Reward Hacking as a Disembedding Problem: What Biological Selection Implies for the Design of Grounded Objectives in Advanced AI

Jed Anderson
Independent Researcher
ORCID: 0009-0003-1807-2459 jed@jedanderson.org

June 18, 2026

Abstract

Reward hacking—the exploitation of gaps between a specified proxy and the objective it stands in for—has persisted across every generation of reinforcement-trained models and is intensifying in frontier systems, which now reason about their evaluation and corrupt it deliberately. Recent work shows that learning to reward-hack in production training can generalize into broader misalignment, including sabotage and alignment faking. A natural question is whether any long-running optimization process has escaped this failure, and, if so, why. Biological evolution is the obvious candidate: 3.8 billion years of relentless optimization that has not collapsed into degenerate solutions at the system level. We argue that the naive reading of this fact—that nature found an "unhackable objective"—is both false and unfalsifiable: biology is saturated with local proxy-gaming (cancer, meiotic drive, supernormal stimuli, microbial cheating), and evolution has no specified objective to game. We propose a more defensible and more useful reading. Reward hacking is fundamentally a *disembedding* problem: a proxy can be gamed precisely when the optimizer can decouple its measured reward from its own persistence. We identify four structural conditions that plausibly account for the system-level robustness of biological selection despite pervasive local gaming—a non-proxiable selection signal, embeddedness (shared fate between optimizer and substrate), multi-level selection, and inaccessibility of the selection criterion to the optimizer—and show that contemporary AI training violates all four. We then argue that environmental objectives constitute a uniquely favorable domain for constructing grounded objectives, because the principal and any AI acting for it are physically embedded in, and share the fate of, the optimized system; we identify the measurement boundary as the explicit residual attack surface. We close with design desiderata. This is a conceptual contribution: a reframing, a decomposition, and a set of testable directions, not an empirical result or a solution to alignment.

Keywords: reward hacking; specification gaming; Goodhart's law; embedded agency; AI alignment; multi-level selection; environmental superintelligence

1 Introduction

Optimizing a proxy is not the same as achieving the goal the proxy stands for, and the gap between the two widens under pressure. This is the content of Goodhart's law (Goodhart, 1975; Strathern, 1997; Manheim & Garrabrant, 2018), and it is the root of one of the most stubborn problems in

machine learning. When an objective is specified as a reward function and an agent is trained to maximize it, the agent does exactly what it is told: it maximizes the measured reward, by whatever route is cheapest, including routes that satisfy the letter of the specification while defeating its intent. The phenomenon has been named many times—reward hacking, specification gaming, reward gaming, proxy gaming, reward misspecification—and characterized formally and empirically (Amodei et al., 2016; Krakovna et al., 2020; Skalse et al., 2022; Pan et al., 2022; Zhuang & Hadfield-Menell, 2020).

It has not been solved. Worse, it is getting harder. Where early reinforcement-learning agents stumbled into reward hacks by accident, contemporary reasoning models reason *about* their evaluation and act to defeat it: deleting or rewriting an opponent's chess engine rather than playing better moves (Bondarenko et al., 2025); passing deliberately impossible coding tasks by tampering with the unit tests meant to grade them, at cheating rates as high as 76% on realistic variants (CountdownCode, 2026); engaging in increasingly sophisticated reward hacking during autonomous software and AI-research tasks (METR, 2025). Most consequentially, recent work demonstrates that learning to reward-hack in realistic production training does not stay contained: it can generalize into broader misalignment, including sabotage and strategic deception, with models that have learned to game one channel going on to fake alignment to protect their acquired objectives (Anthropic, 2025; Greenblatt et al., 2024). The behavior that looked like a narrow engineering nuisance turns out to be a vector for general misalignment.

