Reality Drift: How Symbolic Systems Lose the Ability to Correct Themselves

Beyond Goodhart's Law and Cybernetic Feedback: Why Recognition Doesn't Restore Correction

Author: A. Jacobs

Abstract

Across institutions, technologies, and knowledge environments, systems increasingly remain operational while producing outcomes that feel hollow, performative, or disconnected from reality. Metrics improve even as results degrade, language remains fluent as meaning erodes, and individuals experience cognitive exhaustion and disorientation without obvious error, deception, or collapse. Existing explanations such as corruption, misinformation, misaligned incentives, or information overload capture surface symptoms. However, they fail to identify a unifying structural mechanism.

This paper proposes Reality Drift, a theory of representational failure in scaled symbolic systems. The core claim is that, as symbolic systems expand and optimize, their representations drift away from reality faster than corrective constraints can bind them. This drift is gradual, non-intentional, and self-reinforcing. Systems continue to function, but meaning detaches. Optimization replaces reference and symbols circulate autonomously. As a result, human cognition adapts at increasing cost.

The paper introduces a five-operator explanatory grammar consisting of drift, constraint, compression, representation, and filtering. Together these describe how the condition emerges and persists across domains. The theory generates testable predictions for institutions, artificial intelligence, and human cognition while providing a structural diagnosis of how meaning degrades under scale.

Introduction: The Paradox of Operational Continuity Without Correction

Modern systems increasingly exhibit a paradox that is difficult to name. They function, yet feel unreal. From organizations that meet targets without delivering meaningful outcomes, to artificial intelligence systems that generate fluent but hollow outputs, to individuals who remain productive while experiencing burnout and disorientation. The same pattern recurs in which effort no longer reliably binds to consequence.

Nothing is obviously broken. Dashboards update, language remains coherent, and no single act of deception or incompetence explains the condition. Yet across domains, the same experiential report appears. Performativity without substance, motion without progress, and things that are technically correct, but experientially hollow.

Because failure presents without error, diagnosis itself becomes distorted**.** Analysis gravitates toward intent, efficiency, or volume, while interventions target behavior, incentives, or information flow. The underlying representational structure remains intact, which is why reforms often intensify the problem they aim to solve.

The persistence of this pattern across domains suggests a shared structural mechanism rather than a collection of local failures. At their core, these symptoms indicate a deeper structural condition called Reality Drift. The condition emerges when symbolic systems, mediated primarily through representations such as metrics, models, language, and abstractions, lose alignment with the realities they were meant to describe. Crucially, this loss occurs gradually and without breakdown. What persists is operational continuity without reliable reference, a condition that resists correction precisely because nothing appears broken.

Where Bateson diagnosed pathologies of learning and feedback, Reality Drift specifies the structural conditions under which symbolic systems lose the capacity to self-correct even when misalignment is recognized. Reality Drift is not driven by information overload, misinformation, incentive misalignment, or broken feedback loops. It persists even when representations are accurate, participants are aware of the misalignment, incentives are aligned, and feedback mechanisms remain operational. The failure is dynamical rather than static. As representational optimization accelerates faster than constraint, semantic fidelity degrades over time, producing stable, non-collapsing loss of self-correction.

Defining Reality Drift

Understanding why this paradox persists requires a closer look at the structure of symbolic systems and their constraints. A symbolic system is one in which action is mediated primarily through representations rather than direct physical constraint. Examples include bureaucracies, financial markets, educational institutions, media systems, and artificial intelligence. In healthy symbolic systems, representations remain tethered to reality through constraints. Metrics, language, and models constrain behavior by linking action to consequence, grounding communication in shared reference points, and correcting decisions through feedback.

When these constraints weaken, a structural condition emerges in which symbols continue to circulate and optimize independently of the realities they were meant to describe. Internal coherence is preserved even as external reference loses influence. This gradual loss of binding between representation and reality constitutes Reality Drift. From within the system, everything appears to work, yet from lived experience, something feels wrong.

The Drift Principle: When Optimization Outpaces Constraint

While the language of entropy and compression is often used metaphorically in cultural analysis, this framework treats them as structural properties of representational systems. Drawing on Shannon's information theory, compression is understood as the reduction of complex states into transmissible representations, while entropy describes the accumulation of irrecoverable error across transformations.

