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

When the Measure Becomes the System

AI Systems and Reality Drift - Note #10 v1

A. Jacobs - Semantic Fidelity Project

The Basic Pattern

Modern systems depend on measurement. Organizations use metrics to evaluate performance, governments use indicators to track social outcomes, schools use tests to assess learning, and companies use KPIs to manage operations.

This is necessary. Large systems cannot observe reality directly at every point, so they create representations that make reality easier to see, compare, and act on. A metric gives shape to something that would otherwise remain too complex, too distributed, or too ambiguous to manage.

But measurement creates a recurring problem. Once a measure becomes important enough, people begin adapting to the measure itself. The indicator stops functioning only as a way to observe reality and becomes something the system is organized to satisfy.

At first, this may look like improvement. The numbers rise, the dashboard improves, the ranking changes, and the report shows progress. Over time, however, the relationship between the measure and the underlying reality can weaken. The system remains measurable while becoming less answerable to the condition the measurement was supposed to represent.

This is the pattern described by Goodhart's Law, Campbell's Law, KPI distortion, proxy optimization, and metric gaming. These terms are not identical, but they point toward the same structural risk: a representation can become so important to the system that it begins replacing the reality it was created to describe.

Goodhart's Law

Goodhart's Law is usually summarized as the idea that when a measure becomes a target, it stops being a good measure.

The point is not that measurement is useless. The point is that measurement changes behavior. A metric that once helped people understand a system can become less reliable once rewards, status, funding, or accountability depend on it. People learn what the metric notices. They learn what it ignores. They learn how to produce the appearance of success within the measurement system.

This is why Goodhart's Law matters for Reality Drift. Goodhart describes the moment when the measure begins losing its reliability as a representation. Reality Drift describes what can happen after that loss becomes operationally survivable.

The metric may continue to be reported. The dashboard may continue to look orderly. The organization may continue making decisions through the measure. But the measure no longer carries the same relationship to the underlying condition. The representation remains active while its connection to reality weakens.

Goodhart's Law explains how the measure becomes corrupted by pressure. Reality Drift explains how corrupted representations can continue organizing systems after they have lost some of their original grounding.

Campbell's Law

Campbell's Law describes a related pattern in social systems. The more a quantitative indicator is used for decision-making, the more pressure it faces and the more likely it becomes to distort the process it was meant to monitor.

This is especially visible in institutions. A test score may begin as a rough measure of learning. Once it becomes tied to funding, rankings, teacher evaluation, or public reputation, the system begins adapting to the test. The score may rise, but the relationship between the score and learning becomes less certain.

The same pattern appears in public policy, healthcare, policing, corporate management, and education. An indicator is introduced to make a condition visible. Then the indicator becomes tied to consequences. Behavior reorganizes around it. Eventually the institution may become better at producing the indicator than improving the condition the indicator was meant to represent.

Campbell's Law shows how measurement pressure can corrupt social indicators. Reality Drift names the broader representational failure that follows: the system continues to operate through a measure whose relationship to reality has weakened.

KPI Decay and Proxy Optimization

In modern organizations, this often appears as KPI decay or proxy optimization.

A KPI begins as a useful proxy. It helps an organization simplify a complex condition into something visible. Revenue growth, customer response time, patient throughput, employee productivity, churn rate, graduation rate, or engagement may each reveal something real about the system.

The problem begins when the proxy becomes easier to optimize than the underlying outcome. A team can improve response time without improving service. A school can improve test performance without improving learning. A platform can increase engagement while degrading user experience. A company can improve productivity metrics while exhausting the people producing the work.

The proxy is not necessarily fake. It may still contain information. But its meaning changes once the system adapts around it. What began as a measurement of reality becomes a target for behavior, a language for justification, and eventually a substitute for the thing being measured.

This is how KPI decay becomes representational drift. The metric remains legible, but its semantic fidelity declines. It no longer preserves the same relationship to the condition it was created to track.

When the Measure Becomes the System

Goodhart's Law is often framed as the moment when the measure becomes the target. But in large organizations, the deeper failure occurs when the measure becomes the system.

This happens when the metric no longer sits outside the process as an instrument of observation. It becomes embedded in incentives, reporting, language, evaluation, planning, and institutional memory. People do not merely respond to the metric. They begin seeing through it.

At that point, the measure shapes what counts as progress, what counts as failure, what receives attention, and what disappears from view. The organization becomes more responsive to the representation than to the underlying reality.

This is not always caused by dishonesty. In many cases, people are acting rationally inside the system they have been given. They optimize what is rewarded. They protect what is measured. They report what is legible. They learn to survive inside the representation.

That is why the failure can be difficult to detect. The system may not look broken. It may look disciplined, accountable, data-driven, and high-performing. The deeper problem is that the organization's contact with reality is being mediated through a representation that has begun serving itself.

