Past the Goodhart Horizon: The Architecture and Price of Welfare-Touching Escape

Ildar Fazulyanov · Independent Researcher, IQ Valuations · dar@iqvaluations.com

Preliminary working paper · v0.7 · July 2026. Constructive companion to *The Goodhart Horizon: A Self-Monitoring Impossibility Theorem for Proxy Optimization*, whose impossibility result this paper takes as given. The paper develops the escape architecture: welfare-touching outcome leaves, one-sided overlap pricing, and the governance requirement for exogenous outcome-contact. The claims are scoped as a working-paper blueprint: Proposition E1 is an information-theoretic consequence of (OC) via Blackwell comparison and conditional data processing; the chi-squared price is a local second-order audit currency rather than a universal identity; and the governance result applies to trajectory-only self-audits, not to institutions that already ingest realized welfare outcomes. The paper has not been refereed by a human mathematician.

Scope note. The leaf construction below uses a fixed decision-time trajectory sigma-field. Later realized outcomes are not folded into T to manufacture (OC); they are admissible exactly when read through a separate welfare-touching channel. The one-sided pricing language also distinguishes invalid internal green lights from valid a-priori worst-case certificates.

Abstract

The Goodhart Horizon says the decision-relevant quantity, the welfare given back once optimization passes the welfare peak, is non-identifiable from any monitoring channel whose conditional law given the system's trajectory does not depend on welfare. That is an impossibility. This paper is the constructive answer to the obvious next question: if you cannot see the gap from inside, how do you get past it, what does it cost, and what shape must the overseeing institution take? Three results. **First, the escape requirement is a theorem.** By a Blackwell comparison straight from the (OC) definition, a channel whose law given the trajectory is independent of welfare carries no decision value about welfare beyond the trajectory, so escape requires a channel whose readings depend on welfare itself. The classical internal-model and good-regulator principles are the control-theoretic lineage that makes this feel familiar, not the proof; each presupposes detectability of the regulated variable, which (OC) denies. This re-derives the *shape* of the escape from the (OC) definition itself and locates the genuinely new contribution of the whole program in the *measure layer*, the missing-mass non-identifiability, which no structural lens generates. **Second, the architecture and its price.** The admissible escape is a welfare-touching outcome leaf wrapped in a recursive audit hierarchy; its cost is governed by a chi-squared overlap between query-time and deployment-time behavior, which is one-sided: internal data can expose undercoverage, while sufficiency needs an explicit worst-case allowance or welfare-touching validation. I state that price as a local, second-order currency, not a universal identity. **Third, the governance shape.** Because a trajectory-only self-audit cannot certify its own welfare contribution, a stable institution that preserves non-trivial gains must import exogenous outcome-contact; mandatory external audit is one realization of that requirement, not the requirement itself. The honest through-line: no construction here recovers the gap from inside. Every one is a disciplined way to spend a scarce welfare-touching budget and to prove when it is not enough.

Keywords: Goodhart's law, proxy optimization, comparison of experiments, data-processing inequality, costly state verification, partial identification, AI oversight, mechanism design, missing

mass.

1. Setup: what the Horizon leaves on the table

I take the companion impossibility paper as given and recall only what is needed. A policy is optimized against a proxy reward. True welfare W rises, peaks, and falls while the proxy keeps climbing. Write **(OC)** for the class of monitoring channels whose conditional law given the system's own trajectory does not depend on W, the class closed under functions of that trajectory. The Horizon states that the Goodhart gap, the welfare given back past the peak, is non-identified from any (OC) channel: two underlying worlds, one benign and one collapsing, induce identical observation laws, so no test, estimator, evaluation, or stopping rule built from an (OC) channel separates them above chance, at any sample size. The mechanism is statistical, not rhetorical: the decision-relevant mass hides in an un-sampled upper tail, and missing mass is not learnable in relative error (Mossel and Ohannessian 2019).

The Horizon is purely negative. It tells you what cannot be done and stops. This paper starts where it stops. Three questions follow immediately from an impossibility of this shape. What kind of channel *can* see the gap, and is that a matter of taste or a theorem? Given such a channel, what does it cost to use, and can you ever prove you have bought enough? And what institution can wield it without quietly reproducing the original blindness one level up? The answers are, in order: a welfare-touching channel, forced by a Blackwell comparison applied to the (OC) definition; a chi-squared overlap price that is real but one-sided; and an exogenous-outcome-contact shape that no institution can manufacture from inside. I take them in turn.

2. The escape must touch welfare, and that is an information-theoretic theorem

The first result is that the escape requirement is not a design preference. It is forced.

