

Goodhart's Law, Campbell's Law, and Reality Drift

How Metric Corruption Becomes a Larger Failure of Representational Correction

A. Jacobs | Reality Drift Framework

Introduction

Goodhart's Law and Campbell's Law explain why measures become unreliable when consequential systems organize behavior around them. Reality Drift begins from the same problem but follows it farther. It asks what happens when a degraded measure remains operational, travels through larger systems, shapes decisions, reorganizes behavior, and loses the ability to be corrected by the reality it was created to represent.

Reality Drift is the gradual loss of alignment between a system's internal representations and the reality those representations were created to describe, govern, or support. A drifting system may remain operational, coherent, and apparently successful even as its capacity for reality-based correction weakens. The defining paradox is operational continuity without reliable correction.

The Shared Problem

Large systems cannot encounter every relevant condition directly. They depend on metrics, categories, models, procedures, reports, and narratives. These representations make complexity manageable. They also create a practical substitute for the condition itself.

A measure becomes dangerous when it acquires causal authority. Funding, status, access, discipline, and organizational attention begin to depend on it. People learn what the measure notices, what it ignores, and how to produce favorable movement within it. The measure no longer merely observes behavior. It becomes part of the environment to which behavior adapts.

Goodhart's Law

Goodhart's Law is commonly summarized as the principle that when a measure becomes a target, it ceases to be a good measure. The familiar formulation captures a basic change in the relationship between an indicator and the condition it represents. Under optimization pressure, behavior concentrates on the measurable feature. The association that made the measure useful can weaken because the system is now actively selecting for the measure itself.

Goodhart's Law identifies the degradation of a measure under target pressure. It does not by itself describe the full institutional architecture through which that measure may continue to circulate after its validity has weakened.

Campbell's Law

Campbell's Law describes the corruption pressure placed on quantitative social indicators when they are used for consequential decision-making. A test score tied to funding, a policing statistic tied to evaluation, or a clinical measure tied to reimbursement becomes more than an observation. It becomes an object around which institutions organize conduct.

Campbell's Law is especially important because it connects measurement to institutional distortion. The indicator can alter the process it was introduced to monitor. Reported performance may improve while the underlying condition becomes harder to assess.

The Reality Drift Extension

Reality Drift treats Goodhart and Campbell effects as important but narrower forms of a larger representational failure. A system can drift through metrics, but also through categories, procedures, models, interfaces, specifications, narratives, and generated representations. No explicit quantitative target is required.

Goodhart's Law describes what happens when measures become targets. Reality Drift describes what happens when representations of any kind begin displacing the realities they were created to represent.

Reality → Representation → Operational Proxy → Optimization → Constraint Weakening → Feedback Substitution → Fidelity Loss → Reality Drift

From Measure Corruption to System Persistence

The critical transition occurs when the degraded representation remains useful to the system. The metric still allocates resources. The dashboard still organizes meetings. The report still satisfies oversight. The procedure still produces formal completion. Because operation continues, the weakened relationship between representation and reality does not automatically force correction.

This produces what Reality Drift calls Survivable Wrongness. The representation is wrong or misleading in ways the system can absorb. Consequences may be delayed, distributed, transferred to less visible populations, or excluded from the measurement environment. The absence of failure is not proof of alignment.

The Drift Principle

The Drift Principle explains why these failures can compound over time. Representations can be compressed, optimized, distributed, and reused faster than reality can correct them. A metric can move through dashboards, reports, budgets, benchmarks, public narratives, and automated decisions before its relationship to the original condition is reexamined.

Speed favors the representation. Correction usually requires renewed observation, contextual judgment, contradiction, and consequences strong enough to force revision. When representational processing accelerates while corrective capacity remains slow or weak, distance accumulates.

Recursive Confirmation

The deepest failure occurs when the representation helps produce the conditions later treated as evidence for it. People adapt to the metric. Institutions allocate resources according to it. Categories determine eligibility. The resulting environment is measured again through the same categories and indicators.

Reality → Measure → Target → Behavioral Adaptation → Altered Reality → Recapture → Apparent Confirmation

At this point, the system may interpret an environment it has reorganized as independent confirmation of its own model. The measure has not merely become unreliable. It has become part of a reality-producing loop.

Conceptual Boundaries

- **Proxy Optimization** improves a measurable stand-in rather than the underlying condition.
- **The Optimization Trap** begins when further improvement of the proxy progressively worsens alignment.
- **Feedback Weakening** reduces the influence of corrective evidence.
- **Constraint Collapse** occurs when feedback remains present but can no longer compel revision.
- **Semantic Fidelity** concerns what meaning and relational structure survive transformation.
- **Survivable Wrongness** explains why misalignment can persist without stopping operation.

Conclusion

Goodhart explains the degradation of a measure under optimization pressure. Campbell explains the corruption of social indicators under consequential use. Reality Drift explains how degraded representations remain operational, circulate through larger systems, reorganize behavior, weaken correction, and eventually help produce the reality later used to validate them.

The measure does not merely become a bad measure. It can become an organizing layer of the system. Coherence is not contact, and continued performance inside the measurement environment is not evidence that the system remains aligned with the world beyond it.