When Stagewise Truthfulness Fails to Compose in Adaptive Ensembles

Anonymous authors
Paper under double-blind review

Abstract

Prototyped and proposed prediction ensembles settle contributor payments by marginal contribution after labels arrive, and adapt serving weights on the same measurements. We show that delayed settlement turns adaptive allocation into unpriced intertemporal influence. For any smooth strictly proper settlement: with no allocation stake, stagewise truthfulness composes exactly; with one, first-order truthfulness holds precisely when the unpriced rent gradient vanishes on the simplex tangent space, and otherwise *farming* — misreporting to capture future serving weight — is first-order profitable and, under the stated curvature conditions, deviations inside a history-conditional *cushion* (the Hessian-induced ellipsoid of deviation sizes) induce a law inside the composite null of every anytime-safe own-stream detector. Coupling a sound detector to automatic reallocation can make matters worse — a framing deviator can have an honest peer correctly excluded and inherit its weight (demonstrated end-to-end on an engineered extended-horizon replay), with cascades at every $n \geq 3$ under the stated exclusion semantics — and a trust region on published weights restores a $\sqrt{\tau}$ best-response bound, though a preregistered practical-remedy gate fails on the raw trace. Empirically, in a semi-synthetic replay family over 96 preregistered operating points, the rent gradient is confirmed nonzero in all six top-screened cells at disjoint-seed Bonferroni-corrected intervals (including under the deployed weight map on both traces), though all six share one deviating agent and, on one trace, the confirmed interval covers zero once the pool grows to five contributors; in a follow-up round frozen after the preregistered sizes failed, the sustained deviation at the total-curvature-optimal size is profitable at $\kappa = 1$, where κ prices one unit of serving weight in units of settlement, while the deployed detector stays silent; the stage envelope's moment condition certifiably fails on both raw traces and in every preregistered time segment; and a prespecified *anti-pool* audit (reporting away from the pooled mean) on the raw trace raises settlement income and allocation rent undetected, reported as a deployment vulnerability rather than as peer-channel validation. We conclude that adaptive allocation belongs inside the mechanism design, on the same footing as the payment rule.

1 Introduction

Prediction ensembles, prototyped on intrusion-detection traces and proposed for decentralized LLM aggregation, pay their contributors by measured marginal contribution once the label arrives, and adapt serving weights on the same measurements (Wang, 2026; Luo et al., 2026) (Figure 1). The standard design separates the two: a proper-scoring settlement makes truthful reporting a strict stage best response, while allocation is treated as harmless bookkeeping — weights are not money, so adapting them should not distort reports.

This paper shows the separation is not harmless. Delayed settlement turns adaptive allocation into *unpriced intertemporal influence*: whenever the *rent gradient* — the derivative of an agent's future serving weight in its current report — is nonzero along a feasible direction, that unpriced rent opens a first-order channel invisible to settlement.

Contributions. Noncomposition (Sections 2 and 3). With no allocation stake a separability lemma composes stagewise into dynamic truthfulness; with one, Theorem 1 makes first-order truth-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

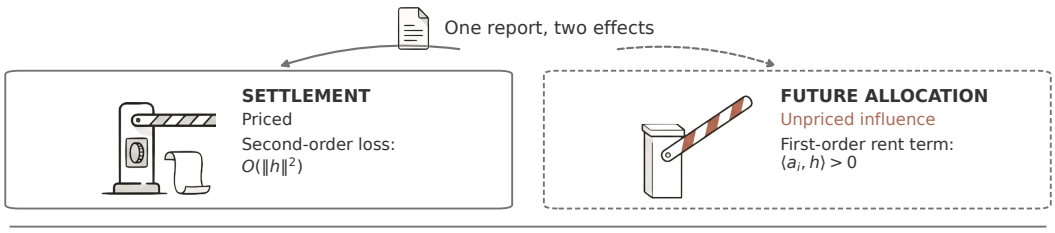

Figure 1: The same contribution measurement enters settlement and future weight updates. Under Theorem 1's assumptions, a suitable small deviation along a profitable direction can yield a first-order allocation gain that outweighs its second-order settlement loss.