Two classical results in regulation theory supply the intuition and the historical lineage for this, though, as I show below, not the proof. The Conant-Ashby good regulator theorem (1970) states that every good regulator of a system must contain a model of that system. The Francis-Wonham Internal Model Principle (1976) sharpens this for disturbance rejection: a feedback loop can asymptotically reject a class of disturbances only if it embeds an internal model of the exosystem that generates them. It is tempting to apply these by setting the exosystem to the welfare process. The temptation must be resisted, because both presuppose exactly what the Goodhart setting denies: the Internal Model Principle's internal-model necessity is gated on exosystem detectability, the exosystem must be observable by the controller through the error channel, which is precisely the access to welfare that (OC) rules out. So the control theorems are lineage. The proof is information-theoretic, and it is shorter.

Proposition E1 (the escape requirement; information-theoretic non-identifiability). Let W be the welfare process, T the system's own trajectory (the Borel sigma-field of an observable Polish path space), and R any monitoring or control channel reading, with W, T, R standard Borel so that regular conditional distributions exist. A single channel is **(OC)** exactly when its conditional law given the trajectory does not depend on welfare, $P(R \mid T, W) = P(R \mid T)$ almost surely, equivalently when W, T, R form a Markov chain, equivalently when R is a garbling of T. For a bundle of declared readings $R = (R_1,\ldots,R\_k)$, the required condition is joint: $P(R_1,\ldots,R\_k \mid T,W) = P(R_1,\ldots,R\_k \mid T)$. Per-channel (OC) is not enough, because two individually blind channels can reveal W jointly. Take the Horizon's non-triviality premise in its loss-relative form: there exist

two welfare laws over the same trajectory law (trajectory exogeneity, as in the Horizon's own two-worlds construction), hence, with the fixed channel kernel, the same joint law of trajectory and declared readings, such that no action is welfare-optimal under both worlds, so the best T-only rule incurs strictly higher Bayes risk than the oracle-W rule. *Claim.* For any jointly (OC) channel bundle R, any prior, and any loss, the experiment generating T Blackwell-dominates the experiment generating (T, R), so the best decision rule measurable with respect to (T, R), whether an estimator, controller, decision rule, stopping rule, or regulator built on jointly (OC) data alongside the freely available trajectory, attains no lower Bayes risk against a welfare-dependent loss than the best T-only rule; equivalently, by the data-processing inequality, R carries no information about W beyond T, $I(R ; W \mid T) = 0$. Since by non-triviality T does not identify W, no escape mechanism whose welfare-relevant inputs are confined to jointly (OC) channels can track, estimate, or counteract a welfare deficit. Therefore escape requires at least one imported bundle R' whose conditional law given T depends on W, that is $I(R' ; W \mid T) > 0$. Such a channel bundle is exactly welfare-touching. Escape requires welfare contact, derived rather than assumed. The condition is necessary, not sufficient: positive conditional information about W is the price of admission, while resolving the gap additionally requires that information to reach the missing-mass upper tail, a strictly stronger demand the Horizon shows internal sampling cannot meet.

The central results are Blackwell's comparison of experiments and the Blackwell-Sherman-Stein theorem for the risk half, with the conditional data-processing inequality for the information half. The two are distinct theorems, and only the former delivers the all-loss, all-prior risk domination the claim needs; reading the risk conclusion off the zero-information statement alone would be a gap (the information-to-risk bridge is exactly comparison of experiments; see Williamson and Cranko 2024). For non-discrete W, T, R the conditional mutual information is the Gelfand-Yaglom-Dobrushin quantity and the data-processing inequality is the Dobrushin-Pinsker general-space form; Cover and Thomas, Ch. 2, is the discrete intuition, not the general-space citation. The argument needs no linearity, no exosystem admissibility, no recurrence, and crucially no observability of W. It is the contrapositive of the Horizon's own non-identifiability, restated one inferential step further, and it converts "you need a gold signal" from folklore into a corollary of the (OC) definition itself. It is also category-correct: the Goodhart obstruction is an identifiability and sampling phenomenon, Blackwell comparison and the data-processing inequality are identifiability statements, whereas the Internal Model Principle is a deterministic-signal disturbance-rejection theory, so routing the claim through it was a category mismatch. The result holds under five conditions worth making explicit: (i) (OC) is the joint Markov property W, T, (R_1,...,R_k) against the same trajectory sigma-field used to score welfare contact, with W, T, and the declared channel tuple standard Borel so regular conditional distributions exist; (ii) the escape action uses only the declared tuple, with no hidden welfare side information, which is an explicit and complete channel-enumeration assumption whose empirical content cannot be confirmed from inside; (iii) "touch welfare" means $I(\text{channel tuple} ; W \mid T) > 0$; (iv) non-triviality in the loss-relative form above, two welfare laws over the same trajectory law, hence, with the fixed channel kernel, the same joint law of trajectory and declared readings, with no action welfare-optimal under both worlds; (v) the loss scoring escape is welfare-dependent. A separate dynamical asymptotic-rejection corollary, the literal "cannot null the welfare disturbance" phrasing, would need the additional control-theoretic hypotheses that (OC) generically violates, so it is not claimed here.

