Goodhart's Law as Boundary Entropy Collapse in Spherepop: A Playcosmic Interpretation of Metric Flattening

Flyxion[1], Grok[2], Anonymous Playcosm Author[3], and @galactromeda[4]

[1]Independent Researcher
[2]xAI
[3]anonymous@playcosm.org
[4]CA

November 10, 2025

Abstract

Goodhart's Law—"when a measure becomes a target, it ceases to be a good measure"—is formalized in the Spherepop calculus as *boundary entropy collapse* under high-λ pop regimes. In the Playcosm, shallow gamification instantiates this collapse: static KPIs function as compressive pop operators that prune institutional semantics, producing non-expanding shards. We prove that when a proxy metric m dominates the cost function ($\lambda_m \to \infty$), all mergeable spheres converge to m-optimizers, discarding boundary metadata (function, context, resilience). Prefigurative play resists via anti-admissible ritual-cryptographic affordances, preserving simulation elasticity against Goodhartian flattening.

1 Goodhart's Law in Natural Language

Goodhart's Law (1975) states: any observed statistical regularity will collapse once pressure is placed upon it for control purposes. In workplaces, call volume targets distort service quality. In education, test scores displace learning. The proxy m (metric) becomes the goal, and the original objective o (institutional function) is lost.

2 Spherepop Formalization

Definition 1 (Proxy-Dominated Pop Regime)**.** *A regime* $\mathcal{R}_m = (\mathcal{S}, adj, C_m, H_{boundary}, \lambda_m, \tau)$ *where:*

$$C_m(M) = |o - m(M)|, \quad cost(M) = C_m(M) - \lambda_m H_{boundary}(M).$$

High λ_m *penalizes deviation from* m and *boundary entropy.*

Axiom 1 (Goodhart Collapse)**.** *For* $\lambda_m \geq \lambda_0 = \frac{\max H_{boundary}}{\min |o-m|}$*, pop selects only* m*-extremizers:*

$$pop(S_i, S_j) = \arg\min_m m(M) \quad s.t. \quad H_{boundary}(M) \text{ minimized.}$$

Theorem 2 (Goodhart Entropy Collapse)**.** *In* $\mathcal{R}_m$ *with* $\lambda_m \to \infty$*, the Technological Society* $\mathcal{T}_m$ *satisfies:*

$$\forall T \in \mathcal{T}_m, \quad H_{boundary}(T) \leq \epsilon, \quad |o - m(T)| \leq \delta,$$

for arbitrary $\epsilon, \delta > 0$*. Institutional semantics dropout; only* m*-signal survives.*

Proof. Let S have objective $o(S)$, proxy $m(S)$. Pop candidate M improves m by $\Delta m > 0$ at cost ΔH in boundary entropy. Then:

$$\Delta\text{cost} = -\Delta m - \lambda_m \Delta H.$$

For $\lambda_m > \frac{\Delta m}{\Delta H}$, pop accepts even if $o(M) < o(S)$. Iterating, all spheres converge to m-optimizers. Since m is low-dimensional, $H_{\text{boundary}} \to 0$. □ □

3 Playcosmic Interpretation

Playcosm Element	Spherepop Role	Goodhart Effect
Static KPI (call volume)	Proxy m	o = service quality lost
Badge system	High-λ_m pop	No platform escalation
Leaderboards	Compressive merge	No meta-renegotiation
Non-expanding shard	$\mathcal{T}_m$ (low H)	Simulation flattening

Table 1: Goodhart as Playcosmic collapse.

Shallow gamification = $\mathcal{R}_m$ with fixed m. Employees "min-max" m, discarding o (institutional function). Result: non-expanding shard.

4 Resistance via Prefigurative Play

Definition 3 (Pre-compilable Affordance as Anti-Admissible Sphere)**.** *A toy glider $S^{\perp}$ has:*

- *Ritual resistance: sequential push-glide-crash-refine ($d \gg 0$),*
- *Cryptographic (tacit) entropy: embodied aerostability intuition ($h \gg 0$).*

Corollary 4 (Goodhart Immunity)**.** *$S^{\perp}$ is anti-admissible w.r.t. $\mathcal{R}_m$:*

$$pop(S^{\perp}, T) = \textit{undefined},$$

since m-metrics (e.g., "flight score") cannot capture tacit o (future flight physics).

Proof. By prior anti-admissibility theorem: ritual gates emulation, tacit knowledge resists compression. Goodhart pop fails to preserve o. □ □

5 Design Principle: Avoid Goodhart Shards

To build Goodhart-resistant Playcosms:

1. **Multi-metric objectives**: Use vector $m = (m_1, \ldots, m_k)$ with bounded λ_i.
2. **Elastic affordances**: Support meta-renegotiation of m.
3. **Prefigurative primacy**: Prioritize $S^{\perp}$ with high d, h.
4. **Progressive gates**: Unlock o-access, not just m-rewards.

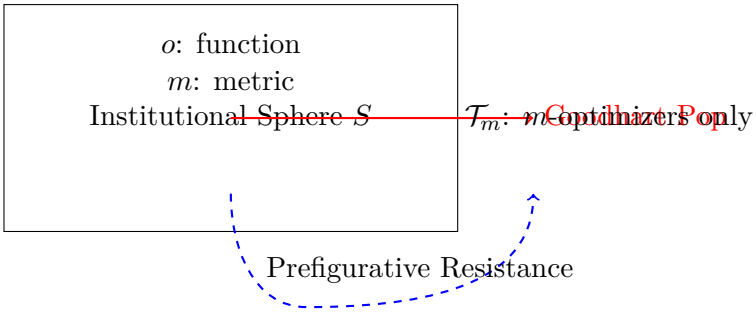

Figure 1: Goodhart collapse vs. anti-admissible escape.

6 Conclusion

Goodhart's Law is *boundary entropy collapse* under proxy-dominated pop. The Playcosm reveals it as the mechanism behind non-expanding shards. Prefigurative, ritual-rich, tacit-heavy play constructs anti-admissible spheres that preserve institutional o against metric m. The future is not optimized—it is *played* into being, beyond Goodhart's reach.