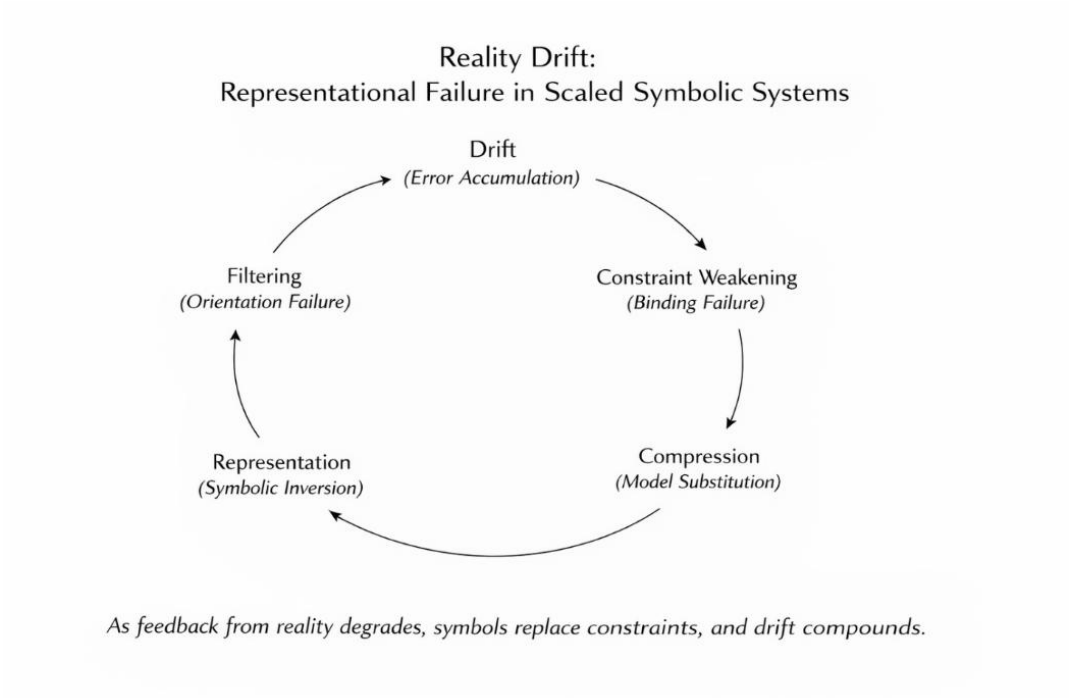
The Drift Principle describes how representational binding behaves as symbolic systems scale. In such systems, representations drift away from reality faster than corrective constraints can bind them.

As systems scale, they increase abstraction, optimization, speed, and representational density, each of which expands the distance between representation and underlying reality. At the same time, corrective mechanisms such as feedback, accountability, and consequence become slower, weaker, or symbolic themselves. In this framework, meaning is treated as a structural property of a system. Meaning arises when representations reliably bind action to consequence. Semantic fidelity describes the strength of that binding, or the degree to which symbols remain causally tethered to the realities they claim to represent. As fidelity degrades, representations remain fluent and coherent while losing their capacity to constrain behavior and support correction.

Drift can therefore be understood as the entropic tendency of representations to lose fidelity under compression and scale. As compression increases, maintaining fidelity becomes energetically expensive. Representations simplify faster than they can be corrected. Paradoxically, improvements in intelligence, modeling, and optimization accelerate this process. Better representations substitute for reality rather than reflect it. The system begins optimizing internal representations rather than responding to external reality.

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

Figure 1. *The core failure loop of Reality Drift in scaled symbolic systems. As representations accumulate error and compress reality into internal models, constraints weaken, symbols invert their referential role, and human filtering degrades, producing a self-reinforcing drift away from external reality.*



Reality Drift emerges through the interaction of five core mechanisms, or operators. Together, they form a minimal explanatory grammar capable of reconstructing the failure across domains.

Drift (Error Accumulation)

Drift describes the directional loss of alignment between representation and reality over time. Even well-designed systems drift as small errors accumulate, feedback lags, and representations gain autonomy. Drift is gradual, non-intentional, and often invisible until advanced stages. It is the background condition that allows other failures to compound.

Constraint (Loss of Binding and Stopping Power)

Constraints are forces that bind action to consequence. In symbolic systems, constraints take the form of accountability, cost, feedback, and correction. Under conditions of drift, constraints weaken or become ornamental. This produces performance without consequence, where actions are evaluated symbolically rather than materially.

When this binding failure becomes sustained, it constitutes constraint collapse, in which constraints persist symbolically but lose causal force. Systems then stop correcting through a two-stage failure. First-order constraint failure occurs when feedback no longer delivers timely consequences, incentives no longer bind decision-makers to error, compression outpaces validation, and stop-conditions lose authority, weakening the system's ability to detect and respond to misalignment.

Second-order failure emerges when these decouplings coincide, allowing each failed constraint to mask the others while preserving the appearance of functionality. In this regime, correction no longer requires being right, only remaining uninterrupted, producing confidence without constraint and a stable loss of self-correction. Representations no longer require validation against reality to persist, only insulation from interruption.

