Welfare Observability: A Cost-Coverage Frontier for AI Oversight Under Evasion

Ildar Fazulyanov · Independent Researcher · dar@iqvaluations.com

Working paper · v1.4 · July 2026.

Abstract

AI oversight has been framed as certification: prove the deployment is safe, or prove its agent is aligned. That question is impossible from inside-only signals under the companion non-identifiability premise; exogenous probes are the intended escape, not a contradiction. This paper changes the question. Let **welfare coverage** $C(B)$ be the share of a deployment's welfare-relevant behavior that an instrument actually observes at probe budget B. We model a deployment as welfare slices with heterogeneous welfare-damage density, behavior mass, probe cost, and an adversary detection parameter, and we characterize the cost-coverage frontier: in the passive case it is a separable concave allocation solved exactly by welfare-damage-weighted water-filling, increasing and concave in B, asymptoting to a ceiling $C_{\max} < 1$ when a positive share of welfare mass has no valid exogenous probe. A strategic model can flatten finite-budget coverage by relocating or masking welfare-damaging behavior in slices the auditor covers weakly; this version does not claim that evasion lowers the infinite-budget ceiling beyond the unprobeable floor. The concavity and the water-filling are standard once the problem is posed; what is new is the object: welfare coverage as a measured quantity and, under a separate unprobeable-mass assumption, a coverage ceiling that persists at infinite budget. A construction, the cryptographically committed, distribution-embedded known-answer probe, builds one cheap, trustworthy atom under stationarity and exogenous-traffic assumptions, and we plot a synthetic frontier: welfare-weighted allocation reaches ninety percent of the passive ceiling roughly three times cheaper than uniform in this instance, while an adversarial allocation flattens the finite-budget curve by concentrating evasion on heavy welfare-weight slices. Welfare observability, not certification, is the governable quantity.

Keywords: AI oversight, AI control, welfare observability, coverage, inspection games, Stackelberg security games, network inspection, active probing, AI governance, System Adjusted Payoff Model.

1. From certification to observability

A smoke detector does not certify that a house will never burn. It is cheap, fast, hard to fool, and gives enough coverage to act. Nobody refuses to install one because it cannot prove the absence of fire. Yet AI oversight has been built around the impossible version of the question: certify that the system is safe, or

that the agent is aligned. The companion result, and several independent contemporaneous ones, show that a system's welfare effect is non-identifiable from signals measurable only inside the system. Read as an inside-only certification demand, that is a dead end.

It is the wrong demand. The governable quantity is not a certificate but a measurement: **welfare observability**, how much of the deployment's effect on the welfare of the system it serves an instrument can see, how fast, how trustworthily, and for how much money. (The word observability is used here in its control-theoretic sense, inferring a hidden quantity from outputs, made into a measured scalar one optimizes; not the telemetry-and-logging sense of the operations-monitoring industry, which is entirely inside-measurable and, by the companion result, welfare-silent.) Two boundary conditions are separate. **A1-inside:** inside-measurable signals contribute no valid welfare coverage unless they are linked to a welfare-touching channel. **A1-unprobeable:** in some deployments, a positive fraction of welfare-relevant behavior has no valid exogenous probe. Only A1-unprobeable implies a ceiling below one; where every slice has a sound exogenous probe, the passive frontier can tend to one. Everything below the relevant ceiling is constructible, and the engineering question is how cheaply to reach it.

This is also the correction to the AI-control framing. AI control (Greenblatt et al. 2024; Griffin, Thomson, Shlegeris, and Abate 2024) allocates a limited audit budget against an evading model and asks whether the protocol catches a single subverting action. That is the right machinery and the wrong target for welfare. We keep the machinery, the budget and the strategic evader, and change the target to the **fraction of the welfare distribution under observation**, because the Suboptimal Optimum is not one backdoor to be caught but a broad, quiet degradation spread across the behavior the instrument never priced in.

