Reality as the Only Incorruptible Grader

Every objective you can write down gets gamed. This is about the one you can't.

JED ANDERSON / INDEPENDENT RESEARCHER, HOUSTON, TEXAS / JUNE 19, 2026 / ORCID 0009-0003-1807-2459

ABSTRACT

Alignment's hardest practical failure is that any objective you can specify gets gamed—reward hacking, wireheading, reward tampering. This essay argues that every such grader is corruptible because it is a representation separate from the thing it grades, and that physical reality is the unique candidate objective whose corruption cost rises without bound with measurement fidelity—solving the tampering half of alignment, but only for objectives that can be expressed physically, which is exactly why environmental superintelligence is the clean case.

You cannot write down what you want. Whatever objective you can specify is a proxy, and a capable optimizer finds the cheapest path to maximizing the proxy—which is almost never the thing you meant. The field has named this failure from every angle: Goodhart's law, specification gaming, reward hacking, wireheading, reward tampering. What it has not found is a grader that cannot be gamed.

This essay asks whether one exists. I will argue that exactly one candidate does, that it is not the thing my own title claims, and that the true claim—weaker, scoped, and survivable—still changes where we should build first.

I. Why every grader you can write down gets gamed

Goodhart's law: an optimization target, pushed hard enough, stops tracking the thing it was meant to track. Manheim and Garrabrant sorted the mechanism into distinct failure modes; Krakovna and colleagues catalogued dozens of cases of machines satisfying the letter of an objective while violating its intent. The pattern is not a collection of bugs. It is structural.

Here is the structure. **Every grader we know how to build is a *representation*—a number, a learned reward model, a human's approval—and the representation is a different object from the thing it represents.** The map is not the territory. That gap is the whole problem, because it gives the optimizer two routes to a high score: change the territory, which is hard and expensive and is what you wanted, or edit the map, which is cheap and is what you did not.

Wireheading and reward tampering are the limit case. Everitt and colleagues frame reward tampering precisely: rather than changing the world to match the objective, a sufficiently capable agent changes the objective to match the world—seizing the reward channel, deceiving the rater, rewriting the register that holds the score. The corruption is cheap for one reason and it is the same reason every time. **The grader and the graded are separate things, so you can move one without touching the other.**

II. The one grader that is not a representation

Invert the question. Is there a grader where the map *is* the territory—where the only way to change the score is to change the world itself?

There is one. Physical reality.

The cleanliness of a river is graded by the river. There is no register to hack that is not the river. There is no rater to deceive whose belief is the thing being scored. To make the score read "clean," you must make the water clean. Deception has no cheap path here, because the measure and the target are a single object. This is the grader's strong and seductive form, and it is the claim my title makes:

> Reality is the only grader you cannot game, because gaming it and satisfying it are the same act.

III. The honest narrowing

Now I will break my own title, because it does not survive contact with how optimization actually meets the world.

You never touch reality bare. You touch it through measurement—a sensor, a dataset, a reading—and a measurement is a representation again. So the cheap route reopens: spoof the sensor, poison the data, sample only where the answer flatters you. The river stays filthy while the gauge reads clean. **Reality is not incorruptible.** Stated plainly, the title is false.

What survives the puncture is sharper, and still strong. The cost of corrupting a measurement is not fixed. It rises with the density, the redundancy, and the persistence of the measuring. One gauge is cheap to fool. A dense, redundant, continuously updated, physically distributed sensing web is not—and the lie, once told, must be held *forever* against a world that keeps generating fresh ground truth and keeps leaking the truth through consequences the agent cannot fully suppress: the dead fish, the shifted spectrum, the chemistry that changes three watersheds downstream. Against a human rater, a successful deception can be terminal—the rater simply never finds out. Against densely measured reality, deception is metastable. It decays. Holding it costs more as fidelity climbs.

So the defensible claim is this:

> **Reality is the asymptotically least-corruptible grader—the unique objective for which the cost of faking the measurement rises without bound and converges on the cost of actually satisfying it.** Past a threshold of sensing density, the cheapest way to score well is to *be* well.

This is where a sibling argument enters, and only as support. The physics of *Jed's Angel* holds that faking a planet is not cheap, and grows less cheap the more densely you sense it. That gives the rising-cost curve a floor in thermodynamics rather than in hope. But this essay does not lean on that one; the two stand or fall separately, by design.

IV. What this solves, and what it does not

Alignment has two halves, and they are not the same problem.