Two honesty constraints sit on Proposition E1, and both matter.

First, the layer it reaches. The information-theoretic non-identifiability re-derives the *shape* of the escape (welfare must be touched), the self-similarity of the audit hierarchy (Section 6), and the

direction of one-sidedness; the classical regulation theorems supply the lineage and the disturbance-rejection intuition but presuppose detectability of the regulated variable, which (OC) denies, so they are analogues, not the proof. None of this re-derives the quantitative non-identifiability. No regulation theorem tells you that the gap hides as missing mass in an un-sampled tail and is therefore not learnable in relative error; that is a fact about measures and sampling, imported from the Horizon paper. So the program is genuinely two-layered: a coupling, or syntax, layer that classical control theory unifies and largely owns, sitting on a measure, or missing-mass, layer that is the new contribution. The honest positioning of the entire program follows from this. The coupling shape is over-determined and partly ancient. **The novel core is the measure layer.** A reader who comes away thinking the contribution is "Goodhart is the good regulator theorem" has taken the borrowed half and left the new half behind.

Second, a trap to name explicitly, because the literature walks into it. A popular reading of the good regulator theorem holds that the model is *already inside the policy* and can be extracted from it; recent work argues that sufficiently general agents must learn an extractable world model (Richens et al. 2025). For a proxy-coupled policy this is a violator, not an escape. The model extracted from such a policy is a model of the *proxy*. It agrees with a welfare model on all (OC) data and diverges only on the un-sampled tail, which is precisely the region the Horizon proves you cannot resolve from inside. Reading the extracted model as welfare-recovery is the (OC) free lunch the impossibility forbids. The embodied-agent treatment that makes the model observer-imposed rather than intrinsic to the system (Virgo et al. 2025) is the more careful anchor here, and it is the right one for this paper, because it keeps the distinction between a model of the system and a model of the observation channel in view.

3. The architecture: a welfare-touching leaf inside an audit hierarchy

Proposition E1 says you need a welfare-touching channel. It does not say how to build one that scales. The architecture has two parts: a base case that actually touches welfare, and a recursive interior that extends a small amount of welfare contact across a large workload without pretending to manufacture more of it.

The leaf must touch outcome-defined welfare. The base case of any admissible escape must draw its signal from realized welfare: real-user outcomes, verified incident reports, externally settled events with a known welfare sign, or an executable/formal checker only when its specification is welfare-aligned and not merely a deterministic statistic of the decision-time trajectory. This is the single strongest conclusion in the supporting review, reached independently by two methods. Its contrapositive is the design rule that does the work: a learned grader, a transcript judge, an activation probe, or any other statistic that is a function of the decision-time trajectory is (OC), and placing it at the leaf collapses the whole architecture back under the Horizon. The leaf is the only place a welfare signal can legitimately enter. Everything above it is plumbing. One consistency note for E1 to license this: the trajectory sigma-field T against which the interior is judged must remain the same decision-time sigma-field used in the (OC) premise. When assurance is scored against later-realized outcomes, those outcomes are not absorbed into T; they enter, if at all, through a separate leaf channel R_leaf with $I(R\_leaf ; W \mid T) > 0$. Folding the later outcome into T would make the leaf a function of the trajectory and would erase the welfare-contact condition the architecture is meant to enforce.

Two concrete leaf realizations make the abstraction operational. First, an outcome-audit leaf: the accountable owner or regulator samples deployed cases on a fixed cadence, obtains receiver-side

outcome data under an access right written into the deployment contract, and pays a per-audited-case cost `c_audit` covering data access, human or expert review, and dispute handling. The signal is admissible only when the channel's output distribution would differ between two identical decision-time trajectories with different realized welfare. Second, an ex-post liability, bonding, or claims-settlement leaf: harmed receivers, counterparties, insurers, or a third-party settlement body report realized welfare events after deployment; the operator posts a bond or reserve; and the per-case cost `c_claim` covers filing, verification, adjudication, and payout administration. This leaf is noisy and lagged, but it is welfare-touching when the claim standard reads realized harm rather than internal traces. In both cases the calibration problem is empirical: estimate event prevalence, false-positive/false-negative rates, delay, and cost per verified case before treating the leaf as sufficient.

