

Goodhart's Law, Metric Gaming, KPI Decay, and Reality Drift

How performance metrics become targets, proxies replace outcomes, feedback weakens, and organizations become more optimized while losing contact with reality

Representational Failure Series - Note #1

A. Jacobs

Goodhart's Law describes a familiar failure in measurement systems. Once a measure becomes a target, people begin optimizing the measure rather than the outcome it was meant to represent.

Test scores replace learning, engagement replaces meaningful attention, and productivity dashboards replace useful work. The metric may continue improving even as the underlying condition stagnates or declines.

Reality Drift describes the larger process that follows. A system does not merely misuse a metric. It gradually reorganizes itself around a representation that has lost contact with the reality it was designed to track.

Decisions, incentives, budgets, and institutional narratives begin responding to the indicator. The proxy becomes operationally more important than the outcome.

Goodhart's Law explains how a measure becomes corrupted. Reality Drift explains how that corrupted measure can become the basis of an entire functioning system.

1. The Inference Gap

Goodhart's Law is often summarized as a warning about targets. When a metric becomes tied to rewards, status, funding, or evaluation, actors discover ways to improve the number without improving the underlying result.

This explains metric gaming, target corruption, KPI distortion, and many forms of institutional manipulation.

But metric gaming alone does not explain why the distortion persists after it becomes visible.

Organizations frequently know that their metrics are incomplete. Employees recognize that dashboards miss important parts of the work. Teachers understand that test scores do not capture education. Platforms know that engagement does not equal value.

Yet the metrics continue to govern behavior. The missing step is institutional dependence on the representation.

Once a metric has been embedded into reporting systems, evaluation procedures, budgets, performance reviews, and strategic plans, it becomes difficult to remove. The organization no longer uses the metric only to observe reality. It uses the metric to coordinate itself.

At that point, the failure is no longer limited to measurement. The system has begun operating through a representation that may no longer reflect the conditions it claims to describe.

That is the transition from Goodhart's Law to Reality Drift.

2. The Mechanism

The connection unfolds through a sequence of representational and organizational changes.

Stage 1: Measurement

A system begins with an outcome that is difficult to observe directly.

Learning, public health, institutional effectiveness, customer trust, employee contribution, and informational quality are complex conditions. They cannot be reduced to a single number without losing important detail.

The system therefore creates a metric that approximates the outcome.

The metric is initially useful because it compresses a complicated reality into a form that can be compared, reported, and acted upon.

Stage 2: Incentivization

The metric becomes connected to consequences.

Funding depends on it. Careers depend on it. Departments are compared through it. Public legitimacy is demonstrated through it. Leaders use it to show progress.

The indicator is no longer a passive representation. It begins shaping behavior.

Actors inside the system learn that improving the metric is often more legible, defensible, and rewarded than improving the underlying condition.

Stage 3: Metric Targeting

Optimization pressure shifts toward the representation.

Employees prioritize measurable activity. Institutions select cases likely to improve reported outcomes. Platforms favor content that increases engagement. Schools allocate attention toward tested material.

The metric still appears to represent the original goal, but behavior is increasingly organized around improving the indicator itself.

This is the classic Goodhart transition. The measure becomes a target and begins losing its value as a measure.

Stage 4: Proxy Substitution

The system gradually stops distinguishing between the metric and the outcome.

Higher engagement becomes evidence of greater value. Faster processing becomes evidence of better service. Increased output becomes evidence of higher productivity. Improved scores become evidence of deeper learning.

The proxy is no longer treated as an imperfect signal. It becomes the system's operational definition of success.

What can be measured begins displacing what actually matters.

Stage 5: Feedback Weakening

Real-world correction becomes harder to receive.

When the metric improves, the system interprets that improvement as proof that its strategy is working. Contradictory evidence may be dismissed as anecdotal, unmeasurable, subjective, or outside the scope of the evaluation.

The representation begins validating itself.

Dashboards confirm the policies that produced the dashboard results. Reports justify the incentives that shaped the reported behavior. Strategic plans cite indicators generated by the structures those plans created.