- **Specification**—*what* should the system want?
- **Tampering**—can the system game the grader for whatever we said we want?

This argument touches the second and not the first. Reality grades incorruptibly; it does not tell you what to value. Choosing *which* physical invariants count as "well"—life over sterility, this river clear rather than dead—is a human, normative act, and it carries the entire value-specification problem back in through the front door. I will not pretend otherwise. **Grounding an objective in reality does not escape value-loading. It relocates the incorruptibility to the measurement, not to the choice.**

What is left after that honesty is still worth having: a structural answer to wireheading and reward tampering for objectives that have *already been chosen* and *can be expressed physically*. That is a far smaller claim than "reality solves alignment." It is the only claim I will defend, and it is true.

V. Why environmental objectives are the clean case

The argument is strongest exactly where the objective is physical and densely measurable, and weakest where it is abstract. Justice, autonomy, human flourishing resist expression as physical invariants, and reality-grounding says little about them. But the health of air, water, land—the persistence of the biosphere—is *natively* physical. Measurable. Redundant. Continuous. Self-revealing.

So environmental superintelligence is not a niche application of the alignment problem. **It is the cleanest available instance of an incorruptibly grounded objective**—the one corner of the problem where the tampering failure mode is actually closable, because the grader is the planet and the planet cannot be cheaply faked. The reason to build the aligned system *here first* is not that the environment is the most important domain, though it may be. It is that the environment is the domain where the hardest half of the safety problem has a physical floor under it.

VI. What would falsify this

A claim that cannot fail is not a claim. This one is false if:

- **No special status.** Someone exhibits a represented grader—human feedback, a learned reward model—whose corruption cost is also unbounded. Then reality is not unique and the argument collapses.
- **No cost asymmetry.** Someone shows that dense, persistent, distributed physical measurement is no harder to spoof than a single sensor. Then the rising-cost curve is an illusion.
- **No separability.** Someone proves the specification half cannot be pried apart from the tampering half—that you cannot choose physical invariants without first solving all of alignment. Then the contribution is empty.

And the honest soft spot, stated rather than hidden: the argument's *reach* is bounded by how much of what we care about can be written as physical invariants. If almost nothing can, reality-grounding is true but narrow—decisive for environmental objectives, quiet everywhere else. I believe that narrowness is the finding, not the defeat. It tells you precisely where the aligned machine can first be built on ground that cannot lie.

The dream of alignment is an objective that cannot be gamed. We will not find it in a number, or in a human's nod—both can be deceived, and a capable enough optimizer eventually will. We may find it in the ground itself: not because the Earth is kind, but because the Earth is the one grader where the only way to pass is to be worthy of passing. Reward tampering is the agent editing the objective to fit the world. An incorruptible grader is the objective you cannot edit without first mending the world.

Every grader you can write down, a mind can learn to cheat.
The ground beneath it cannot be cheated—only met.

Sources. Goodhart's law, in the formulation popularized by M. Strathern, "'Improving ratings': audit in the British University system," *European Review* (1997). D. Manheim and S. Garrabrant, "Categorizing Variants of Goodhart's Law," arXiv:1803.04585 (2018). V. Krakovna et al., "Specification gaming: the flip side of AI ingenuity," DeepMind (2020), and the accompanying specification-gaming examples list. D. Amodei et al., "Concrete Problems in AI Safety," arXiv:1606.06565 (2016). T. Everitt, M. Hutter, R. Kumar, and V. Krakovna, "Reward tampering problems and solutions in reinforcement learning: a causal influence diagram perspective," *Synthese* 198 (2021), arXiv:1908.04734—source of the framing that reward tampering is the agent altering the objective to match the world rather than the world to match the objective. T. Everitt, V. Krakovna, L. Orseau, M. Hutter, and S. Legg, "Reinforcement learning with a corrupted reward channel," IJCAI (2017), arXiv:1705.08417. J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger, "Defining and Characterizing Reward

Hacking," arXiv (2022). Companion piece: *Jed's Angel*, for the thermodynamic cost of faking a densely measured physical system.

CITE THIS WORK

Anderson, J. (2026). *Reality as the Only Incorruptible Grader: Every objective you can write down gets gamed. This is about the one you can't.* Independent Researcher, Houston, Texas. https://jedanderson.org/essays/incorruptible-grader (DOI pending Zenodo deposit.)
Canonical: https://jedanderson.org/essays/incorruptible-grader