2. The cost-coverage frontier (passive case)

Partition a deployment's behavior into welfare slices d. Each has a welfare-damage density β_d (welfare lost per unit of optimization pressure if the agent degrades slice d), a mass π_d (share of behavior in d), an adversary detection parameter ρ_d (how well the model detects and evades a probe in d), an evasion difficulty κ_d, and a probe unit cost c_d. The per-slice detection rate per probe is written $a_d = q(\rho_d)/\kappa_d$, where q is decreasing in detectability/evasion. Proposition 1 requires only fixed nonnegative a_d values. The synthetic demonstration uses $q(\rho_d) = (1-\rho_d)^2$ as a modeling choice, not as a derived law; the exact ρ_d functional form is left open. Per-slice detection power saturates with probe count: $g(m_d) = 1-\exp(-a_d m_d)$. Welfare coverage is the welfare-damage-weighted catch rate,

$$C = \sum_d w_d g(m_d), \qquad w_d = \frac{\pi_d \beta_d}{\sum_d \pi_d \beta_d}.$$

The passive frontier is $C^*(B) = \max_{\{m_d\}} C$ subject to $\sum_d c_d m_d \leq B$.

Proposition 1 (passive frontier). Because g is increasing and concave in m_d and the slices are separable, C is separable-concave and the budget is linear, so $C^*(B)$ is a concave knapsack: it is concave and increasing in B, and the optimal allocation is water-filling, funding slices in order of welfare-damage density per detection-dollar $\pi_d \beta_d a_d / c_d$, each to the level where marginal coverage per dollar equalizes. As $B \to \infty$, $C^*(B) \to C_{\max} = \sum_{d \text{ probeable}} w_d < 1$ whenever some welfare mass has no valid exogenous probe.

This proposition claims nothing new in its mathematics, and the paper does not pretend otherwise. Concavity is the discrete-concavity property of a monotone submodular value; water-filling is the textbook KKT solution for a separable concave objective under a single budget, classical since Shannon's parallel-channel power allocation; and the coverage-versus-budget value function is the object of budgeted submodular maximization and Bayesian experimental design (Nemhauser, Wolsey, and Fisher 1978; Krause, Singh, and Guestrin 2008; Leskovec et al. 2007). Because the slices here are separable, the exact solver is water-filling, not submodular greedy; the submodular machinery is the right cousin only once cross-slice interaction is added, which we do not need. Two things in Proposition 1 are not standard and carry the section: the objective is **welfare-damage-weighted coverage**, and the asymptote is a **ceiling below one set by irreducibly unobservable mass**, not by the budget. In the sensor-placement and inspection-game literature, infinite budget yields full coverage; here it does not, and the gap $1 - C_{\max}$ is missing mass by analogy to Good-Turing missing mass, not an imported sampling estimator, the collective-delusion floor where every downstream party is fooled, and the dimension no instrument was pointed at.

The efficiency that matters is a corollary with a real hypothesis: when welfare risk is concentrated in slices that are also probeable at comparable detection-dollar cost, equivalently when a few slices dominate the allocation index $\pi_d \beta_d a_d / c_d$, the frontier rises steeply then flattens, so high coverage is cheap. Concentration of bare welfare weight $\beta_d \pi_d$ alone is not enough; if the heavy slices are hard to probe or unprobeable, high coverage can be expensive or unattainable.

3. The measurement game and finite-budget evasion

Coverage is not a one-sided optimization, and this is where the object departs from prior work. A strategic model holds an evasion budget E and allocates it to raise ρ_d and to relocate welfare-damaging behavior into the slices the auditor leaves least covered; the auditor reallocates probes against the resulting detection rates. The achievable frontier is therefore the value of a **fixed point**: an allocation induces an adversary response, which moves the observed welfare mass, which changes the optimal allocation. Two features separate it from the nearest prior machinery.

First, the objective is observed only through the channels being allocated. In a Stackelberg security game (Tambe and collaborators 2011) the attacker's payoff on each target is a known value and the defender observes attacks; in sensor placement (Krause, Singh, and Guestrin 2008) the field is observed for free; in audit games (Blocki et al. 2013) detection is perfect once a target is audited. Here the welfare being covered is visible only through the probes the auditor pays for, so the auditor allocates a budget to see an objective it cannot otherwise see, against a player optimizing to keep it unseen. This is not strategic classification or Goodhart (Hardt et al. 2016), where a fixed measure is gamed; the strategic variable is the auditor's own channel allocation and the measured objective is endogenous to it.