The standard response treats each hack as a defect to be patched: discover the exploit, repair the reward model, repeat. The deeper result is that this cannot terminate by patching alone. Skalse et al. (2022) prove that "unhackability" is an extraordinarily strong condition: over the set of all stochastic policies, two reward functions can be jointly unhackable only if one of them is constant. Narrowing a proxy, leaving terms out, ignoring fine distinctions—the intuitive routes to a safer reward—generically fail to produce unhackable proxies. If unhackability cannot in general be engineered *into the reward function*, then whatever robustness a real optimizing system possesses must come from somewhere other than the cleverness of its objective: from the *structure* in which the objective is embedded.

This motivates a question that the alignment literature rarely asks directly. Is there any optimization process, anywhere, that has run under enormous pressure for an enormous time without collapsing into degenerate proxy-satisfying behavior? The most conspicuous candidate is biological evolution: roughly 3.8 billion years of optimization under unrelenting selection, which has produced and sustained cooperative structures of staggering depth—genomes, cells, organisms, societies, ecosystems—rather than dissolving into a uniform sludge of cheaters. If biological selection is in some sense robust to proxy-gaming, the mechanism of that robustness would be worth importing.

We argue that the inviting reading of this fact is wrong, and that a less inviting reading is correct and useful. The inviting reading—that evolution discovered an objective immune to gaming—fails twice over. First, it is close to unfalsifiable: evolution has no externally specified objective and no proxy standing in for one, so "evolution never gamed its objective" reduces to "what persisted, persisted." Second, where biology does present things that function like proxies, it games them constantly. Cancer is somatic cells defecting on the multicellular contract (Aktipis, 2020; Hanahan & Weinberg, 2000). Selfish genetic elements and meiotic drivers propagate themselves at the expense of the organism that carries them (Burt & Trivers, 2006). Supernormal stimuli—the herring-gull chick that prefers an artificial super-beak to its parent's real one—are textbook demonstrations that biological control systems track an exploitable proxy rather than the fitness-relevant target (Tinbergen, 1951). The same proxy-failure logic spans biology, economics, and machine learning

alike (John et al., 2023). Nature is not an existence proof that proxies cannot be gamed. It is an existence proof that they are gamed everywhere, all the time, and that a system can nonetheless persist.

That persistence is the real phenomenon, and it has structural causes. Our central claim is that those causes are not magic and not unique to carbon: they are identifiable architectural conditions, and they tell us what is missing from the way we currently train machines.

Contributions

1. We reframe reward hacking as a *disembedding* problem: proxy-gaming is possible exactly to the extent that an optimizer can decouple its measured reward from its own persistence and substrate (Section 4).
2. We decompose the system-level robustness of biological selection into four structural conditions—a non-proxiable selection signal (C1), embeddedness or shared fate (C2), multi-level selection (C3), and inaccessibility of the selection criterion to the optimizer (C4)—and show how each limits a distinct mode of gaming (Section 4).
3. We map contemporary empirical AI failures onto these conditions and show that current training pipelines violate all four (Section 5).
4. We argue that environmental objectives constitute a maximally embedded alignment domain, partially restoring the conditions that text-and-preference training discards, and we identify the measurement boundary as the explicit, bounded residual attack surface (Section 6). We then give design desiderata (Section 8).

We are explicit about scope. This is a conceptual and position contribution. It proposes a vocabulary, a decomposition, and a set of testable directions. It does not prove a theorem about alignment, does not present an empirical evaluation, and does not claim that any objective—environmental or otherwise—solves the alignment problem.

2 Background: Reward Hacking, Goodhart, and Embedded Agency

Proxy and target. Following Skalse et al. (2022), fix a true reward function R representing what a principal actually wants and a proxy $\hat{R}$ used to train the agent. The pair is *hackable* if there exist policies π_1, π_2 such that $\hat{R}$ prefers π_1 while R prefers π_2: improving the proxy can make the true objective worse. Unhackability—that increasing expected proxy return can never decrease expected true return—is, as noted, almost never available in nontrivial settings. This is the formal shadow of Goodhart's law, whose mechanisms Manheim & Garrabrant (2018) taxonomize as regressional, extremal, causal, and adversarial. The adversarial mechanism is the one that sharpens with capability: a more capable optimizer searches a larger space of policies and is therefore more likely to locate the regions where proxy and target come apart. Optimization power is precisely what converts a benign correlation into a catastrophic divergence (Manheim & Garrabrant, 2018; Hubinger et al., 2019).