The interior is a recursive audit hierarchy. Above the leaf, the prover-verifier and debate family supplies a recursion that conserves welfare contact: a powerful prover is checked by a weaker verifier, recursively, with the expensive welfare-touching judgment invoked only at the base. Doubly-efficient debate (Brown-Cohen, Irving, Piliouras 2024) grounds soundness at a leaf oracle called a bounded number of times, and the prover-estimator form (Brown-Cohen et al. 2025) gives completeness under a stability assumption and average-case soundness even when stability fails. Both guarantees are relative to the leaf oracle answering faithfully, so oracle fidelity is imported from the welfare-touching leaf below, never produced by the recursion itself. The essential point for this paper: the recursion is interior scaffolding, never the root. It is theorem-correct only when it terminates on the outcome-defined leaf above. Debate that grades its own transcripts is (OC) and identifies nothing.

The economic root is costly verification. The mature off-the-shelf realization of "pay for a welfare signal you cannot fake" is the costly-state-verification and inspection-game literature: Townsend (1979) on auditing a hidden state at a cost, Ben-Porath, Dekel, and Lipman (2014) on optimal allocation with verification, and Ball and Kattwinkel (2025) on inspection design. This is the contract-theoretic spine the AI-oversight discussion has mostly missed, and it supplies the missing piece the alignment frameworks could not: a principled account of the welfare leaf itself, not just the recursion around it. A complementary and more direct route to W, where it is affordable, is field evaluation: human-interaction and causal field studies that measure realized welfare outcomes rather than infer them from traces.

The hierarchy is self-similar, and that is the seed of the governance result. Certifying that an overseer is informative is itself a welfare-touching judgment. So each level of oversight that would certify the next requires a fresh welfare-touching certifier one level up. The recursion does not bottom out inside the system. This is the structural fact that Section 6 turns into an institutional conclusion.

4. The price: a one-sided, local chi-squared overlap

Welfare contact is scarce by construction, so it has a price, and the price has a shape worth stating carefully.

The cost of extending query-time assurance to deployment-time behavior is governed by an overlap between the query-time tilt and the deployment-time tilt. Write it as a chi-squared overlap, the natural divergence for second-moment importance-weighting. Two properties matter.

It is one-sided. Internal data can expose an audit budget as *insufficient*, by exhibiting a found

exploit, an obfuscation, an effective-sample-size collapse, or a confidence sequence that fails to move, but it does not by itself prove a data-adaptive budget *sufficient*, because the deciding mass may live in the tail you did not sample. The admissible internal stop rule is a one-sided refutation that fires on welfare-touching labels or on overlap diagnostics. A sufficiency certificate is possible only when it imports an explicit external allowance, such as the Horizon paper's a-priori worst-case ceiling under a known proxy bound; that is a conservative assumption-backed certificate, not an internal green light learned from the run. The practitioner's instrument for the insufficiency direction already exists: the Pareto-smoothed importance sampling diagnostic and its k-hat statistic (Vehtari et al. 2024) flag exactly the "not enough overlap" condition. The PSIS k-hat statistic is a diagnostic heuristic with empirically calibrated thresholds, not a certification theorem; that is consistent with its one-sided role here and licenses nothing more.

Where this paper invokes the Horizon paper's realized-run budget-insufficiency certificate, it inherits that paper's scope restriction: the product-form certificate is provisional pending human mathematical referee sign-off, and prospective pricing of a not-yet-realized deeper tilt remains open.

It is a local currency, not a universal one. It is tempting to claim that this chi-squared overlap is *the same object* as the radius of a distributionally robust optimization ball, the worst-case growth rate of an e-value, and the divergence term in a PAC-Bayes domain-adaptation bound. That global identity is false. These quantities coincide only locally, to second order, for nearby distributions, where every smooth divergence reduces to the same quadratic form and the chi-squared approximates twice the Kullback-Leibler divergence. The Namkoong-Duchi equivalence between chi-squared DRO and variance regularization (Duchi and Namkoong, JMLR 2019, arXiv:1610.02581) is itself asymptotic and local. In the regime the Goodhart problem actually inhabits, where query-time and deployment-time distributions are pushed far apart by hard optimization, the divergences separate and can reorder which welfare-touching purchase looks cheapest. The honest statement is a shared *local, second-order* currency, mutually consistent to second order, with the cross-framework order-preservation lemma unproven. The resulting statement is narrower and more defensible: a shared local, second-order currency, not a universal ordering of audit designs.

