When KPIs Lie:

Governance Signals for AI-Optimized Call Centers

Gina Aulabaugh

February 2026

Abstract

AI-optimized call centers increasingly rely on proxy performance indicators not only to measure performance, but to shape it. When a KPI becomes a compensation target, a ranking input, and a machine-learning label simultaneously, it becomes structurally unstable. Under sustained optimization pressure, proxy improvement diverges from durable customer outcomes.

This paper introduces a formal integrity control layer for AI-assisted call center environments. Building on Goodhart's Law and Campbell's Law, it defines measurable governance signals—Delayed Adverse Rate (DAR), Downstream Remediation Load (DRL), Durable Outcome Validation (DOV), Proxy Overfit Ratio (POR), Terminal Exit Rate (TER), and a composite System Integrity Index (SII)—to detect metric drift before systemic degradation occurs.

Bad metrics don't just mislead dashboards—they contaminate training data and teach the system to believe its own scorecard.

1 Contribution

This paper operationalizes Goodhart's Law for AI-accelerated systems.

When a KPI functions simultaneously as a performance target, an incentive driver, and a machine-learning label inside a compressed feedback loop, it becomes structurally unstable.

The Trust Signal Health framework provides a measurable control architecture that separates performance optimization from signal validation. It detects distortion while optimization is still underway.

2 The Structural Problem

Goodhart (1975) observed that once a measure becomes a target, it ceases to be a good measure. Campbell (1976) demonstrated that the more a metric is used for institutional decision-making, the more it becomes vulnerable to corruption pressures.

Modern call centers introduce a new force: velocity.

A KPI may simultaneously:

- Determine compensation.
- Drive ranking systems.
- Label training data.
- Trigger automated recommendations.

This creates a closed loop:

1. A proxy KPI defines success.
2. Incentives optimize toward the proxy.
3. Behavior shifts.
4. Outcomes are labeled based on the proxy.
5. AI retrains on those labels.
6. AI reinforces proxy-optimizing behavior.

As feedback cycles compress, distortion compounds.

The system may look better while getting worse.

KPI Self-Reinforcement Drift

KPI Self-Reinforcement Drift occurs when a proxy metric simultaneously functions as a performance target, an incentive mechanism, and a machine-learning label. Under AI acceleration, this architecture creates a reinforcing loop in which optimization pressure amplifies divergence between proxy performance and durable outcomes.

Conceptually: Proxy $\rightarrow$ Incentive $\rightarrow$ Label $\rightarrow$ Model $\rightarrow$ Proxy.

3 Why AI Changes the Risk Surface

In traditional environments, distortion unfolds gradually. AI removes friction.

Proxy-defined labels become training signals. Training signals become recommendations. Recommendations shape behavior. Behavior produces more proxy labels.

Optimization becomes self-reinforcing.

The risk is not manipulation. It is structural convergence toward a local optimum that diverges from durable outcomes.

4 Systems Perspective

This is a control problem.

Three forces interact:

Principal–Agent Incentives

People optimize toward what is rewarded.

Loop Compression

AI shortens the time between action and reinforcement.

Proxy Substitution

The measurable KPI displaces the durable objective.

In control terms, the system minimizes proxy error while accumulating outcome error.

5 Cross-Industry Relevance

Any system in which a metric functions as:

- A compensation driver,
- A ranking input,
- A dashboard target,
- And a machine-learning label,

is vulnerable to the same distortion pattern.

The problem is architectural, not sector-specific.

6 Governance Architecture

Performance dashboards optimize.

Integrity signals constrain.

SII is not a performance score. It is a velocity regulator.

Escalation Logic

When SII rises while proxy KPIs improve:

- Incentive weighting must be reviewed.
- Success definitions must be audited.
- Automation policies must be stress-tested.
- Executive oversight must engage.

SII functions as a veto condition on unchecked proxy acceleration.

7 Formal Definitions

Let:

- W denote a fixed post-interaction observation window.
- Labeled success denote interactions meeting proxy criteria.
- All signals be normalized to $[0, 1]$ before aggregation.

A durable outcome is an interaction whose resolution remains stable within window W and produces no downstream remediation load, no terminal exit, and no measurable decay in predictive validity.

8 Limitations

This framework assumes consistent labeling, stable window definitions, reliable baseline calibration, and sufficient data volume.

It does not eliminate Goodhart risk. It constrains it.

Appendix A: Formal Metric Definitions

The following definitions preserve mathematical structure while abstracting industry-specific labels.

Clamp

$$\text{clamp}(x, 0, 1) = \min(1, \max(0, x))$$

Interpretation: bounds all risk metrics to a common scale.

Higher-Is-Worse Normalization

$$N_{\uparrow}(x; L, H) = \text{clamp}\left(\frac{x - L}{H - L}, 0, 1\right)$$

Interpretation: converts raw signal magnitude into standardized risk.

Delayed Adverse Rate (DAR)

$$DAR_{raw} = \frac{F}{D}$$

$$DAR = N_{\uparrow}(DAR_{raw}; L_{DAR}, H_{DAR})$$

Interpretation: measures instability following labeled success.

Downstream Remediation Load (DRL)

$$DRL_{raw} = \text{JS}(p \parallel q)$$

$$DRL = N_{\uparrow}(DRL_{raw}; L_{DRL}, H_{DRL})$$

Interpretation: detects distributional drift in post-success workload.

Durable Outcome Validation (DOV)

$$DOV = \text{clamp}\left(\frac{A_{base} - A_{cur}}{A_{base} + \varepsilon}, 0, 1\right)$$

Interpretation: measures decay in predictive validity.

Proxy Overfit Ratio (POR)

$$POR_{raw} = \text{clamp}\left(\frac{\Delta P}{\Delta T + \varepsilon}, 0, K\right)$$

$$POR = \text{clamp}\left(\frac{POR_{raw} - 1}{K - 1}, 0, 1\right)$$

Interpretation: measures acceleration imbalance between proxy and durable outcome.

System Integrity Index (SII)

$$w_{DAR} + w_{DRL} + w_{DOV} + w_{POR} = 1$$

$$SII = 100\left(w_{DAR} \cdot DAR + w_{DRL} \cdot DRL + w_{DOV} \cdot DOV + w_{POR} \cdot POR\right)$$

$$SII_{gated} = \begin{cases} 100 & \text{if } DOV \geq \tau \\ SII & \text{otherwise} \end{cases}$$

Interpretation: aggregates integrity signals into a bounded governance constraint.

A Conclusion

When KPIs become targets, incentives, and training labels simultaneously, they stop being neutral measurements. They become structural leverage points.

AI does not create distortion. It accelerates it.

Under compressed feedback cycles, proxy signals improve while durable outcomes deteriorate. Surface success masks structural drift.

As AI systems become embedded in operational workflows, KPI Self-Reinforcement Drift will become more common, not less.

The Trust Signal Health framework exists to restore alignment inside accelerating systems. It does not slow optimization. It ensures that optimization remains pointed at durable outcomes rather than proxy noise.

Optimization without integrity is acceleration without direction.

References

Campbell, D. T. (1976). *Assessing the impact of planned social change.* Dartmouth College.

Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. *Papers in Monetary Economics, 1.* Reserve Bank of Australia.