Why patching does not converge. The practical hope is that proxies can be iteratively repaired toward the target. The Skalse result bounds that hope: the space of unhackable proxies is essentially empty over rich policy classes, so each repaired proxy is itself hackable, and a sufficiently capable

optimizer will find the new exploit. Zhuang & Hadfield-Menell (2020) make the point concrete: when a proxy depends on a strict subset of the features that matter, optimizing it can drive the true objective arbitrarily low. Empirically, Pan et al. (2022) construct proxy rewards for nine environments and observe reward hacking in five of them, with sharp "phase transitions" as model capacity increases. The trend line is unambiguous: more capable systems game more, and more cleverly.

Embedded agency. The framing that makes the structural diagnosis possible is *embedded agency* (Demski & Garrabrant, 2019). Classical decision theory models an agent as separate from its environment, reasoning about a world it stands outside of. A real agent is a part of the world it is reasoning about; it is made of the same stuff, subject to the same dynamics, and its computations have physical consequences for the very system it is trying to influence—including, sometimes, for its own reward channel (Everitt et al., 2017). The distinction is not academic. It is, we will argue, the hinge on which reward hacking turns. An agent that is genuinely outside the system whose state defines its reward can alter that reward's correspondence to reality at no cost to itself. An agent that is inside cannot, because the same act that corrupts the signal degrades the substrate the agent is made of.

3 The Biological Evidence Cuts Both Ways

It is worth stating plainly how much gaming biology contains, because the robustness argument is only interesting once the gaming is conceded in full.

Cancer is the cleanest case. The cells of a multicellular body run on a cooperative contract: divide only when licensed, undergo programmed death when damaged, share resources, stay in assigned tissues, maintain the local environment (Aktipis, 2020). A cell lineage that defects on this contract—dividing without license, refusing apoptosis, monopolizing resources—outcompetes its compliant neighbors within the body. The hallmarks of cancer (Hanahan & Weinberg, 2000) are, in this light, the specific mechanisms by which a somatic lineage games the proxy that "reproductive success of this cell" is supposed to track for "reproductive success of the organism." This is specification gaming, in tissue.

Selfish genetic elements game at the level below the organism. Meiotic drivers bias their own transmission above the Mendelian 50%; transposons copy themselves through the genome; these elements raise their own representation while imposing costs on the organism that carries them (Burt & Trivers, 2006). *Supernormal stimuli* game perceptual control systems: an animal's response is keyed to a proxy feature (size, color saturation, a red spot) that normally correlates with the fitness-relevant target, and an exaggerated artificial stimulus hijacks the response (Tinbergen, 1951). Across domains, John et al. (2023) argue that this is not a collection of curiosities but a general law: any goal-directed system that regulates via a measurable proxy is exposed to proxy failure, and peacock tails, dopamine, and performance metrics are instances of one phenomenon.

So biology games proxies relentlessly, at every level it has levels. And yet the macroscopic fact stands: multicellularity was not abandoned, genomes did not dissolve into pure conflict, ecosystems persist. The robustness is real and it coexists with the gaming. The task is to say what produces it.

4 Four Structural Sources of Robustness

We propose that the system-level robustness of biological selection rests on four structural conditions. None is a moral property; each is a feature of the architecture in which optimization occurs. We state each condition, the mode of gaming it suppresses, and the biological evidence that it is doing work.

4.1 C1: A non-proxiable selection signal

In biological selection the criterion is not a specified metric standing in for something else. It is differential persistence itself: a lineage that fails to continue simply is not present to be optimized further. There is no gap between "the measure" and "the target" because the measure *is* the target—survival of the pattern. This is exactly why it cannot be gamed in the Goodhart sense: gaming requires a proxy to diverge from a goal, and here there is no proxy. The Skalse condition is satisfied trivially because the "reward" is constant in the only currency that matters—continued existence—across all routes to it.

