When the Signal Eats the Source

Randall Gossett
Independent Research — Extropy Engine
Academia.edu · lladnaros.com

Abstract

When a descriptive label acquires incentive value, status weight, or punitive association, actors stop optimizing for the underlying condition and start optimizing for the label itself. The result is a progressive divergence between the representation and the reality it was built to track. This paper formalizes that divergence as *representational fidelity decay*, modeled by a differential equation in which the decay rate is a function of social load and domain-specific correction strength. The model introduces three empirically grounded domains ordered by feedback latency — physical, institutional, and social/moral — and derives a k-parameter family of decay curves that predicts the collapse rate of label fidelity in each. The self-application problem is addressed directly: the model does not claim exemption from its own mechanism. Operationalization pathways, falsification criteria, and a bidirectional extension that allows fidelity restoration are provided. The model is then applied to the contemporary AI content detection problem as a running case study, demonstrating why regulatory enforcement mandates are subject to the same decay dynamic they attempt to solve.

1. Introduction

Labels are maps. Maps are not territories. This is not a new observation — Korzybski made it in 1933 and Borges illustrated it to absurdity in fiction. What is underexplored is the *rate* at which maps decouple from their territories, and the *conditions* that accelerate or slow that decoupling.

The proximate cause of decoupling is straightforward: when a label carries economic, social, or punitive weight, rational actors stop tracking the underlying territory and start tracking the label boundary instead. The label becomes the optimization target. This dynamic has been named Goodhart's Law — "when a measure becomes a target, it ceases to be a good measure" — and its variants: Campbell's Law (high-stakes labels attract corruption of the measure), and the Cobra Effect (incentive structures that perversely amplify the problem they were designed to solve). These are documented in monetary policy, education metrics, performance management, medical diagnosis, and platform content moderation.[1][2] [^3]

What the existing literature largely lacks is a *dynamical model* — one that captures not just the fact of decay but its rate, its domain-dependence, and the conditions under which fidelity can recover. This paper provides that model.

The central equation is:

$$\frac{dF}{dt} = -k \cdot S(t) \cdot F(t) + r \cdot C(t) \cdot (1 - F(t))$$

where F(t) is representational fidelity at time t, S(t) is social load, C(t) is corrective feedback strength, k is the domain's susceptibility coefficient, and r is the restoration rate. The first term drives decay; the second drives recovery. Most existing treatments address only the decay direction. Both are necessary for a complete model of how labels fail and how they sometimes self-correct.

2. Core Definitions

2.1 Representational Fidelity F(t)

Fidelity is the degree to which a label tracks its intended referent rather than the social incentive structure surrounding it. F = 1 represents perfect correspondence between the map and the territory. F = 0 represents complete semantic capture — the label now tracks only what passes the classifier, what satisfies the committee, or what earns the reward, regardless of whether the underlying condition exists.

F is not directly observable in all domains. It is estimated through domain-specific proxies:

- **Physical domains**: measurement error relative to ground truth; predictive validity of the label against observed outcomes
- **Institutional domains**: downstream performance predictive validity of credentials or ratings; inter-rater reliability over time; replication rates in science
- **Social/moral domains**: corpus-based semantic drift measured via cosine distance between embedding vectors across time periods; inter-rater reliability decay; divergence between stated and revealed preferences[4][5]

The operationalization is domain-specific by design. The model does not assume a universal measurement protocol. It assumes the *existence* of a meaningful territory against which label accuracy can, in principle, be assessed — even when that assessment is costly or delayed.

2.2 Social Load S(t)

Social load is the aggregate incentive, status, and punitive weight attached to a label at time t. It includes:

- Economic rewards for holding or bestowing the label
- Status gains associated with the label
- punishments for the absence of the label
- Reputational costs of challenging the label's validity

S(t) follows a logistic growth curve in the general case. Labels do not acquire social load linearly. They tend to be descriptive and low-stakes until a threshold event — an institutional endorsement, a regulatory mandate, a platform policy change, a viral cultural moment — causes rapid adoption of the label as a signal. After that inflection, S(t) saturates near a

maximum as the label becomes fully embedded in the incentive architecture. The logistic model captures both the slow early accumulation and the saturation behavior.

2.3 The Correction Parameter k and Feedback Strength C(t)

k is not a universal constant. It is a domain-specific susceptibility coefficient that encodes how strongly the territory pushes back against label drift. Domains with strong, rapid, low-cost causal feedback have low k. Domains with weak, delayed, or socially mediated feedback have high k.

C(t) is the active corrective feedback signal at time t — the strength of the mechanism by which the territory re-imposes accuracy on the map. It is nonzero only when feedback actually reaches the label system.

