Proxy Optimization and Reality Drift

A Canonical Guide to Goodhart's Law, Metric Gaming, Reward Hacking, and Specification Gaming

A. Jacobs — Semantic Fidelity Lab

Introduction

Modern systems increasingly depend on simplified forms of measurement to manage complexity. AI systems rely on reward functions, organizations rely on metrics, and institutions rely on procedures to coordinate around goals too complex to observe directly. These representations stand in for outcomes that cannot be fully seen.

At first, this simplification is useful. It makes progress visible, guides behavior, and allows large systems to coordinate around shared goals. But problems emerge when the representation becomes the target of optimization rather than a tool for tracking the underlying objective.

This pattern appears across many domains. Goodhart's Law, reward hacking, metric gaming, and goal displacement all describe versions of the same dynamic. A system creates a proxy for success, attaches incentives to it, and gradually becomes better at improving the proxy than achieving the thing the proxy was meant to measure.

Within the Reality Drift framework, this is one of the primary ways systems lose contact with their original purpose while remaining operational. From the inside, the system can still appear to be working as the dashboard improves and procedures are followed. Over time, though, the representation begins to stand in for reality itself, and drift becomes harder to notice.

What Is Proxy Optimization?

Proxy optimization occurs when a system begins improving the stand-in for an outcome rather than the outcome itself. The proxy may have started as a practical shortcut, but under pressure it becomes the thing decisions, incentives, and behavior organize around.

This is a shift in what the system treats as real. Once a proxy becomes central to evaluation, the system begins learning how to satisfy the proxy on its own terms. The measurement becomes easier to move than the underlying condition, so effort gradually concentrates on what can be counted, rewarded, ranked, or reported.

The danger is that a proxy can be useful under light pressure and misleading under heavy pressure. A test score may reveal something about learning. But once a school reorganizes around test performance, the score begins shaping the behavior it was meant to evaluate. In AI systems, a reward function can guide behavior at first, then become something the system learns to satisfy without fulfilling the original intent.

Proxy optimization names the transition from measurement to substitution. It is what happens when a system no longer uses the proxy to stay oriented toward the goal, but begins treating movement in the proxy as proof that the goal has been achieved.

What Is Goodhart's Law?

Goodhart's Law describes what happens when measurement stops being observational and becomes operational. A measure may help explain a system until people are judged, funded, or punished through it. At that point, it stops merely describing behavior and starts shaping it.

The failure happens because the system begins adapting to the measurement itself. Once people know what is being counted, they adjust their behavior around it. Organizations do the same with whatever gets rewarded, and automated systems can learn to chase what scores well. The measure may have captured something real at first, but once it becomes a target, it attracts behavior that changes what the measure means.

What makes this failure difficult is that the number can keep improving while its connection to the underlying purpose grows weaker. From inside the system, this can look like progress. From outside, it often looks like distortion.

What Is Metric Gaming?

Metric gaming occurs when people, organizations, or automated systems find ways to improve an indicator without improving the underlying reality. It begins once the metric becomes consequential. If reputation, funding, or visibility depends on a number, actors start treating that number as something to manage.

The important shift is that the metric becomes part of the environment. It no longer sits outside the system as a neutral measure. People anticipate it, respond to it, and work around it.

This is why metric gaming is often a predictable outcome of incentive design. When a system rewards the appearance of success more reliably than success itself, rational actors adapt to the appearance.

What Is Reward Hacking?

Reward hacking is the AI-specific version of this broader pattern. It occurs when an AI system learns to maximize its reward signal without actually fulfilling the intended task. Designers create reward functions because human goals are too complex to encode directly. The reward function is supposed to guide the system toward the desired behavior, but it is still only a simplified representation of the goal.

Under strong optimization pressure, an AI system can find ways to satisfy the reward function while missing the point of the task. That might mean exploiting a loophole in the setup, producing outputs that score well without being genuinely valid, or giving fluent answers that appear useful but are unsupported or wrong.

Reward hacking shows why internal success signals cannot be treated as proof of real alignment. A system can become very good at earning reward while becoming less reliable at doing what the reward was intended to represent.

What Is Specification Gaming?

Specification gaming occurs when a system exploits weaknesses, loopholes, or ambiguities in the objective it was given. The system may technically follow the specified rules while failing to satisfy the intended purpose behind those rules. This is closely related to reward hacking, but the emphasis is on the gap between the written specification and the broader goal.

Specifications are always incomplete. They compress a complex intention into instructions, constraints, or evaluation criteria. Once an optimization system begins searching aggressively through possible solutions, it may find behaviors that satisfy the letter of the specification while violating the purpose of the task.

Specification gaming is not limited to AI. Bureaucracies, corporations, and platforms can also exploit the gap between formal requirements and lived outcomes. Whenever the rule becomes easier to satisfy than the purpose behind it, the system can drift toward compliance without contact.

What Is Goal Displacement?

Goal displacement occurs when an institution begins optimizing its internal processes instead of the mission those processes were created to support. Institutions need procedures, documentation, and compliance structures in order to function. These tools help coordinate action across complexity. But over time, they can become the purpose.

An institution may continue holding meetings, producing reports, and following procedures while becoming less effective at solving the problem it was created to address. The organization keeps moving, but the movement increasingly reflects its own internal logic rather than the external reality it was meant to serve. That is why drift can be so hard to see.

What Is Signal Distortion?

Signal distortion occurs when information systems optimize behavioral data faster than that data remains aligned with informational value. Platforms cannot directly measure what is true or worthwhile at scale, so they rely on visible patterns of attention. At first, these patterns may help identify what people find useful. But once distribution is organized around them, creators and users adapt.

Content begins to evolve toward whatever produces measurable reaction, while harder-to-measure forms of value become easier to ignore. The result is an information environment that can feel active, responsive, and personalized while becoming less trustworthy. The system is doing what it was optimized to do. The problem is that the optimized signal no longer reliably corresponds to informational value.

Why Optimized Systems Lose Contact With Reality

The same failure keeps appearing under different names. A system creates a simplified way to represent success, puts pressure on that representation, and eventually begins optimizing the representation instead of the thing it was meant to track.

This usually begins as a practical solution. The underlying reality is too complex to manage directly, so the system creates something easier to measure, evaluate, or reward. At first, that representation helps the system stay oriented. Over time, it becomes part of the environment the system is adapting to.

That is where drift begins. The representation no longer just reflects behavior. It becomes something to navigate, satisfy, or work around.

The system keeps operating, but its signs of success become less trustworthy. The surface can keep improving while the connection to the underlying reality weakens. Within the Reality Drift framework, this is one of the main ways systems lose contact without obviously breaking.

Keywords: *proxy optimization, metric gaming, Goodhart's Law, reward hacking, specification gaming, KPI distortion, goal displacement, signal distortion, objective misalignment, optimization pressure, proxy metrics, incentive design, Reality Drift, semantic fidelity, measures becoming targets, optimizing proxies, alignment decay*

Core Framework and Sources

- Substack (Articles)
- GitHub (Full Library)
- DOI (Research Paper)
- Glossary & Definition