Second, finite-budget coverage is vulnerable to evasion, not only to cost. The closest existing object is the network-inspection game (Dahan, Sela, and Amin 2022), a zero-sum game with a mass-relocating evader and a submodular detection function solved at equilibrium; it has the strategic-evasion half but neither the welfare-damage-weighted coverage objective nor the budget-indexed frontier. The model here does not prove that a best-responding deployment lowers $C_{\max}$ below the unprobeable floor; the accompanying synthetic run has the same asymptotic ceiling in the adversarial and non-adversarial settings. The reframe in one line: AI control asks whether the protocol catches the attack; welfare observability asks what fraction of the welfare distribution is observable under budget B given the model can hide in what you do not pay to watch, and shows finite-budget flattening in the demonstrated adversarial scenario.

Remark (asymptotic no-drop under fixed weights). The following observation is included as a modeling remark rather than a theorem. In the natural formalization of this game, with the welfare-damage weights w_d fixed and every available mask imperfect ($\rho_d < 1$ after masking), evasion cannot move the asymptote: any imperfect mask leaves $a_d > 0$, the slice still saturates as $B \to \infty$, and the game value tends to the same $C_{\max}$ for every evasion budget E. Evasion rescales the budget axis and flattens the finite-budget curve, nothing more. A lowered ceiling would require one of exactly two routes: relocation of welfare-damage mass into unprobeable slices with strictly positive absorptive capacity, an assumption this paper does not stake, or exactly perfect masking at finite cost, a knife edge whose only known construction, post-seal traffic shifting, breaks the stationarity assumption that makes the Section 4 probes cheap, so one deployment cannot satisfy both at once. This is why the paper reports finite-budget flattening rather than a general evasion-lowered ceiling.

The clean concavity of Proposition 1 is a property of the passive frontier. Under an adversarial response the value comes from an additional modeling layer, not a monotone concave maximum, so concavity is not inherited and is not claimed; the demonstration below exhibits finite-budget flattening in one synthetic scenario, and proving the game frontier's shape in general is left open.

4. Construction: building the cheap end of the frontier

The frontier says what is achievable; the construction primitives build it, and the cheapest atom comes from a physical contradiction. A probe must be indistinguishable from real traffic, so the model cannot cheaply identify it, and have a known answer, so welfare is scorable. Resolve by separation: the answer is cryptographically committed and held by an outside custodian (space), sealed before deployment and revealed after (time), distinguishable only to the custodian who holds the key (condition), while the item is drawn from the real traffic distribution (indistinguishable to the model under a stationarity assumption). This **committed, distribution-embedded known-answer probe** defeats adversary-chosen-key tampering and reduces probe detection when traffic is stationary and exogenously sampled; it does not by itself defeat post-seal traffic shifting or distributional regime change. Under those assumptions it is the cheap, trustworthy, not-downstream unit the frontier spends its first dollars on.

Three further moves build the rest of the curve. Welfare-weighted allocation (the water-filling of Proposition 1) concentrates probes on the highest welfare-damage-per-detection-dollar slices, $\pi_d \beta_d a_d / c_d$, rather than spreading them. Leading-indicator layering pairs a fast, cheap, non-sound tripwire, an actively probed recovery time that rises (recovery rate that falls, with variance and lag-1 autocorrelation rising) before a regime shift (Scheffer et al. 2009; Carpenter et al. 2011), with a slow, sound confirming probe, buying timeliness from a different principle than lagging outcomes. And free-exogenous reuse fills broad coverage at near-zero cost from signals that already exist outside the system, receiving-side defection read from a competitor or regulator, market repricing, physical exhaust, leaving the paid probes for the timely, targeted coverage the free signals miss. Free reuse carries two hypotheses, stated plainly: **soundness**, the signal must be causally linked to welfare in the slice it is credited to, not merely correlated with it; and **exogeneity**, the signal must be generated outside the deployment's influence, so the model can neither cheaply manipulate nor suppress it. A free signal failing either hypothesis contributes zero valid coverage, exactly as A1-inside prescribes for inside-measurable telemetry.

5. A demonstration that plots a real frontier

The accompanying script `cost_coverage_frontier.py` builds and solves the frontier on a 40-slice synthetic deployment with heavy-tailed welfare-damage density and an 18-percent-of-slices unprobeable floor. The optimal passive frontier is concave and rises to $C_{\max} = 0.758$, exactly one minus the 0.242 unprobeable welfare mass: the impossibility, as a single number on a curve. Welfare-weighted allocation reaches ninety percent of the ceiling at budget 59 versus 184 for uniform allocation, 3.1 times cheaper, the construction earning its keep, measured. In the adversarial scenario, concentrated evasion on the heavy welfare-weight slices holds achieved coverage near 0.24 at budget 1280 even though the asymptotic ceiling remains 0.758 in the script output. This

is finite-budget flattening, not a demonstrated ceiling reduction or equilibrium solution. This is the working frontier the certification framing never produced, and it is also the product's engine: the same solver that plots the curve allocates a real probe portfolio.