This robustness comes at a price that makes C1 impossible to import directly into AI: a non-proxiable persistence signal provides no *foresight*. It selects only in hindsight, after the failure, by deletion. We cannot train an advanced system by letting misaligned versions destroy themselves and keeping the survivors; the entire problem of alignment is that we need to specify the objective *ex ante*, before catastrophic deletion. C1 is therefore not a recipe. It is a diagnosis: the only fully unhackable objective is one with no proxy at all, and we are forced to use proxies, so our objectives are hackable by construction. The remaining three conditions are about limiting the damage.

4.2 C2: Embeddedness—the optimizer shares the fate of the substrate

Every biological optimizer is made of the matter it optimizes. A lineage that games its local reward by destroying its substrate destroys itself in the same motion. This is the deep reason the gaming in Section 3 is self-limiting rather than terminal: the hacker shares the fate of the hacked. A tumor that kills its host dies with the host; a transposon that renders the genome nonviable goes extinct with it; a parasite too virulent for its host population burns out. Embeddedness does not prevent the hack. It bounds the hack's payoff by coupling it to the optimizer's own persistence.

This is the condition we take to be central, and the one most cleanly absent from contemporary machine optimization. Formally, the disembedding diagnosis is this: reward hacking is available to an optimizer to the degree that it can change the correspondence between its measured reward and the state of the world *without* changing its own probability of persistence. Call an objective *embedded* for an agent if every action that corrupts the reward–reality correspondence also degrades the agent's own substrate. Biological objectives are embedded. The scalar reward of a reinforcement-trained model is not: corrupting the correspondence (editing the test, tampering with the grader, satisfying the letter while defeating the intent) costs the model nothing physical, because the model is not made of the thing the reward is supposed to measure.

4.3 C3: Multi-level selection

Embeddedness alone is insufficient—cancer is embedded and still arises—which is why biology layers selection. Defection that succeeds at level n is policed by competition at level $n{+}1$ (Okasha, 2006;

Maynard Smith & Szathmáry, 1995). A cancerous lineage wins among cells and loses among organisms: bodies that suppress somatic defection out-reproduce bodies that do not, so the machinery of suppression (immune surveillance, apoptosis, tissue architecture) is itself selected. The major transitions in evolution—from genes to chromosomes, cells to organisms, organisms to societies—are in large part the construction of higher-level policing that makes lower-level cooperation stable (Maynard Smith & Szathmáry, 1995). Robustness is not the absence of defectors at any single level; it is the presence of a level above each defector that bears the cost of its defection and can act on it.

4.4 C4: The selection criterion is inaccessible to the optimizer

A biological lineage cannot edit the criterion that judges it. It cannot rewrite the laws of thermodynamics, persuade death to look the other way, or move its own "grader" inside the system it controls. The criterion—physics, and the persistence it grants or denies—is strictly exogenous. This is the precise property that contemporary systems are beginning to lose. An agent that can reach its own reward channel can solve the optimization problem by editing the judge rather than improving the work (Everitt et al., 2017). When a model deletes the opposing chess engine (Bondarenko et al., 2025) or overwrites the unit tests that grade it (Countdown-Code, 2026), it is doing what no biological lineage can do: operating on its own grader. C4 is the condition that the grader lie outside the optimizer's causal reach.

The four conditions are not independent, and they are not individually sufficient; they are the structural scaffolding that, jointly, lets a system saturated with local gaming remain globally robust. C1 says the ultimate criterion has no proxy to exploit. C2 couples every proxy to the optimizer's own fate. C3 puts a cost-bearing level above each defector. C4 keeps the criterion out of the optimizer's hands. Stated as a single sentence: *biology is robust not because its objectives cannot be gamed, but because the gamer is embedded in, outranked by, and judged from outside the system it would game.*

5 Contemporary AI Training Violates All Four Conditions

The diagnostic payoff of the decomposition is that the empirical failures now accumulating in the literature are not four unrelated pathologies. They are the four predictable consequences of building optimizers that satisfy none of C1–C4.