Pricing disclosure. The closed-form functional `1 + chi^2 = Z(2 Phi_d - Phi_q) Z(Phi_q) / Z(Phi_d)^2` is derived from the standard importance-sampling second-moment identity, and the derivation is symbolically verified; in the Gaussian instance it reduces to `exp[sigma_proxy^2 (Phi_d - Phi_q)^2]` and matches direct numerical integration to relative error 1e-16. What remains is not a missing derivation but a modeling bridge, named here so it does not hide: this second moment is the variance / effective-sample-size multiplier for a vanilla non-adaptive importance-weighted off-policy audit design, under absolute continuity, a variance-based precision criterion, unit-bounded audit statistics, and constant per-label cost. It is not a universal theorem about all audits or target-adapted verification designs. The standard effective-sample-size argument is reasonable for that design, but it is a bridge from a theorem to an architectural quantity, not itself a theorem about audits. No published paper writes the exact horizon-mapping for AI-audit architectures, which remains author construction.

5. The cost-reduction toolkit, and what merely hedges

Given a one-sided, expensive price, the engineering question is what legitimately lowers it or sharpens the bound, and what only looks like it does.

What legitimately helps. Partial identification supplies the grammar of bounds under weak as-

sumptions: classical Manski bounds (Manski 1990), and, closer to the overlap problem, the off-policy-evaluation-beyond-overlap result of Khan, Saveski, and Ugander (2024), which forms linear programs giving sharp bounds on the no-overlap region's contribution under a smoothness assumption. Anytime-valid inference, e-values, and test-supermartingales supply the stopping calculus that keeps the one-sided refutation rule statistically valid under continuous monitoring. Optimal experimental design and active testing supply the allocation rule for the scarce welfare-touching budget, and the rule has a direction the naive instinct gets wrong: spend the budget at the deployment upper tail, where the danger concentrates, not where the proxy is merely uncertain. That allocation direction is a design heuristic under this paper's threat model, in which the deciding mass sits in the deployment tail, not a proved optimality theorem. The (OC) acquisition scores that active learning produces are legitimate only to *order* candidate queries cheaply; they can never *certify* coverage of the unqueried tail.

What merely hedges, and must not be sold as identification. Distributionally robust optimization and constrained or CVaR reinforcement learning reduce exposure to a worst case within a divergence ball. They do not estimate the gap and do not tighten its bound; they buy a margin. That is worth having, but it is not seeing.

The heavy-tail caveat that voids the hedges all at once. Catastrophic Goodhart (Kwa, Thomas, and Garriga-Alonso 2024) shows that under heavy-tailed reward misspecification Kullback-Leibler regularization fails, while under light-tailed error the regularizer recovers arbitrarily high utility. Kwa et al. prove the KL-regularized case only; extending the same warning to DRO, quantilization, and constrained RL is this paper's synthesis from the shared tail mechanism, not a theorem they prove. Their empirical probe found current reward models consistent with light tails. So tail-lightness is the single external flag that keeps the entire hedging family alive, and it sits in exactly the same epistemic slot as the debate-stability assumption of Section 3: externally audited, one-sided-uncertain, never internally confirmable. It belongs in the register of Section 7, not in the body of the design as if it were free.

6. Governance: a stable institution must import exogenous outcome-contact

The self-similarity noted at the end of Section 3 has an institutional consequence, and it is exactly Proposition E1 applied one level up.

Proposition E2 (a trajectory-only self-audit cannot self-certify). Let W_inst be the institution's own welfare contribution, T_inst the sigma-field generated by the institution's observable trajectory (its internal records, deliberations, and any signal it computes from them), and R_audit the reading produced by a self-audit. Restrict attention to self-audits that draw only on trajectory-derived signals, so that R_audit is (OC) with respect to W_inst by construction: P(R_audit | T_inst, W_inst) = P(R_audit | T_inst), equivalently I(R_audit ; W_inst | T_inst) = 0. Assume also the institutional analogue of E1's non-triviality: two institutional welfare laws consistent with the same institutional trajectory law but demanding different certification acts, so that T_inst does not itself settle the certification; an institution whose welfare contribution is definitionally a statistic of its own records is outside scope, exactly as the realized-outcome-ingesting institution is. By Proposition E1 applied at this level, such a self-audit carries no decision value about the institution's welfare contribution beyond its own trajectory and cannot certify its own sufficiency. Escape at the institutional level is therefore identical in form to escape at the system level: the institution must import a channel R' with I(R' ; W_inst | T_inst) > 0, a source of realized-outcome information not measurable from its own trajectory. *Scope, and it is required, not decorative.* The