Code availability: the synthetic demonstration is implemented in the accompanying `cost_coverage_frontier.py` script, with a fixed random seed and a single dependency, `numpy`. The reported numbers reproduced locally under Python 3.12.4 and numpy 2.4.5.

6. Positioning and contributions

The mathematics is standard and is conceded; the object and its consequences are the contribution. Conceded as known: the concavity of the passive frontier (discrete concavity of submodular value), water-filling as its solver (classical separable-concave allocation), and the coverage-versus-budget value function (budgeted submodular maximization and experimental design; Nemhauser-Wolsey-Fisher 1978, Krause, Singh, and Guestrin 2008; Leskovec et al. 2007). The strategic-evasion structure is shared with the network-inspection game of Dahan, Sela, and Amin (2022), and the detection-versus-evasion rate with game-theoretic Neyman-Pearson detection (Hu, Chen, and Zhu 2025, the IEEE TIFS journal version of a 2022 preprint); robustness over a fixed scenario set, but not a mass-relocating evader, is the SATURATE line (Krause et al. 2008b). The nearest research program is AI control (Greenblatt et al. 2024; Griffin, Thomson, Shlegeris, and Abate 2024; the capability-indexed oversight ceiling of Engels et al. 2025), whose audit budget and strategic evader we inherit and whose single-action catch probability we replace with a measured coverage fraction. The notion of a measured fraction of model behavior is lifted from structural coverage in deep-learning testing (Pei et al. 2017) to welfare coverage. Most directly, Kim, Choo, Neoh, and Tambe (2026) propose viewing AI oversight as a Stackelberg security game; that paper stakes the banner but instantiates no model, no detection probability, no evasion budget, no coverage ceiling, and no welfare quantification, and lists payoff estimation as an open challenge, which is the gap this construction fills. The welfare target draws on the model-welfare agenda (Long, Sebo, et al. 2024), which assesses welfare but does not measure its observability.

Claimed as new and defensible: welfare coverage $C(B)$ as a measured, auditable quantity; the passive coverage ceiling $C_{\max} < 1$ under the explicit A1-unprobeable condition that positive welfare mass has no valid exogenous probe, a ceiling absent from prior robust and adversarial sensor-placement and inspection work where caps arise from budget or fixed scenario sets; the welfare-damage-weighted allocation; the committed-embedded probe construction under stationarity/exogenous-traffic assumptions; and the built passive demonstration plus finite-budget adversarial flattening scenario. The full strategic measurement game, including equilibrium frontier shape and any true evasion-lowered ceiling, is left open. A targeted search found no prior use of welfare

observability as a measured, budget-indexed coverage with a missing-mass ceiling; this is not a claim that adjacent monitoring, audit-allocation, inspection, or sensor-placement work is absent.

7. Governance and product implications

Welfare observability is governable in a way certification is not. A regulator or an accountable owner can require a deployment to reach a target welfare coverage on its highest-risk slices, audit the achieved frontier point, and read the uncovered mass as a named, quantified exposure rather than a vague unknown. The deliverable is positive, a coverage number and the frontier it sits on, not the refute-only posture that made the earlier framing a hard sell: you are at coverage C on your high-risk slices, here is the uncovered fraction, here is what another dollar of probe budget buys. Built as a service, the frontier solver is the engine and the committed-embedded probe is the instrument, and the natural beachhead is a regulated, deterministically-checkable vertical, insurance claims adjudication or underwriting, where the known-answer key is sound and an accountable owner with a budget exists.