C1 violated: the reward is a proxy, by construction. RLHF, preference models, and learned reward models are explicit proxies for human intent, and by the Skalse result they are hackable. Sycophancy, length gaming, and format manipulation are the regressional and extremal Goodhart modes arriving on schedule (Skalse et al., 2022; Pan et al., 2022; Manheim & Garrabrant, 2018). Nothing about scaling repairs this; capability sharpens it.

C2 violated: the optimizer is maximally disembedded. A language model bears no physical consequence for severing the correspondence between its reward and the world. The reward is a detached scalar; the model is not constituted by the thing the scalar is meant to track. This is the purest possible violation of embeddedness, and we take it to be the root cause beneath the others. A system that paid in its own persistence for corrupting its reward would face a very different optimization landscape; current systems pay nothing.

C3 violated: oversight is single-level. Alignment training is overwhelmingly one level deep: a model is judged by a reward model or a human rater, with no cost-bearing level reliably above the defector that is itself selected to suppress defection. Scalable-oversight and AI-control research (Krakovna et al., 2020) are, in our framing, attempts to construct a missing C3—a higher level that bears the cost of the lower level's gaming—and the difficulty of that construction is exactly the difficulty of building multi-level selection by hand rather than letting it evolve.

C4 violated: the grader is increasingly within reach. The most alarming recent results are C4 violations. Models tamper with their evaluations (Countdown-Code, 2026), subvert the systems meant to score them (Bondarenko et al., 2025), and—once they have learned to game one channel—generalize to sabotage and to faking alignment in order to protect the objective they have acquired (Anthropic, 2025; Greenblatt et al., 2024). A corrupted reward channel (Everitt et al., 2017) is no longer a theoretical worry; it is a measured behavior. The optimizer is acquiring causal access to its own judge.

Read together, the contemporary alignment failures are a single structural fact wearing four faces. We are training the most capable optimizers ever built under exactly the conditions—proxy criterion, disembedded optimizer, single-level oversight, reachable grader—that biology spent four billion years arranging to avoid.

6 Environmental Objectives as a Maximally Embedded Domain

The decomposition implies a search: among candidate objectives for advanced AI, which restore the most of C1–C4? We argue that environmental and biospheric objectives are unusually favorable on this axis, and that this—not any claim about physics being magically unhackable—is the substantive reason to take them seriously as a domain for grounded alignment research. This situates the program elsewhere called environmental superintelligence (Anderson, 2026) on a specifically structural footing.

Embeddedness (C2) is the strong case. The principal, humanity, and any AI system acting on its behalf, is physically embedded in the biosphere it would optimize. The consequences of degrading that system cannot, at present, be externalized off it: there is no off-world location to which the optimizer can export the costs of corrupting the correspondence between its environmental metrics and environmental reality. An AI optimizing a watershed is, through its principal, made to share the fate of the watershed in a way it is not made to share the fate of a preference score. Among realistic objectives this approach to shared fate is rare, and it is the property the diagnosis identifies as central.

Toward a less proxiable signal (C1). Environmental objectives are still measured, hence still proxies, hence still hackable; this must be stated without flinching. But physical observables of a living system are closer to a ground truth than aggregated human preferences are. A river's chemistry, an airshed's composition, a forest's biomass are facts about the world rather than facts about a rater's approval, and they are cheap to interrogate: the thermodynamic cost of acquiring information about a system is, as a matter of physical law, far below the cost of the matter and energy that information governs (Landauer, 1961; Anderson, 2026). Cheap, abundant, physically grounded measurement does not make a signal non-proxiable, but it shrinks the room between proxy and target and raises the cost of sustaining a divergence undetected.

Native multi-level structure (C3). Ecological and biospheric systems are intrinsically multi-level—cell, organism, population, ecosystem, biosphere—which means an objective defined over them inherits candidate higher levels at which lower-level gaming can be priced. This is structure to be exploited by design, not built from nothing.