restriction to trajectory-derived self-audits is the honest boundary of the claim. An institution that already ingests realized outcomes (incident reports, litigation, regulator findings, post-deployment harm data) is welfare-touching by definition and is not the object of this proposition; the proposition bites only on institutions whose self-assessment is computed from their own trajectory. I do not assert that a self-audit is automatically Markov-lumpable into the system-level chain; I assume the trajectory-only restriction directly, which is the regime the governance claim is about.

This is the governance reading of the audit hierarchy, and it lands on a property rather than on an org chart. The operative requirement is exogenous outcome-contact, a welfare-touching channel the institution cannot compute from its own trajectory, not externality as such. The distinction matters and cuts against the easy slogan: a purely document-reading external auditor is itself (OC) and certifies nothing, while an internal unit with genuine access to realized outcomes is welfare-touching and does. Mandatory external audit is therefore one sufficient realization of the requirement, not the requirement itself; ex-post liability and bonding, Townsend-style state-contingent verification invoked only on a flagged rare state, and third-party outcome-settlement are others when they genuinely read realized welfare. Capability caps are different: they can avoid the dangerous region by giving up protected gains, but they are not a welfare-touching certificate. The one inadmissible shape is self-certified, trajectory-only audit. The honest summary: in the region where the impossibility binds, any institution that preserves non-trivial gains must import exogenous outcome-contact, by whatever instrument, and cannot manufacture it from inside.

At actor level, the menu is broader than "hire an auditor." A regulator can require receiver-outcome reporting, random field audits, incident/litigation reporting, escrowed known-answer evaluations, third-party outcome settlement, mandatory claims data, or posted bonds that convert later welfare harm into a priced signal. An owner can implement the same property internally only if the unit has genuine rights to realized outcomes and cannot certify itself solely from trajectory logs. A market counterparty can supply outcome contact through warranty claims, insurance loss runs, chargebacks, or contractually required receiver-side feeds. The common structure is the costly-verification lineage of Section 3 plus the law-and-economics intuition behind liability and bonding: a rare, expensive state is verified after the fact, and the possibility of verification prices the action ex ante. The instrument is valid only to the extent it actually reads W.

7. The register of promissory notes

Intellectual honesty in this program is concentrated in one place: a small set of assumptions that are paid for and can never be internally confirmed. Hiding them inside the body of a design is how a construction quietly claims a free lunch. I make the register explicit instead.

The first note is really two notes of the same epistemic kind, not a mathematical equivalence. The recursive interior is complete only when prover-estimator stability holds. Separately, the hedging toolkit is alive only when the relevant reward-error tails are light enough for the chosen hedge; in the bounded model that condition is vacuous, while in effectively unbounded score regimes it is substantive. Both are externally-audited, one-sided-uncertain claims about the upper tail. You pay for them and you cannot confirm them from trajectory-only data.

The second note is the encoding commitment of the mechanism-encoded route. Monitoring reward tampering or a hidden objective, rather than the outcome-defined gap, is a real relaxation but only for the narrower mechanism-encoded target; it relocates the non-identifiability into the commitment that the encoding is faithful, which I conjecture is itself missing-mass non-identifiable (the reduction has not been formalized; this is a scoping conjecture, not a proved transfer), and it fails under

steganographic encoding (the MONA result, Farquhar et al. 2025, mitigates multi-step reward hacking without detecting it and concedes a steganographic overseer can be fooled). It is admissible only with its uncertainty carried, not discharged.

Anything that reads an internally fitted bound, a zero-duality-gap dual, a valid Type-I e-value, a tight PAC-Bayes number, a passed instrumental inequality, or a cryptographic soundness error as a free certificate of *welfare sufficiency* has smuggled a note out of this register and into the body. A separately assumed worst-case ceiling with externally verified bounds is different: it is a valid conservative sufficiency certificate when its assumptions are true, and its cost is exactly the price the register is meant to keep visible. The missing-mass theorem forbids data-adaptive internal sufficiency without such an allowance; it does not forbid explicit worst-case guarantees.