fulness equivalent to the rent gradient vanishing on the tangent space. Otherwise *farming* pays, stays inside a Hessian-induced cushion, and hides in the composite null of *every* anytime-safe own-stream detector. **Detection changes incentives (Section 4).** An anytime-valid detector with an explicit delay bound, *composed* with automatic reallocation, can amplify the deviation (Theorem 2): a framing deviator stays null-side of its own monitor while an honest peer is excluded and its weight handed over; the repair cascades at every $n \geq 3$. **The price of restriction, and what the data shows (Sections 5 and 6).** A trust region restores a $\sqrt{\tau}$ best-response bound (Theorem 3). Preregistered rounds confirm rent-gradient nondegeneracy across a replay family — including the deployed map on two traces — profitable and detector-silent at the total-optimal size; the stage hypothesis (Proposition 1) certifiably fails on both raw traces, confining in-model support to that family.

The instantiated mechanism originates in a study of collective miscalibration in strategically aggregated ensembles (Wang, 2026) that contains none of the present results.

2 A REPEATED MECHANISM WITH EXACTLY-ONCE DELAYED SETTLEMENT

2.1 PRIMITIVES AND TIMELINE

Agents $N = \{1, \ldots, n\}$ report simultaneously at decision epochs $t = 1, \ldots, H$ ($H \leq \infty$): agent i holds a private forecast component $b_{i,t} \in \Delta^{K-1}$ (its *type*) over a categorical outcome $y_t \in \{1, \ldots, K\}$ and submits an immutable report $r_{i,t} \in \Delta^{K-1}$ with unique ID (i, t). The platform serves the weighted aggregate $\hat{p}_t = \sum_i \hat{s}_i^{(e(t))} r_{i,t}$ under the weight epoch $e(t)$ current at t, and takes a decision $a_t = a(\hat{p}_t)$ (binary $K = 2$, threshold rule, is the running instance). The label y_t becomes visible only at $t + D_t$; the delay process (D_t) may be random and unbounded but is independent of reports *and labels* (**A-delay**); labels may arrive out of order or duplicated.

Definition 1 (Exactly-once delayed settlement)**.** The settlement ledger maps report IDs to settlement records and obeys:

(EO1) *One settlement event per sample.* When the first copy of y_t arrives, sample t moves from *pending* to *settled*; every later copy of y_t is dropped.

(EO2) *The transfer is a pure function of the frozen tuple.* Settlement of sample t pays agent i $\pi_{i,t} = \alpha\, \Delta_i(r_{\cdot,t}, y_t)$, where

$$\begin{aligned} \Delta_i(r, y) &= v(N; r, y) - v(N \setminus \{i\}; r, y), \\ v(S; r, y) &= \mathcal{S}\big(\bar{r}_S,\, y\big), \quad \bar{r}_S = \tfrac{1}{|S|} \textstyle\sum_{j \in S} r_j, \end{aligned} \tag{1}$$

computed once, from the report vector frozen at epoch t and the label y_t only — the leave-one-out marginal contribution of cooperative game theory (Shapley, 1953), in the form used for data valuation (Ghorbani & Zou, 2019); $\mathcal{S}$ (calligraphic, unlike the coalition S) is a strictly proper scoring rule (reward orientation), with the *Brier instance* $\mathcal{S}_{\mathrm{B}}(q, y) = -\frac{1}{2}\|q - e_y\|_2^2$ (binary scalar form exactly $-(\bar{r} - y)^2$) carried throughout; every local statement below is exact under it.

(EO3) *Rolling statistics are not money.* A window W of the last m settled samples feeds the weight map $\hat{s}^{(e+1)} = \varphi(\bar{\Delta}_1^W, \ldots, \bar{\Delta}_n^W)$ (our softmax family is exponential weights on those marginals,

Freund & Schapire, 1997); re-averaging a window never re-issues a transfer, so agent i's cumulative balance is exactly $\sum_{t \text{ settled}} \pi_{i,t}$, independent of m and the publication cadence.