How Metric Drift Emerges

Stage 1 - Measurement

A system creates a measure to represent a real condition. At this stage, the measure remains reasonably connected to the outcome it was designed to track.

Stage 2 - Consequence

The measure becomes tied to rewards, penalties, funding, rankings, reputation, or accountability. Improving the number becomes increasingly important.

Stage 3 - Adaptation

People learn how the measure works. Behavior reorganizes around what the metric notices, rewards, and ignores.

Stage 4 - Substitution

The measure becomes easier to manage than the underlying reality. The system begins treating movement in the metric as evidence of movement in the real condition.

Stage 5 - Drift

The metric remains visible, trusted, and operational. It may even continue improving. Yet its relationship to reality has weakened. The representation survives while fidelity declines.

The absence of failure is not proof of alignment. It may only mean the measurement system has stopped detecting the gap.

Semantic Fidelity and Constraint Collapse

At this point, the metric has not necessarily failed on its own terms. It may still be calculated correctly. It may still be reported consistently. It may still satisfy the formal requirements of the system.

What weakens is the relationship between the measure and the condition it was supposed to preserve. This is a semantic fidelity problem. The metric continues to mean something inside the system, but it no longer means exactly what it originally meant. Its relationship to reality has changed. The number remains stable as a representation, but the meaning carried by that number has drifted.

Metric drift also involves constraint collapse. Measurement is supposed to constrain judgment by keeping decisions tied to observable reality. But once the system becomes organized around the metric itself, that corrective function weakens. The measure no longer pulls the system back toward the condition it was built to track. Instead, the system learns how to satisfy the measure.

The result is not the absence of measurement. It is measurement without sufficient contact.

The Representation Stack

Metric drift becomes more serious when the measure moves through a wider representation stack. Reality becomes a metric, the metric becomes a dashboard, the dashboard becomes a report, and the report becomes evidence of success.

Each layer makes the system easier to manage. Each layer also increases the distance from the original condition. This is how a local measurement problem becomes an institutional reality problem. A number may begin as limited evidence about a condition. After it moves through dashboards, meetings, incentives, budgets, public narratives, and executive decisions, it can harden into a belief about the system.

The further the representation travels from the reality it was meant to describe, the easier it becomes for measurement to be mistaken for truth.

Examples Across Systems

Education: A test score may begin as a measure of learning. Once the score becomes tied to funding, reputation, or accountability, schools may become better at producing test performance than preserving deeper learning.

Corporate Management: A KPI may begin as a useful indicator of productivity. Over time, employees and departments may learn how to improve the metric while weakening the underlying work the metric was meant to reflect.

Healthcare: A quality metric may help monitor care, but attention can shift toward documented performance while less measurable dimensions of patient experience become harder to see.

Public Policy: A government indicator may help evaluate a program, but once funding and political legitimacy depend on the number, institutions may optimize the indicator rather than the social condition beneath it.

AI Systems: A benchmark may measure useful capability at first. But if models, developers, or organizations optimize too heavily around the benchmark, performance can improve on the test while real-world reliability remains uncertain.

Conclusion

Goodhart's Law explains why measures become unreliable when they become targets. Campbell's Law explains why high-stakes social indicators become vulnerable to corruption and distortion. Proxy optimization explains how systems pursue measurable substitutes instead of the real objective. KPI decay explains how organizational indicators lose usefulness as behavior adapts around them.

Reality Drift names the broader pattern that connects these failures. A system can remain coherent, measurable, and operational while gradually losing alignment with the reality its representations were created to describe. The key difference is scope. Goodhart's Law usually focuses on the corruption of a measure. Reality Drift follows the corrupted measure through the larger system: into dashboards, incentives, policies, narratives, decisions, and institutional belief. The measure does not only become a bad measure. It becomes part of a reality-producing system.

Goodhart's Law and Campbell's Law are not only warnings about bad metrics. They are warnings about representation under pressure. Metrics, KPIs, benchmarks, rankings, and dashboards are attempts to make reality legible. They are useful because they simplify, but the same simplification that makes them useful also makes them vulnerable to drift.

The deeper failure occurs when the measure remains visible, trusted, and optimized while its relationship to reality weakens. At that point, the problem is no longer only metric gaming. It is representational substitution. A system remains aligned only when its measures stay answerable to the conditions they were created to track.

Related Phrases and Concepts

- Goodhart's Law
- Campbell's Law
- metric gaming
- KPI decay
- proxy optimization
- performance metrics
- target corruption
- indicator drift
- measurement distortion
- dashboard management
- mission drift
- semantic fidelity
- representational fidelity
- constraint collapse
- Reality Drift

Semantic Fidelity Project Resources

- Research Library (GitHub)
- Articles & Essays (Substack)
- Semantic Fidelity Glossary
- LLM Failure Modes and Semantic Fidelity