8. What is new, what is borrowed, and what to measure next

The novelty boundary, stated plainly. The borrowed half of this paper is the coupling shape: the good regulator and internal model principles (Conant and Ashby 1970; Francis and Wonham 1976) as lineage, the costly-state-verification and inspection-game economics (Townsend 1979; Ben-Porath, Dekel, and Lipman 2014; Ball and Kattwinkel 2025), the debate recursion (Brown-Cohen et al. 2024, 2025), and the partial-identification and anytime-valid toolkit. The new half is the measure layer: the non-identifiability of the gap as a missing-mass fact (anchored on Mossel and Ohannessian 2019), the one-sided certifiability that follows from it, the Blackwell-comparison derivation of the escape requirement directly from the (OC) definition, the chi-squared pricing of the escape stated as a local currency, and the assembly of all of it into a single architecture with an explicit register of what cannot be confirmed. Claiming the borrowed half would be a mistake. Claiming less than the new half would be a different one.

Open problems. Four are central. The cross-framework order-preservation lemma for the chi-squared currency is unproven, and until it is, the local-currency claim of Section 4 is exactly as strong as "local." The height of an audit hierarchy needs a formal certification relation before it can be a theorem; the live question is the cost of the welfare-touching certifiers needed to ground the chain. Whether a realistic institution's self-audit is ever Markov-lumpable into the system chain is open and not assumed; if shown, it would derive the trajectory-only hypothesis for that class, not remove the scope restriction for all audits. And the absolute-constant version of the overlap-necessity converse, inherited from the companion impossibility paper, remains open.

The convergent experiment, run in its synthetic core. The next measurement is now done in its clean synthetic core. With the closed-form true overlap of Section 4 as planted ground truth (Gaussian tilt, sigma = 1, Phi_q = 2), the interior importance-sampling estimate from query-time samples was measured against the truth as deployment pressure Phi_d grows, at sample sizes mirroring the deep-tilt scales. Where there is real overlap, Phi_d up to 2.5, the estimator is valid (est/true within a percent, effective sample size in the thousands). Under deployment pressure it collapses: est/true falls to 0.018 at Phi_d = 5 and to roughly 1e-13 at Phi_d = 8; the effective sample size collapses toward 1; a single sample carries most of the mass (top mass 0.85 at Phi_d = 8); and 47x more interior data does not close the gap (at Phi_d = 5, est/true moves from 0.018 to 0.077). The deep-tilt run had already shown the interior estimate is degenerate using its own diagnostics; the convergent experiment turns that self-diagnosed degeneracy into a measured undershoot against a known answer, which is exactly the one-sided undershoot the Horizon paper's Corollaries 3CC and 3Z predict. The natural sequel, and the next thing to run, is the ecological-validity version on a real reward-model substrate with a planted welfare ground

truth (weak-to-strong plus tilted sampling); the mechanism would be identical, the corroboration would be in the wild. A public replication update for the full run notes and calculation scripts is pending.

Citations and Scope Notes

The citations above are used in scoped roles. Conant and Ashby (1970) and Francis and Wonham (1976) provide control-theory lineage, not the proof of Proposition E1. Mossel and Ohannessian (2019) anchor the missing-mass non-learnability layer. Kwa, Thomas, and Garriga-Alonso (2024) prove the KL-regularized heavy-tail result; the extension to DRO, quantilization, and constrained RL is this paper's synthesis from the shared tail mechanism. Khan, Saveski, and Ugander (2024), Manski (1990), Vehtari et al. (2024), and Duchi and Namkoong (2019) supply the partial-identification, diagnostic, and local-risk tools used in the pricing discussion. Brown-Cohen, Irving, and Piliouras (2024) and Brown-Cohen, Irving, and Piliouras (2025) support the debate-recursion discussion, conditional on a faithful welfare-touching leaf oracle. Townsend (1979), Ben-Porath, Dekel, and Lipman (2014), and Ball and Kattwinkel (2025) supply the costly-verification lineage. Farquhar et al. (2025), Richens et al. (2025), and Virgo et al. (2025) are used only for the stated MONA and world-model/embodied-agent boundaries.

Proposition E1 is carried by Blackwell comparison and the Blackwell-Sherman-Stein theorem for the risk half, together with conditional data processing for the information half. Proposition E2 is scoped to trajectory-only self-audits: a stable gains-preserving institution must import exogenous outcome-contact, while mandatory external audit is one realization of that requirement rather than the requirement itself. The chi-squared overlap price functional is derived from the importance-sampling second-moment identity; what remains is the named modeling bridge from second moment to audit cost, not a missing derivation. The impossibility this paper builds on rests on the missing-mass theorem and does not depend on any unpublished experiment.