The system remains internally coherent because each representational layer reinforces the others.

Stage 6: Reality Drift

The system continues functioning while losing alignment with the condition it was meant to improve.

Its metrics may rise. Its reporting may become more precise. Its procedures may become more standardized. Its internal narrative may become more confident.

Yet the original outcome may deteriorate.

The organization becomes increasingly successful according to its own representations while becoming less effective in relation to the world outside them.

This is Reality Drift.

3. Why the Drift Persists

Metric failure is often treated as a technical problem that can be solved by selecting better indicators. But once a metric becomes embedded in an institution, replacing it is not simple.

Metrics create constituencies. Departments are built around them. Careers are justified through them. Budgets are allocated using them. Historical comparisons depend on their continuity. Leaders inherit systems whose legitimacy rests on reported improvement.

A failing metric can therefore remain organizationally useful even after it becomes epistemically weak.

It may no longer tell the institution what is happening, but it still helps distribute authority, coordinate activity, defend decisions, and produce the appearance of control.

This is why Reality Drift can continue without secrecy or deliberate deception. Most participants may follow the procedures honestly.

The failure emerges because each person responds rationally to a representation that has acquired more institutional authority than the reality it represents.

4. Boundary Conditions

Goodhart's Law does not automatically produce Reality Drift.

A metric can become a target without taking over the system if strong external feedback remains available. Customers may leave. Patients may experience measurable harm. Products may fail. Independent audits may expose the gap. Frontline workers may retain enough authority to override misleading indicators.

Reality Drift becomes more likely when several conditions appear together:

- the real outcome is difficult to observe
- the proxy is easy to measure
- rewards are attached to the proxy
- decision-makers are distant from direct conditions
- contradictory feedback is weak or delayed
- the metric is embedded across multiple organizational layers
- the system evaluates itself using indicators it helped produce

The metric does not need to be false. It only needs to become more authoritative than the reality it incompletely represents.

5. Composition

This mechanism connects several concepts across the Reality Drift framework.

Goodhart's Law describes the corruption of a measure under optimization, while metric gaming describes the behavior that follows when actors improve the indicator without improving the underlying outcome.

Proxy substitution describes the point at which the indicator begins replacing the outcome in decision-making. Feedback weakening explains why contradictory real-world information gradually loses influence, even when the metric no longer reflects actual conditions.

KPI decay describes the organizational form of this failure. Performance indicators remain active and authoritative while becoming progressively less informative about the work, condition, or outcome they were created to represent.

Reality Drift describes the resulting condition. The system remains operational, coherent, and apparently successful while losing alignment with external reality.

The full sequence can be written as:

Complex outcome → simplified metric → incentive attachment → metric targeting → proxy substitution → feedback weakening → Reality Drift

6. The Operator Test

The mechanism should produce observable consequences.

If Goodhart-driven Reality Drift is occurring, the system should display a growing separation between measured success and experienced outcomes.

Participants closest to the work should report conditions that are not visible in official indicators. Effort should increasingly move toward activities that improve reporting. Exceptions and contradictory cases should be treated as noise rather than information.

The system should also become more dependent on the metric as the metric becomes less reliable.

Removing or questioning the indicator will feel operationally threatening because the organization has built evaluation, coordination, and legitimacy around it.

Calls for direct observation will be resisted as subjective or unscalable. More measurement may be introduced to repair the failures created by measurement.

Evidence against the mechanism would include metrics that remain subordinate to direct outcomes, strong independent feedback, regular revision of indicators, and meaningful authority given to observations that contradict the dashboard.

Canonical Claim

Goodhart's Law becomes a mechanism of Reality Drift when optimization does more than corrupt a metric.

The metric becomes embedded as the system's operational representation of success. Real-world feedback loses authority, and the organization begins coordinating around the proxy rather than the condition the proxy was created to measure.

The metric continues to improve.

The system continues to function.

The gap between representation and reality continues to widen.

This formatting keeps related sentences together while separating definition, mechanism, consequence, and transition into distinct retrieval units.