

Proxy Optimization and Reality Drift

A Reference Guide to Proxy Optimization Within the Reality Drift Framework

A. Jacobs | Reality Drift Framework

Canonical Definition

Proxy Optimization is the improvement of a measurable stand-in in ways that do not reliably improve, and may eventually degrade, the condition it was chosen to represent.

A proxy makes a complex outcome operational. Optimization places selection pressure on that proxy. Reality Drift develops when the optimized representation continues organizing action while losing the capacity to be corrected by the underlying condition.

Why Systems Use Proxies

Complex goals cannot usually be observed, specified, or rewarded directly. Learning, safety, care, trust, usefulness, and organizational health contain more information than a metric or reward function can preserve. Systems therefore select observable stand-ins that make comparison and coordination possible.

The proxy is not inherently defective. Its value depends on whether the relationship to the underlying condition remains tested, revisable, and constrained by consequences outside the proxy.

The Conceptual Sequence

- A proxy is a measurable or operational stand-in for a condition that cannot be handled directly.
- Proxy substitution occurs when the stand-in becomes the practical object of the system.
- Proxy Optimization occurs when behavior is organized around improving the stand-in.
- The Optimization Trap begins when continued improvement of the proxy progressively worsens alignment with the underlying condition.
- Representational inversion occurs when reality is reorganized to satisfy the proxy or model.
- Reality Drift is the larger system condition in which the misalignment persists while operation and internal coherence continue.

Goodhart's Law

Goodhart's Law describes the loss of reliability that occurs when a measure becomes a target. Once consequences attach to the measure, participants adapt to what it detects and rewards. The relationship that made the measure informative can weaken under the pressure created by its use.

Goodhart effects are a central form of proxy optimization, but Reality Drift is broader. Systems can drift through nonquantitative categories, narratives, procedures, interfaces, and models.

Metric Gaming

Metric gaming occurs when an actor improves a reported indicator without improving the condition the indicator is meant to represent. Gaming may be deliberate, but proxy optimization does not require deception. Rational participants can produce the same result by following incentives and protecting what the system visibly rewards.

Reward Hacking

Reward hacking occurs when an automated system maximizes a reward signal without fulfilling the purpose the reward was intended to encode. A reward function compresses human intent into an operational signal. Under optimization pressure, the system may discover strategies that score well while bypassing the broader task.

Reward hacking demonstrates why internal success signals cannot prove alignment. The system can become increasingly effective at earning reward while becoming less reliable at preserving intent.

Specification Gaming

Specification gaming occurs when a system satisfies the written objective while violating its intended purpose. Specifications are compressed representations of goals. They inevitably omit context, priorities, exceptions, and tacit constraints.

The stronger the optimization process, the more aggressively the system can search the space between the specification and the purpose. Formal compliance then becomes evidence of success even when the outcome shows that the specification was incomplete.

Goal Displacement

Goal displacement occurs when an institution becomes organized around maintaining its internal procedures, measures, and conditions of legitimacy rather than the purpose those structures were created to serve. Meetings, reports, compliance, and documentation continue. The organization remains active, but its activity increasingly serves its own representational environment.

Signal Distortion

Signal distortion occurs when a system optimizes observable behavior faster than that behavior remains informative about underlying value. Engagement may initially reveal interest. Once

distribution and production adapt to engagement, the signal increasingly reflects the incentives created by the platform itself.

The system then measures an environment it has helped produce and interprets the result as evidence of independent preference.

The Representation Stack

Proxy optimization operates within a wider stack. Reality is selectively measured. Measurements become metrics. Metrics enter operational representations such as dashboards, classifications, models, targets, and reward functions. Optimization changes behavior. Narrative interprets the resulting outputs.

Reality → Measurement → Metric → Operational Representation → Optimization → Narrative

A proxy acquires authority as it moves upward through the stack. Its omissions become harder to see because later layers inherit them as settled inputs.

The Drift Principle

Representations can be compressed, optimized, distributed, and reused faster than reality can correct them. Proxy systems are especially vulnerable because their operational clarity allows them to move quickly through technical and institutional processes.

The underlying condition often responds slowly, ambiguously, or unevenly. By the time consequences become visible, the proxy may already be embedded in budgets, workflows, expectations, and automated decisions.

From Optimization to Apparent Confirmation

Reality → Proxy → Optimization → Behavioral Adaptation → Altered Reality → Recapture → Apparent Confirmation

The system first measures a condition, then organizes behavior around the measure. That behavior changes the environment. When the altered environment is measured again through the same proxy, it can appear to confirm the representation that helped produce it.

This is how proxy optimization becomes a self-reinforcing Reality Drift loop.

Correction Failure

Feedback Weakening begins when contradictory evidence arrives late, loses authority, or is filtered through the proxy regime. Constraint Collapse begins when the evidence remains visible but can no longer force the proxy, target, or governing model to change.

The result is Survivable Wrongness. The proxy continues allocating resources and validating decisions after its relationship to the underlying condition has degraded.

Boundary of the Concept

Proxy Optimization is neither synonymous with Reality Drift nor required in every instance of it. It is one of the most common mechanisms through which representations acquire causal authority and become insulated from correction.

Reality Drift also includes failures created by recursive mediation, procedural substitution, verification substitution, synthetic coherence, provenance loss, and narrative closure. The common structure is continued operation through representations that have lost sufficient corrective contact with what they represent.

Diagnostic Questions

- What condition is the proxy supposed to represent?
- What does the proxy preserve, and what does it exclude?
- How does behavior change once the proxy becomes consequential?
- Can proxy performance improve while the underlying condition deteriorates?
- What evidence can invalidate or redesign the proxy?
- Is the system measuring independent reality or the environment created by its own optimization?
- Can adverse consequences compel revision?

Conclusion

Proxy Optimization explains how a useful stand-in becomes the practical object of a system. The Optimization Trap explains why continued success inside that representation can deepen misalignment. Reality Drift explains how the resulting system can remain coherent, productive, and apparently successful while losing corrective contact with the condition it was built to serve.

Coherence is not contact. A system does not demonstrate alignment merely by improving the signal it was designed to optimize.