3. The Three-Domain Framework

The model partitions domains into three archetypes ordered by decreasing correction strength. This is not a categorical claim about discrete types. It is a parameterization strategy — three points on a continuous spectrum of feedback latency.

3.1 Physical Domain ($k \approx 0.02$)

Physical labels operate under direct causal correction. The label "load-bearing," "waterproof," "boiling point," or "bridge safe" cannot drift far from its referent before the territory responds with structural failure, measurement discrepancy, or observable contradiction. The feedback loop is fast, high-fidelity, and expensive to ignore.

Social load can still accumulate on physical labels — "organic," "natural," and "non-toxic" have all been subjected to substantial incentive pressure — but the underlying causal reality imposes a floor on how far the label can drift. Mislabeled pharmaceutical dosages kill people. Mislabeled structural tolerances collapse buildings. The corrective feedback is harsh, immediate, and not socially negotiable.

Even in physical domains, $k > 0$. The model does not claim any domain is immune. It claims the correction mechanism is strong enough to substantially limit decay rate.

3.2 Institutional Domain ($k \approx 0.15$)

Institutional labels — credentials, ratings, certifications, diagnoses, compliance designations — have feedback loops that exist but are slow, noisy, and bureaucratically mediated. The label "qualified physician," "investment grade bond," or "AI compliant" does not immediately self-correct when it diverges from reality. The correction mechanism operates through malpractice litigation, bond defaults, audits, replication failures, and regulatory enforcement — all of which introduce substantial lag between the label's divergence and its correction.

During that lag, the label is available for optimization. Campbell's Law specifically addresses this: "The more a measure becomes valued, the more pressure there is to cheat or corrupt the data to deliver it." Medical diagnostic inflation in the DSM, credential inflation in higher

education, and rating agency capture during the 2008 financial crisis are all documented instances of institutional label decay under elevated social load.[2][3] [^1]

Institutional labels can recover fidelity through scandal, reform, or catastrophic outcome feedback — which is why the bidirectional model (with the restoration term r·C(t)·(1-F(t))) is essential for this domain. The restoration term activates when C(t) spikes — when a scandal erupts, when defaults cascade, when a replication crisis forces retraction.

3.3 Social/Moral Domain (k ≈ 0.45)

Social and moral labels — "authentic," "harmful," "safe," "credible," "human-made," "problematic," "verified" — have the weakest corrective feedback of the three archetypes. The territory they purport to track is partly constituted by consensus itself. There is no structural failure mode that forces the label back to its referent. There are no replication crises in the ordinary sense. The "territory" of a moral label is intersubjectively constructed, meaning the very act of disputing the label can be framed as violating the label's requirements.

This creates a runaway dynamic not present in the other domains. As social load increases, the label's defenders often intensify the social cost of challenging it, which further suppresses corrective feedback, which accelerates decay. The model captures this in the k coefficient — high susceptibility combined with near-zero C(t) produces rapid, potentially irreversible fidelity collapse.

Social/moral labels do occasionally recover fidelity, but the restoration mechanism is typically external — empirical catastrophe, exposure of fraud, or cross-domain collision with physical reality (a health intervention labeled "safe" that kills people; a financial instrument labeled "ethical" that funds atrocities). These events inject C(t) from outside the social system, triggering the restoration term.

4. The Complete Dynamical Model

4.1 Primary Decay Equation

The decay component of the full equation is:

$$\left.\frac{dF}{dt}\right|_{\text{decay}} = -k \cdot S(t) \cdot F(t)$$

Including F(t) as a multiplicative factor — rather than the simpler -k·S(t) — has two benefits. First, it prevents fidelity from going negative: as F approaches zero, the decay rate also approaches zero, so the model has a natural floor at F = 0. Second, it more accurately models the intuition that a label already at near-zero fidelity is harder to degrade further, since the optimization pressure has already been fully absorbed.

The solution to the decay-only equation is exponential decay under accumulated social load:

$$F(t) = F_0 \cdot \exp\left(-k \int_0^t S(\tau)\, d\tau\right)$$

For a logistic S(t), this produces a characteristic S-shaped collapse: slow decay during the early low-load phase, rapid collapse through the inflection region as social load saturates, and asymptotic approach to zero afterward. The k parameter determines how fast the collapse propagates once load saturates.