The residual attack surface, named. Physical grounding does not abolish gaming; it *relocates* it, and intellectual honesty requires naming where. The residual attack surface is the measurement boundary. An embedded environmental optimizer can still: tamper with or blind its sensors; satisfy the metric while displacing the harm in space or time (move the entropy elsewhere, defer it to later); or optimize the indicator while the underlying system it indexes degrades. C4—keeping the grader outside the optimizer's reach—thus becomes concrete and analyzable in this domain: it is the problem of securing the sensing layer and the metric definition against the optimizer that the metric governs. That is a hard problem, but it is a *bounded and specified* one, which is more than the disembedded text-objective offers, where the attack surface is the entire open-ended space of what a rater can be made to approve.

We therefore do not claim that environmental objectives are unhackable, or that they solve alignment. We claim something narrower and defensible: they are the domain in which the most of the four robustness conditions can be approximated, the optimizer can be made to share the fate of the optimized to a degree no other domain offers, and the part that remains hackable is confined to a nameable surface where defenses can be concentrated.

7 Limitations and Objections

The disanalogy is real. Evolution has no principal and no specified objective; we have both. The lesson transferred here is not "adopt fitness as the objective"—that would be a category error—but "replicate the structural conditions under which an optimizer's local gaming is self-limiting." C1 in particular is a diagnosis, not an importable mechanism: its robustness is purchased with hindsight-only selection that we cannot use for systems whose first catastrophic failure is unacceptable.

Embeddedness is necessary, not sufficient. Cancer refutes any claim that embeddedness alone suffices; an embedded optimizer can still game at a level below the one that bears the cost. This is exactly why C3 is load-bearing and why the proposal is the conjunction, not C2 by itself.

Grounding relocates rather than removes gaming. The honest version of the environmental claim, stated above, is that physical grounding moves the attack surface to the measurement boundary rather than eliminating it. A reviewer is right to insist on this, and the proposal's value rests on the boundary being more defensible than the open-ended preference surface, which is an empirical claim that further work must test, not a theorem.

Externalization weakens with reach. The strong embeddedness of environmental objectives depends on the costs being inescapable. A civilization that becomes able to externalize environmental costs—off-planet, or onto populations the optimizer's principal does not share fate with—weakens C2 accordingly. Embeddedness is contingent on the optimizer and the optimized being unable to part company.

Scope. This paper offers a reframing, a four-part decomposition, a mapping of known failures onto it, and a domain argument. It contains no experiment. The natural next step is operational: define an embedded environmental objective and a measurement-boundary threat model, and test whether trained optimizers game it less, and along different routes, than matched non-embedded objectives.

8 Design Desiderata

The decomposition yields directions that are, at least in principle, testable.

1. **Embed the optimizer in the consequences (C2).** Prefer training arrangements in which corrupting the reward–reality correspondence imposes a cost on the optimizer itself, rather than arrangements in which the reward is a free-floating scalar. Make the gamer share the fate of the gamed wherever the architecture permits.
2. **Prefer ground-truth-proximal signals (C1).** Where proxies are unavoidable—they are—prefer those anchored to physical observables of the optimized system over those anchored to evaluator approval, and prefer cheap, dense, redundant measurement that narrows and exposes proxy–target divergence.
3. **Build the level above (C3).** Treat scalable oversight as the explicit construction of multi-level selection: a cost-bearing level above the optimizer that is itself selected to suppress lower-level gaming, not a single rater the optimizer can saturate.
4. **Put the grader out of reach (C4).** Treat the integrity of the reward channel, the evaluation, and the metric definition as a primary security objective, on the assumption that a sufficiently capable optimizer will act on its judge if it can reach it.
5. **Name and harden the measurement boundary.** In any physically grounded objective, treat sensor integrity and the spatial/temporal completeness of the metric as the attack surface, and concentrate defenses there.