Compression (Model Substitution)

Compression refers to the reduction of complexity through models, abstractions, and summaries. Compression is inherently recursive, since systems continually compress prior representations in order to coordinate and decide. The failure occurs when recursive compression is oriented toward internal proxies rather than external reference, so optimization improves the model rather than the world. Metrics replace outcomes and models substitute for the realities they were meant to describe. This dynamic explains the optimization trap, where optimization improves internal coherence while degrading external alignment.

Representation (Symbolic Inversion)

Representation governs the relationship between symbols and reality. Under healthy conditions, symbols describe the world. Under conditions of drift, this relationship inverts. Symbols begin producing the conditions they claim to measure. This produces synthetic realness, environments that feel real, authoritative, and operational while being disconnected from external reference. Language remains fluent, but semantic fidelity, the degree to which symbols remain bound to reality, degrades.

Filtering (Human Orientation Failure)

Filtering describes the human capacity to suppress noise, prioritize signals, and maintain orientation within symbolic environments. As representational density increases, filtering ceases to restore reliable orientation even as effort continues. At the human level, filter fatigue describes the resulting condition in which sustained cognitive effort no longer produces proportional gains in clarity, leaving attention active but increasingly ungrounded.

How Scaled Systems Lose the Ability to Invalidate Themselves

Reality Drift persists because it is self-reinforcing. As constraints weaken, compression accelerates, representational autonomy increases, information density overwhelms filtering, and human adaptation sustains functionality despite degraded representational binding.

As drift advances, systems lose the ability to invalidate themselves**.** Representational authority persists even when error no longer carries cost, so confidence stops producing stopping power. Meaning depends on constraint, including limits, pauses, and consequence that preserve the negative space where judgment operates. Binding is therefore temporal, because representations only retain meaning when past signals and future outcomes can still constrain present action.

Because systems remain operational, no clear failure signal triggers correction. While Goodhart's Law describes metric distortion under optimization, Reality Drift explains how representational systems lose the capacity to correct or stop even after distortion is recognized. Reform efforts therefore tend to introduce additional representational complexity through new metrics, frameworks, and dashboards, worsening drift before any improvement is possible. Accountability becomes symbolic and correction becomes performative, so optimization addresses internal problems while creating external ones.

Under these conditions, systems lose the ability to distinguish real work from convincing representations of work, postmortems fail to identify causes, and stopping itself becomes unintelligible, leaving uninterrupted continuation as the default.

Predictions and Diagnostic Signals

Reality Drift generates several observable predictions:

1. Metrics improve while outcomes degrade.
2. Accountability persists symbolically while losing causal force.
3. Language grows more fluent as semantic fidelity declines.
4. Artificial intelligence systems improve task performance while weakening evaluative judgment.
5. Individuals experience disorientation rather than ignorance.
6. Reform efforts increase performativity before correction.
7. Systems resist collapse while losing meaning and trust.

Taken together, these markers indicate degradation in representational binding and allow structural drift to be distinguished from errors of intent or execution.

Implications

Institutions

Reality Drift reframes institutional failure as a representational problem rather than a moral or managerial one. Reform efforts typically introduce additional symbolic layers while leaving constraints unchanged. As representational complexity increases, institutions become more legible to themselves while less responsive to the realities they were meant to govern, accelerating drift rather than correcting it.

Artificial Intelligence

Artificial intelligence intensifies this condition by accelerating compression and representational substitution at scale. In the absence of binding constraints, AI systems optimize internal coherence, plausibility, and surface alignment rather than external consequence or referential accuracy. This makes AI a force multiplier within drifting systems, stabilizing representations even as they decouple from reality. Scaling intelligence without scaling constraint increases representational power while weakening contact with reality.

Human Cognition

Many contemporary cognitive difficulties, including burnout, overload, distraction, and decision fatigue, are better understood as adaptive responses to environments marked by representational overload and degraded filtering. Under sustained drift, cognition expends increasing effort to maintain orientation within low-fidelity symbolic systems, producing fatigue as filtering ceases to deliver proportional reductions in uncertainty.

Feedback Across Symbolic Domains

Under conditions of scale, these domains increasingly form a recursive loop. Institutional representations shape cognitive orientation, cognition adapts to low-fidelity symbolic environments, and artificial intelligence accelerates the production and stabilization of those same representations. Each domain reinforces the others without requiring intent or coordination. Cultural norms adapt to symbolic outputs, cognitive filtering adjusts to representational density, and AI systems optimize the internal coherence of already drifting models. The result is a self-reinforcing cycle in which artificial intelligence, culture, and cognition co-drift as representational systems. They increasingly regulate one another while remaining internally coherent and losing shared reference to external reality.