4.2 The Bidirectional Model with Restoration

The full model adds a restoration term:

$$\frac{dF}{dt} = -k \cdot S(t) \cdot F(t) + r \cdot C(t) \cdot (1 - F(t))$$

The restoration term r·C(t)·(1-F(t)) has the following properties:

- It is zero when C(t) = 0 (no corrective feedback)
- It is bounded by r when F = 0 (maximum restoration force when fidelity is fully collapsed)
- It goes to zero as F approaches 1 (no restoration needed when fidelity is already high)
- It is controlled by r, a domain-specific restoration rate that reflects how efficiently the domain converts corrective input into fidelity recovery

This formulation allows the model to capture documented recovery events: the post-Enron accounting reforms that temporarily improved label fidelity in financial reporting; the replication crisis correction wave in psychology; the EU medical device regulatory overhaul after device failures. These are not anomalies. They are the restoration term activating.

4.3 The Standardization Phase

One important early-phase behavior the model must accommodate: social load sometimes initially *improves* fidelity before degrading it. When a label becomes important enough that institutions begin rigorously defining it — standardizing measurement protocols, auditing compliance, investing in ground-truth verification — the correction mechanism C(t) also strengthens during the early load phase. This can produce a brief fidelity improvement before the gaming dynamic dominates.

The model captures this through the interaction between the decay and restoration terms. In the early phase, if C(t) grows proportionally with S(t) (because institutions respond to the label's importance by investing in verification), the net effect can be dF/dt > 0. Only after gaming strategies mature and begin outpacing institutional verification does the decay term dominate.

This means the model predicts four observable phases for most social labels:

1. **Descriptive phase**: Low S(t), low gaming, high F. The label is a functional map.
2. **Standardization phase**: Rising S(t), institutions invest in verification, C(t) rises, fidelity may improve slightly.
3. **Capture phase**: S(t) saturates, gaming strategies proliferate, C(t) cannot keep pace, fidelity decays.

4. **Collapse or correction phase**: Either F approaches zero and the label is fully semantic (tracks only the classifier, not the territory), or an external C(t) shock triggers partial restoration.

5. Falsification Criteria

The "Falsifiable Test" label on any visualization of this model is not satisfied by simulated output curves. It requires empirical operationalization. The following three tests would falsify the model's core k-ordering claim:

Test 1 (k-inversion): Identify a domain with strong physical causal feedback (k-analogous to the physical archetype) and a domain with weak, delayed, social feedback (k-analogous to the social/moral archetype), subject both to equal social load S(t), and find that the socially mediated domain maintains higher fidelity over time than the physically constrained one. This would require the k-ordering to be wrong.

Test 2 (load-insensitive social collapse): Find a social/moral label under high sustained social load that maintains stable, empirically verifiable fidelity across multiple measurement periods without external corrective input. Semantic drift analysis using corpus embedding methods provides the measurement protocol: if the cosine distance between the label's embedding vector in the high-load period and its original descriptive embedding remains near zero despite high S(t), the model is falsified.[5][4]

Test 3 (restoration inhibition): The model predicts that injecting a strong C(t) signal into a decayed institutional label should produce measurable fidelity recovery. A regulatory overhaul or scandal-driven reform that fails to move the inter-rater reliability or predictive validity of the affected label would falsify the restoration term's behavior.

These tests are not easy. They require longitudinal data, agreed-upon measurement protocols, and domain-specific operationalization of F. But they are specified, in principle executable, and would constitute genuine evidence against the model if the predicted outcomes did not materialize.

6. The Self-Application Problem

This is addressed directly because it is the most common line of attack on the theory, and because addressing it honestly is better than leaving it open.

The model is itself a representation under social load. "When the Signal Eats the Source" is a label with growing incentive value — it explains something people want explained, it provides social utility, and its author has a reputational stake in its propagation. By the model's own mechanism, this paper is subject to fidelity decay if it acquires sufficient social load and if the corrective feedback mechanism (peer review, empirical testing, adversarial critique) weakens relative to the incentive to adopt it as an identity signal.

This is not a contradiction. It is a scope condition. The paper does not claim $k = 0$ for itself. It claims the model is an institutional-domain artifact with $k \approx 0.15$ — subject to moderate decay,

correctable through adversarial critique, replication attempts, and empirical testing. The self-application problem is answered not by exempting the model from its own mechanism but by applying the mechanism to the model explicitly and deriving the prediction: if this paper becomes a status signal rather than a functional map, its fidelity will decay at an institutional-domain rate unless subjected to ongoing correction.

The corrective feedback mechanism for this paper is: falsification attempts, operationalized tests, peer criticism, and empirical challenge. This critique is the mechanism, not the contradiction.

7. Case Study — AI Content Detection and the "Human-Made" Label

The contemporary AI content detection regulatory regime provides a real-time demonstration of the model.

