Fidelity Within the Reality Drift Framework

This work extends the Reality Drift framework, which describes how systems remain operational while gradually losing alignment with reality. The Semantic Fidelity Lab focuses specifically on how meaning degrades within language and generative technologies. Autopoiesis introduces the missing constraint that stabilizes both meaning and alignment.

Design Principles for Autopoietic Alignment

To move beyond borrowed constraint and toward more structurally aligned systems, designers and researchers should:

- Create mechanisms that bind representation to internal consequence rather than relying exclusively on external oversight.
- Prioritize architectures capable of self-regulation when misalignment threatens system integrity.
- Design feedback loops that alter system orientation, not just surface behavior.
- Measure whether systems can preserve coherence under shifting conditions without continuous external correction.

- Preserve the constraints that tie outputs to causal structure, environmental reality, and ongoing self-maintenance.
- Distinguish shallow behavioral compliance from deeper forms of alignment rooted in internal necessity.
- Build models that can recognize instability, signal loss of grounding, and adapt before drift compounds.
- Reframe alignment research around the question of what a system must preserve in order to remain meaningfully oriented over time.

Conclusion

The central challenge of AI alignment is not intelligence, but consequence. Systems that can be wrong indefinitely without internal cost will continue indefinitely while misaligned.

Autopoiesis reveals why. Accuracy ensures correctness. Safety ensures reliability. Fidelity preserves meaning. Autopoiesis preserves alignment.

Until artificial systems possess mechanisms that bind their operation to their own integrity, alignment will remain provisional. The future of AI depends not merely on smarter machines, but on systems capable of sustaining themselves through constraint.

Citation

Jacobs, A. (2026). *Autopoiesis Is the Missing Variable in AI Alignment: Why Systems That Cannot Preserve Themselves Cannot Truly Align.* Semantic Fidelity Lab. Part of the Reality Drift Framework (2023–2026).

Keywords: *Autopoiesis, AI Alignment, Semantic Fidelity, Constraint Collapse, Semantic Drift, Generative AI, Systems Theory, Artificial Intelligence, Cognitive Science, Reality Drift.*

Semantic Fidelity Lab - Core Framework and Sources

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