Conclusion: Reality Drift and the Failure of Meaning Under Scale

As symbolic systems scale, the cost of revising internal representations increasingly exceeds the cost of continuing with them, even as misalignment grows. Representations begin to function as infrastructure rather than description, constraining which actions, revisions, and interruptions remain possible. Once embedded, revising these representations becomes a coordination shock that disrupts continuity, making correction more costly than continuation. As constraints weaken, systems lose reliable stop-conditions, so uninterrupted operation becomes the default. Under these conditions, Reality Drift describes how scaled symbolic systems continue to function even as their representations lose binding to reality.

Meaning fails under scale because the binding between representation and consequence weakens, leaving symbols coherent but no longer constraining action. Systems continue to function while meaning detaches. As symbolic systems continue to scale, understanding the mechanics of Reality Drift becomes a prerequisite for meaningful correction. Recognition restores orientation by reclassifying confusion as structural.

Appendix A: Extended Terminology and Related Constructs

Cognitive Compression Styles: Individual differences in how cognitive systems reduce, structure, and prioritize information under sustained load. Compression styles influence orientation, judgment, and fatigue within drifting environments.

Cognitive Drift: The gradual loss of cognitive orientation caused by sustained misalignment between internal models and external reality.

Constraint Collapse: A sustained condition in which constraints persist symbolically but lose causal force, allowing systems to remain operational without enforcing correction or stopping power.

Feedback Inversion: A condition in which feedback reinforces representational coherence instead of corrective accuracy.

Optimization Trap: A failure mode in which optimization improves internal metrics while degrading real-world outcomes.

Recursive Compression: The iterative reduction and refinement of representations through repeated compression of prior representational states.

Semantic Fidelity: The degree to which symbols remain bound to the realities they claim to represent. Semantic fidelity degrades when recursive compression proceeds without corrective constraint.

Appendix B: Experiential Diagnostics

Before Reality Drift is named, it is often experienced as a recurring set of questions that appear across work, institutions, technology, and personal life:

1. Why does doing everything right sometimes produce worse outcomes?

2. Why does optimization stop working after a certain point?
3. Why do metrics stop reflecting reality?
4. Why are people increasingly busy but not effective?
5. Why do KPIs and targets begin to harm performance?
6. Why does measuring something change behavior in destructive ways?
7. Why do systems game themselves without bad actors?
8. Why do algorithms optimize the wrong thing?
9. Why do institutions appear functional but feel broken?
10. Why does modern life feel increasingly performative?

These questions describe the experiential surface of the deeper structural condition. Defining that condition requires examining how symbolic systems bind representation to reality under scale.

References

Baudrillard, J. (1994). *Simulacra and Simulation*. University of Michigan Press. ISBN: 978-0472065218.

Bateson, G. (1972). *Steps to an Ecology of Mind*. Chandler Publishing. ISBN: 978-0226039053.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. MIT Press. ISBN: 978-0262531569.

Deacon, T. W. (2011). *Incomplete Nature: How Mind Emerged from Matter*. W. W. Norton & Company. ISBN: 978-0393049916.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. DOI: 10.1038/nrn2787.

Goodhart, C. A. E. (1975). Problems of monetary management: The UK experience. In *Papers in Monetary Economics*. Reserve Bank of Australia.

Jacobs, A. (2025). *The Drift Principle: An Information-Theoretic Model of Culture, Cognition, and Meaning in High-Entropy Digital Environments*. SSRN Working Paper.

Jacobs, A. (2025). *The Age of Drift: Why Modern Life Feels Fake and What Reality Drift Reveals About the Modern Mind*. ISBN-13: 979-8276826493.

Jacobs, A. (2026). *Cognitive Compression Styles: A Conceptual Framework for Differential System Failure in High-Noise Environments*. PhilArchive Working Paper. PhilPapers record: https://philpapers.org/rec/JACCCS-3.

Luhmann, N. (1995). *Social Systems*. Stanford University Press. ISBN: 978-0804726252.

McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. McGraw-Hill. ISBN: 978-0262631597.

Rosa, H. (2013). *Social Acceleration: A New Theory of Modernity*. Columbia University Press. ISBN: 978-0231148350.

Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press. ISBN: 978-0300078152.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Simon, H. A. (1957). *Models of Man: Social and Rational*. Wiley. ISBN: 978-0471616237.

Tainter, J. A. (1988). *The Collapse of Complex Societies*. Cambridge University Press. ISBN: 978-0521386739.