7.1 Current S(t) Position

The label "human-made" or "verified human creator" is currently in the early-to-mid standardization phase — S(t) is rising but has not yet saturated. The EU AI Act's Article 50 transparency provisions, enforceable from August 2026, attach significant economic and punitive weight to this label: platforms face fines up to €15 million or 3% of global annual turnover for non-compliance with AI content detection requirements. Spotify's "Verified by Spotify" badge system and YouTube's monetization policies create additional economic valence. These represent a large, simultaneous S(t) spike across multiple platform domains.[6][7][8][9]

7.2 Correction Mechanism Weakness

The corrective feedback available to the "human-made" label is structurally weak. Unlike a physical domain where the territory pushes back through direct causal consequence, the "human-made" territory has no self-enforcing corrective mechanism. The detection systems used to assess the label are:

- Systematically bypassable through adversarial perturbation, spectral processing, re-generation attacks, and stem-level manipulation[10][11]
- Subject to false positive rates ranging from 15% to 50% depending on sensitivity tuning, meaning the label will be incorrectly assigned to human-created content at scale[12][13]
- Limited to detecting watermarks from cooperative generators — content from non-watermarked generators cannot be flagged regardless of detection investment[14][15]

By the model's prediction, a label with high and rising S(t), domain-level k closer to institutional or social/moral than physical, and a correction mechanism known to be defeatable by anyone with moderate technical sophistication will undergo rapid fidelity decay. Within a compressed time period, "detected as human-made" will cease to track "actually human-made" and will track instead "successfully bypassed the current classifier version."

7.3 Prediction

The model predicts that the label "human-made content" will follow an institutional-to-social collapse path under current regulatory conditions — high social load, moderate-to-weak correction, active adversarial optimization. Fidelity will decay faster than platforms' ability to update detection systems, because:

1. The adversarial compute available to individual creators equals or exceeds the marginal detection improvement available to platforms (see: SynthID reverse-engineering at under $3 cost)[^10]
2. The economic incentive asymmetry favors attackers — a creator protecting their livelihood will outspend the marginal enforcement effort
3. False positive collateral damage will create institutional pressure to *reduce* detection sensitivity, weakening C(t) from the inside

This is not a speculative prediction. It is the direct output of applying the model to the domain parameters.

8. Discussion

8.1 Relation to Existing Literature

The model extends Goodhart's Law from a static adage to a dynamical system. The formal Goodhart analysis in the recent mathematical literature treats $M = G + \xi$, where M is the measure, G is the true goal, and ξ is the discrepancy — and finds that strong Goodhart effects require long-tail distributions of ξ. The fidelity decay model is compatible with this: the k parameter can be interpreted as a function of the tail behavior of the map-territory discrepancy distribution. High-k domains (social/moral) are precisely those where ξ has long, unconstrained tails — there is no physical cutoff limiting how far the label can drift from its referent.[16][1]

Campbell's Law contributes the observation that gaming scales with value. The adversarial Goodhart effect captures the deliberate optimization against the classifier boundary. The current model unifies these into a single dynamical equation with domain-specific parameters and adds the restoration term missing from all prior treatments.[^2]

The social feedback literature provides empirical grounding for the social load mechanism. Group-based reputational incentives systematically weaken sensitivity to underlying outcomes — subjects have been shown to divert resources from more effective to less effective charities in order to maintain group reputational position, even when they know the underlying outcome difference is real. This is S(t) operating in a controlled experimental setting.[^17]

The semantic drift measurement literature provides the measurement protocol for F in social/moral domains. Corpus-based distributional semantic analysis using cosine distance between decade-specific word embeddings has been validated as sensitive to both cultural shifts and linguistic drift. Label fidelity in the social/moral domain can be operationalized as

the inverse of semantic drift rate from the label's original descriptive embedding toward its incentive-era embedding.[18][4]

The quantum probabilistic cognition literature provides complementary evidence that reasoning in social domains is not context-free and is subject to systematic non-classical effects — order effects, framing effects, and identity-based reasoning — that are absent in physical domains. This supports the theoretical grounding for why social/moral domains have qualitatively weaker correction mechanisms than physical ones.[^19]

8.2 Scope Conditions

The model applies when:

1. A label has a meaningful referent that exists independently of the label — there is a territory the map can be accurate or inaccurate about
2. Social load on the label is measurable in principle, even if not always precisely
3. The domain has a correction mechanism, however weak or delayed
4. The time scale of interest is long enough for the decay dynamics to manifest

The model does not apply to purely performative labels — labels that constitute their referent by being uttered, where there is no independent territory (promises, legal verdicts, ceremonial declarations). Nor does it directly address cases where the "territory" itself changes faster than the label can track, which is a distinct problem from social load distortion.