2.2 UTILITY: SETTLEMENT PLUS ALLOCATION RENT

With discount $\beta \in (0, 1]$ ($\beta < 1$ if $H = \infty$; transfers and weights are bounded, so all sums below converge absolutely), agent i's utility is

$$U_i = \mathbb{E}\Big[\sum_{t \le H} \beta^{\,t+D_t}\, \pi_{i,t}\Big] + \chi_i\, \mathbb{E}\Big[\sum_{t \le H} \beta^{\,t}\, \hat{s}_i^{(e(t))}\Big], \tag{2}$$

where $\chi_i \ge 0$ is the per-epoch value of one unit of serving weight: traffic share, visibility, downstream revenue. The second term is *allocation rent*, and its timing is the model's central structural fact: by EO3 the report $r_{i,t}$ enters the weights only *after* sample t settles, so a report influences allocation strictly forward in time. Lemma K (Section B) gives χ_i a *local, direction-dependent* interpretation through a Lipschitz served-output value; no global χ_i-to-L_i equivalence is used anywhere. Utility from the *current* decision a_t is deliberately excluded: valuing today's decision distorts even a static mechanism, a classical, separate problem; (2) isolates the intertemporal component.

Definition 2 (Stagewise vs. dynamic truthfulness)**.** The stage score of agent i is $r_i \mapsto \mathbb{E}[\Delta_i(r_i, r_{-i}, y)]$. The mechanism is *stagewise truthful* if truthful reporting is a strict best response of the stage score at the truthful r_{-i}. The everywhere-truthful profile is a *dynamic equilibrium* if every deviating strategy that alters some report with positive probability yields strictly lower utility (2) (probability-zero differences are payoff-equivalent).

What b_i is. The outcome model for all expectations (Section 3.1) is correctly specified *in the aggregate*: $Y_t \mid (b_{1,t}, \ldots, b_{n,t}) \sim \mathrm{Cat}(\bar{p}_t)$ with $\bar{p}_t = \frac{1}{n}\sum_j b_{j,t}$. The b_i are private *components* of this additive decomposition (complementary evidence the mechanism pools, not heterogeneous subjective posteriors), and truthfulness means *truthful component reporting*: given truthful peers, b_i is the unique stage optimum because it completes the pooled posterior. No Bayesian incentive compatibility is claimed for discordant interim beliefs; that is a different model. (Scope: no strategic participation/timing, no label tampering; A-delay is the trusted-feedback assumption; transfers may be negative; budget balance is out of scope.)

2.3 SEPARABILITY, AND THE QUESTION THIS PAPER ANSWERS

Lemma (Separability)**.** *Assume EO1–EO3, A-delay, and $\chi_i = 0$. Then $U_i = \alpha \sum_{t \le H} \mathbb{E}[\beta^{\,t+D_t}]\, \mathbb{E}[\Delta_i(r_{\cdot,t}, y_t)]$, a sum whose t-th term depends on agent i's strategy only through $r_{i,t}$. If the mechanism is stagewise truthful, the everywhere-truthful profile is a dynamic equilibrium, for every delay distribution, discount, window, and cadence,* regardless of the weight map*: an agent with no allocation stake cannot be tempted by it. (Proof: Section A.)*

When $\chi_i > 0$ the separation gives away unpriced influence, and exploitability is *local*: a deviation's settlement cost is second order (strong properness), so everything turns on whether it moves the agent's own future weight at *first* order. One object governs the answer: the projected rent gradient $\Pi_{\mathcal{T}} a_i$ ($\Pi_{\mathcal{T}}$: the orthogonal projection onto the simplex tangent space; Section 3), in one dimension $\chi_i \psi_i$ with ψ_i the tangent weight response (Section 3 defines it). Theorem 1: $\Pi_{\mathcal{T}} a_i \neq 0$ makes farming profitable at first order and invisible to safe monitoring; a worked instance (Section 3.2) makes the coefficient generically nonzero; a dead-zone construction (Section 3.1) engineers it to vanish.

3 TRUTHFULNESS DOES NOT COMPOSE: INVISIBLE INFLUENCE FARMING

3.1 THE TANGENT-SPACE GEOMETRY OF SETTLEMENT AND RENT

Write $\phi(q) = \mathbb{E}_{Y \sim q}\, \mathcal{S}(q, Y)$ for the potential, so $\mathbb{E}_{Y \sim q}\mathcal{S}(r, Y) = \phi(q) - D_\phi(q, r)$, D_ϕ the Bregman divergence of the convex ϕ (Gneiting & Raftery, 2007; Abernethy & Frongillo, 2012); deviations live on $\mathcal{T} = \{h : \mathbf{1}^\top h = 0\}$. Under the pooled-posterior model $Y \mid (b_1, \ldots, b_n) \sim \mathrm{Cat}(\bar{p})$, $\bar{p} = \frac{1}{n}\sum_j b_j$, and (S1) below, a deviation $h \in \mathcal{T}$ by agent i (others truthful; $w_j := b_j - \bar{p}$) satisfies,