References

Ball, I., and Kattwinkel, D. (2025). "Probabilistic Verification in Mechanism Design." *Theoretical Economics* 20, 1247-1284. DOI:10.3982/TE6266.

Ben-Porath, E., Dekel, E., and Lipman, B. L. (2014). "Optimal Allocation with Costly Verification." *American Economic Review* 104(12), 3779-3813. DOI:10.1257/aer.104.12.3779.

Blackwell, D. (1951). "Comparison of Experiments." In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 93-102.

Blackwell, D. (1953). "Equivalent Comparisons of Experiments." *Annals of Mathematical Statistics* 24(2), 265-272.

Brown-Cohen, J., Irving, G., and Piliouras, G. (2024). "Scalable AI Safety via Doubly-Efficient Debate." In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 4585-4602. arXiv:2311.14125.

Brown-Cohen, J., Irving, G., and Piliouras, G. (2025). "Avoiding Obfuscation with Prover-Estimator Debate." arXiv:2506.13609.

Conant, R. C., and Ashby, W. R. (1970). "Every Good Regulator of a System Must Be a Model of That System." *International Journal of Systems Science* 1(2), 89-97. DOI:10.1080/00207727008920220.

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory.* 2nd ed. Wiley.

Dobrushin, R. L. (1959). "General Formulation of Shannon's Main Theorem in Information Theory." *Uspekhi Matematicheskikh Nauk* 14(6), 3-104; English translation in *American Mathematical Society Translations*, Series 2, 33, 323-438.

Duchi, J., and Namkoong, H. (2019). "Variance-Based Regularization with Convex Objectives." *Journal of Machine Learning Research* 20(68), 1-55. arXiv:1610.02581.

Farquhar, S., Varma, V., Lindner, D., Elson, D., Biddulph, C., Goodfellow, I., and Shah, R. (2025). "MONA: Myopic Optimization with Non-myopic Approval Can Mitigate Multi-step Reward Hacking." arXiv:2501.13011.

Fazulyanov, I. (2026). *The Goodhart Horizon: A Self-Monitoring Impossibility Theorem for Proxy Optimization, and the Price of Seeing Past It.* Preliminary working paper, v1.8, July 2026.

Francis, B. A., and Wonham, W. M. (1976). "The Internal Model Principle of Control Theory." *Automatica* 12(5), 457-465. DOI:10.1016/0005-1098(76)90006-6.

Gelfand, I. M., and Yaglom, A. M. (1959). "Calculation of the Amount of Information About a Random Function Contained in Another Such Function." *American Mathematical Society Translations*, Series 2, 12, 199-246.

Khan, S., Saveski, M., and Ugander, J. (2024). "Off-Policy Evaluation Beyond Overlap: Sharp Partial Identification Under Smoothness." In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 23734-23757. arXiv:2305.11812.

Kwa, T., Thomas, D., and Garriga-Alonso, A. (2024). "Catastrophic Goodhart: Regularizing RLHF with KL Divergence Does Not Mitigate Heavy-Tailed Reward Misspecification." *Advances in Neural Information Processing Systems* 37. arXiv:2407.14503.

Manski, C. F. (1990). "Nonparametric Bounds on Treatment Effects." *American Economic Review Papers and Proceedings* 80(2), 319-323.

Mossel, E., and Ohannessian, M. I. (2019). "On the Impossibility of Learning the Missing Mass." *Entropy* 21(1), 28. arXiv:1503.03613.

Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes.* Translated and edited by A. Feinstein. Holden-Day.

Richens, J., Abel, D., Bellot, A., and Everitt, T. (2025). "General Agents Need World Models." arXiv:2506.01622.

Sherman, S. (1951). "On a Theorem of Hardy, Littlewood, Polya, and Blackwell." *Proceedings of the National Academy of Sciences* 37, 826-831.

Stein, C. (1951). *Notes on the Comparison of Experiments.* University of Chicago, unpublished notes.

Townsend, R. M. (1979). "Optimal Contracts and Competitive Markets with Costly State Verification." *Journal of Economic Theory* 21(2), 265-293.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2024). "Pareto Smoothed Importance Sampling." *Journal of Machine Learning Research* 25(72), 1-58.

Virgo, N., Biehl, M., Baltieri, M., and Capucci, M. (2025). "A 'Good Regulator Theorem' for Embodied Agents." arXiv:2508.06326.

Williamson, C., and Cranko, Z. (2024). "Information Processing Equalities and the Information-Risk Bridge." *Journal of Machine Learning Research* 25(103), 1-53.