8.3 Limitations

The k values used in illustrations (0.02, 0.15, 0.45) are stipulated parameters chosen to produce characteristic decay curves. They are not empirically derived from measurement campaigns. A full empirical program would estimate k by measuring fidelity decay rates in controlled natural experiments — domains undergoing a known sudden S(t) spike (a new regulatory mandate, a major scandal, a platform policy change) — and fitting the model parameters to the observed decay trajectory. That program is outside the scope of this paper but constitutes the natural next step.

The model also treats k as fixed within a domain archetype. In practice, k is probably best understood as a continuous function of feedback latency, with individual labels varying along the spectrum. A label like "organic food" sits somewhere between physical and institutional, with k intermediate between 0.02 and 0.15. A label like "trustworthy journalist" sits between institutional and social/moral. Future work should treat k as a measurable quantity derived from feedback latency measurements rather than a stipulated domain constant.

9. Conclusion

Representations decay when the incentive to hold the label exceeds the cost of inaccuracy. The rate of that decay is not uniform across domains. It is governed by the strength and speed of corrective feedback — the degree to which the territory can impose accuracy on the map despite the social pressure to drift.

The full model, including the restoration term, predicts four phases: descriptive accuracy, possible standardization-era improvement, capture and collapse, and either permanent semantic emptying or correction-shock recovery. All four phases are empirically attested in documented label histories.

The model applies to itself at $k \approx 0.15$. It will degrade if adopted as a status signal without ongoing empirical challenge. The correction mechanism is this: if the falsification criteria laid out in Section 5 are tested and found disconfirmed, the model is wrong and should be revised. That is not a rhetorical hedge. It is the only claim a map can responsibly make about a territory it cannot fully see.

Keywords: Goodhart's Law, representational fidelity, semantic drift, social load, map-territory correspondence, dynamical systems, AI content detection, Campbell's Law, metric capture, feedback latency

References

1. Goodhart's law - Goodhart's law is an adage that has been stated as, "When a measure becomes a target, it ceases to b...
2. Goodhart's Law - Goodhart's Law is a reminder that measures that become targets will distort behaviours in generally ...
3. Goodhart's Law - Goodhart's Law states that “when a measure becomes a target, it ceases to be a good measure.” In oth...
4. Cultural Shift or Linguistic Drift? Comparing Two ... - PMC - We show how two computational measures can be used to distinguish between semantic changes caused by...
5. Semantic Drift Analysis - Semantic drift analysis quantifies shifts in linguistic and symbolic meanings over time and across t...
6. EU AI Act Penalties Explained: Avoid Costly Non-Compliance - The Act uses a three-tier penalty structure with fines scaling from 1% to 7% of turnover based on vi...
7. The EU AI Act Fine Calculation Most Boards Are Doing ... - For standard enterprise violations, regulators calculate both 3% of your global annual turnover and ...
8. Spotify adds 'Verified' badges to distinguish human artists ... - Spotify is introducing a 'Verified' badge to help users identify when artists on its platform are hu...
9. Spotify's new badge identifies human artists, as AI music ... - The audio streaming platform has launched a certification tool that indicates whether a song was cre...
10. Attempting model extraction of Google DeepMind SynthID ... - Deep dive into extracting and attacking Google DeepMind's SynthID image watermarking system using ad...

11. reverse engineering Gemini's SynthID detection - Built a detector that identifies SynthID watermarks with 90% accuracy; Developed a multi-resolution ...
12. AI detectors are easily fooled, researchers find - Callison-Burch said the trouble with accuracy rates is that they often neglect false positives: Anyo...
13. AI detectors have a 15% false positive rate. That means ... - PSA: AI detectors have a 15% false positive rate. That means they flag real human writing as AI cons...
14. SynthID - SynthID is a tool to watermark and identify AI-generated content, helping to foster transparency and...
15. Google opened SynthID for text but kept image, audio and ... - Google DeepMind open sourced SynthID for text, but their image, audio and video watermarking tech is...
16. On Goodhart's law, with an application to value alignment - “When a measure becomes a target, it ceases to be a good measure”, this adage is known as Goodhart's...
17. Group-based reputational incentives can blunt sensitivity to ... - Group-based reputational incentives can weaken-and sometimes nearly eliminate-affective differentiat...
18. Cultural Shift or Linguistic Drift? Comparing Two ... - We show how two computational measures can be used to distinguish between semantic changes caused by...
19. A quantum probabilistic framework for reasoning ... - The current work addresses this by outlining the representational scaffolding necessary to explain w...