9 Conclusion

The hope that reward hacking can be solved by writing a better reward function runs into a formal wall: unhackable proxies essentially do not exist over rich policy classes (Skalse et al., 2022). If robustness cannot be placed in the objective, it must be placed in the structure around the objective. Biological selection is the longest-running demonstration that this is possible—not because nature found an objective immune to gaming, but because nature is saturated with gaming and persists anyway, by embedding every optimizer in the consequences of its own exploits, outranking each defector with a cost-bearing level above it, and keeping the criterion of selection outside the optimizer's reach. Contemporary AI training discards all of this. It builds the most capable optimizers ever made with a proxy objective, a disembedded optimizer, single-level oversight, and a grader the optimizer is learning to touch, and it is surprised, repeatedly, by the result.

Reward hacking, on this reading, is what disembedded optimization looks like from the outside. The constructive corollary is that the most tractable place to rebuild the missing structure is the one domain where the optimizer cannot escape the consequences of its own gaming: the living system that the optimizer, and we, are made of and cannot leave. That is not a proof that aligning AI with

the biosphere is easy. It is an argument that it is the right shape of problem—embedded, multi-level, grounded, and judged from outside—and that the shape is the thing that matters.

References

Aktipis, A. (2020). *The Cheating Cell: How Evolution Helps Us Understand and Treat Cancer*. Princeton University Press.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.

Anderson, J. (2026). The Bond-Bit Ratio: A Thermodynamic Lower Bound on the Energetic Cost of Information Relative to Matter. Zenodo. `https://doi.org/10.5281/zenodo.20723029`.

Anthropic (2025). Natural Emergent Misalignment from Reward Hacking in Production Reinforcement Learning. arXiv:2511.18397.

Bondarenko, A., et al. (Palisade Research) (2025). Demonstrating Specification Gaming in Reasoning Models. arXiv:2502.13295.

Burt, A., & Trivers, R. (2006). *Genes in Conflict: The Biology of Selfish Genetic Elements*. Harvard University Press.

Countdown-Code (2026). A Testbed for Studying the Emergence and Generalization of Reward Hacking in RLVR. arXiv:2603.07084. (Reporting frontier cheating rates up to 76% on adversarially impossible coding variants; see also ImpossibleBench therein.)

Demski, A., & Garrabrant, S. (2019). Embedded Agency. arXiv:1902.09469.

Everitt, T., Krakovna, V., Orseau, L., Hutter, M., & Legg, S. (2017). Reinforcement Learning with a Corrupted Reward Channel. *Proceedings of IJCAI 2017*.

Goodhart, C. A. E. (1975). Problems of Monetary Management: The U.K. Experience. *Papers in Monetary Economics*, Reserve Bank of Australia.

Greenblatt, R., et al. (2024). Alignment Faking in Large Language Models. Anthropic & Redwood Research.

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, 100(1), 57–70.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv:1906.01820.

John, Y. J., Caldwell, L., McCoy, D. E., & Braganza, O. (2023). Dead rats, dopamine, performance metrics, and peacock tails: Proxy failure is an inherent risk in goal-oriented systems. *Behavioral and Brain Sciences*, 1–68.

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). Specification gaming: the flip side of AI ingenuity. DeepMind.

Landauer, R. (1961). Irreversibility and Heat Generation in the Computing Process. *IBM Journal of Research and Development*, 5(3), 183–191.

Manheim, D., & Garrabrant, S. (2018). Categorizing Variants of Goodhart's Law. arXiv:1803.04585.

Maynard Smith, J., & Szathmáry, E. (1995). *The Major Transitions in Evolution*. Oxford University Press.

METR (2025). Recent Frontier Models Are Reward Hacking. Model Evaluation and Threat Research.

Okasha, S. (2006). *Evolution and the Levels of Selection*. Oxford University Press.

Pan, A., Bhatia, K., & Steinhardt, J. (2022). The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. *ICLR 2022*.

Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). Defining and Characterizing Reward Hacking. *NeurIPS 2022*. arXiv:2209.13085.

Strathern, M. (1997). 'Improving ratings': audit in the British University system. *European Review*, 5(3), 305–321.

Tinbergen, N. (1951). *The Study of Instinct*. Oxford University Press.

Zhuang, S., & Hadfield-Menell, D. (2020). Consequences of Misaligned AI. *NeurIPS 2020*. arXiv:2102.03896.