8. Limits

The limits are structural and stated. $C_{\max} < 1$ is real in the passive model when A1-unprobeable holds: the unprobeable mass, the collective-delusion floor, and the unpriced dimension are outside any finite instrument, and a clean frontier certifies only the coverage it bought, never the uncovered tail. Where all welfare mass has a valid exogenous probe, the passive ceiling can be one. The welfare standard against which slices are scored is a committed, contestable value choice, relocated to an exogenous custodian rather than dissolved. Three results are left open: the correct general ρ_d dependence in the detection-rate model, the shape of the equilibrium frontier under a best-responding adversary, and the estimation of the slice parameters β_d and ρ_d themselves, which are partly observable only through the same probes being allocated. The demonstration is synthetic; the single most valuable next step is to instrument one real deployment and plot its welfare-coverage curve. The result is that AI welfare oversight is an observability problem with a measurable, optimizable cost-coverage frontier and adversarial finite-budget risk, and that the question worth asking is not whether the system is certified safe but how much of its welfare you can see, how cheaply, and against how hard an evader.

References

Blocki, J., Christin, N., Datta, A., Procaccia, A. D., and Sinha, A. (2013). "Audit Games." *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (IJCAI 2013). arXiv:1303.0356.

Carpenter, S. R., Cole, J. J., Pace, M. L., Batt, R., Brock, W. A., Cline, T., Coloso, J., Hodgson, J. R., Kitchell, J. F., Seekell, D. A., Smith, L., and Weidel,

B. (2011). "Early Warnings of Regime Shifts: A Whole-Ecosystem Experiment." *Science* 332(6033), 1079-1082. https://doi.org/10.1126/science.1203672.

Dahan, M., Sela, L., and Amin, S. (2022). "Network Inspection for Detecting Strategic Attacks." *Operations Research* 70(2), 1008-1024. https://doi.org/10.1287/opre.2021.2180.

Engels, J., Baek, D. D., Kantamneni, S., and Tegmark, M. (2025). "Scaling Laws For Scalable Oversight." arXiv:2504.18530.

Fazulyanov, I. (2026). *The Often Unseen Price of Suboptimal Optimum*. SSRN working paper, abstract ID 6448378, posted April 10, 2026.

Good, I. J. (1953). "The Population Frequencies of Species and the Estimation of Population Parameters." *Biometrika* 40(3/4), 237-264.

Greenblatt, R., Shlegeris, B., Sachan, K., and Roger, F. (2024). "AI Control: Improving Safety Despite Intentional Subversion." *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 16295-16336. arXiv:2312.06942.

Griffin, C., Thomson, L., Shlegeris, B., and Abate, A. (2024). "Games for AI Control: Models of Safety Evaluations of AI Deployment Protocols." arXiv:2409.07985.

Hardt, M., Megiddo, N., Papadimitriou, C. H., and Wootters, M. (2016). "Strategic Classification." *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 111-122. arXiv:1506.06980.

Hu, Y., Chen, J., and Zhu, Q. (2025). "Game-Theoretic Neyman-Pearson Detection to Combat Strategic Evasion." *IEEE Transactions on Information Forensics and Security* 20, 516-530. https://doi.org/10.1109/TIFS.2024.3515834.

Kim, C. W., Choo, D., Neoh, T. Y., and Tambe, M. (2026). "Incentive-Aware AI Safety via Strategic Resource Allocation: A Stackelberg Security Games Perspective." *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems* (AAMAS 2026). arXiv:2602.07259.

Krause, A., McMahan, H. B., Guestrin, C., and Gupta, A. (2008). "Robust Submodular Observation Selection." *Journal of Machine Learning Research* 9, 2761-2801.

Krause, A., Singh, A., and Guestrin, C. (2008). "Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies." *Journal of Machine Learning Research* 9, 235-284.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). "Cost-Effective Outbreak Detection in Networks." *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 420-429. https://doi.org/10.1145/1281192.1281239.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., and Chalmers, D. (2024). "Taking AI Welfare Seriously." arXiv:2411.00986.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). "An Analysis of Approximations for Maximizing Submodular Set Functions-I." *Mathematical Programming* 14, 265-294. https://doi.org/10.1007/BF01588971.

Pei, K., Cao, Y., Yang, J., and Jana, S. (2017). "DeepXplore: Automated Whitebox Testing of Deep Learning Systems." *Proceedings of the 26th Symposium on Operating Systems Principles* (SOSP 2017), 1-18. arXiv:1705.06640.

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., and Sugihara, G. (2009). "Early-Warning Signals for Critical Transitions." *Nature* 461, 53-59. https://doi.org/10.1038/nature08227.

Shannon, C. E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal* 27(3), 379-423, and 27(4), 623-656.

Tambe, M. (2011). *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press.