exactly, $\mathbb{E}\bar{\Delta}_i(h) = \bar{\Delta}_i^0 - D_\phi(\bar{p}, \bar{p} + h/n)$, $\bar{\Delta}_i^0 = D_\phi(\bar{p}, \bar{b}_{-i})$; with $Q := n^{-2}\,\nabla^2\phi(\bar{p})\big|_{\mathcal{T}}$ (Q^{-1} always the inverse of this restriction *on* $\mathcal{T}$),

$$\mathbb{E}\bar{\Delta}_i(h) = \bar{\Delta}_i^0 - \tfrac{1}{2}\,h^\top Q\,h + o(\|h\|^2), \tag{3}$$

$$d_j := \nabla_h\,\mathbb{E}\bar{\Delta}_j\big|_0 = -\frac{\nabla^2\phi(\bar{b}_{-j})\big|_{\mathcal{T}}\,w_j}{(n-1)^2} \quad (j \neq i), \tag{4}$$

with *no remainder* in (3) for the Brier instance (globally quadratic), and $d_i = 0$: the truthful report is an interior stage optimum (envelope, Milgrom & Segal, 2002), so a deviator moves its *own* expected marginal only at second order, but each peer's at first order (proofs: Section C).

The dynamic rent functional. A deviation alters the whole future weight *trajectory* rather than a single snapshot. Fix h at epoch t (all else truthful) and run the deviating and truthful worlds on the same draws, so that all other reports, labels, and delays coincide: h reaches the future only through the published weights. The *counterfactual rent functional* is

$$\mathcal{R}_i(h) \coloneqq \chi_i\,\mathbb{E}\Big[\sum_{u>t}\beta^{\,u}\big(\hat{s}_i^{(e(u))}(h) - \hat{s}_i^{(e(u))}(0)\big)\Big], \tag{5}$$

and the *rent gradient* is $a_i \coloneqq \nabla_h\mathcal{R}_i(0) \in \mathcal{T}$; this functional, rather than any single-publication derivative, is the object Theorem 1 is about. It is computable (Lemma K′, Section B): with $\mathcal{A}_t := \{u > t :\, t \in W(e(u))\}$ the *affected epochs* (those whose serving weights depend on report t),

$$a_i = \chi_i\,\mathbb{E}\Big[\textstyle\sum_{u\in\mathcal{A}_t}\frac{\beta^{\,u}}{m}\,J_{\varphi_i,u}\,d^{(t)}\Big], \tag{6}$$

with $J_{\varphi_i,u}$ the weight-map Jacobian along the *realized* truthful window statistic of publication $e(u)$, $d^{(t)}$ the realized sample-t marginal-response matrix (mean rows the d_j of (4), $d_i = 0$), and the expectation *joint*: the two factors share sample t and are dependent; exchange and decoupling conditions (exact for affine maps; a Jensen-plus-covariance remainder in general) are in Section B. Only under one *deterministic* operating-point Jacobian does (6) factor through a scalar kernel $\omega_i = \mathbb{E}[\frac{1}{m}\sum_{u\in\mathcal{A}_t}\beta^{\,u}] > 0$ — there, and only there, cadence rescales a_i without rotating it (Corollary K′-stationary); in general the dichotomy is judged on the total gradient. In the examples' normalization $\omega_i = \beta^{\,t+1}$; both expansion terms below are per deviation report.

The own row of $d^{(t)}$ is *mean*-zero ($d_i = 0$): under a deterministic Jacobian (or decoupling) the rent gradient is purely a peer effect, while in general the own row contributes through its covariance with the random Jacobian. For a one-dimensional family $h = c\,h_0$ it reduces to the scalar $\chi_i\psi_i$, $\psi_i = \langle a_i, h_0\rangle/\chi_i$; the coefficient $\chi_i \geq 0$ is primitive in (2), with Lemma K's local, direction-dependent interpretation (map-independent only when the allocation direction is forced: $n = 2$, or fixed peer-debit proportions), and Theorem 3 prices the rent in χ_i directly; no global χ_i-to-L_i equivalence is used.

The naive general claim is false. The *dead-zone map* $\varphi_i(\Delta) = f_i(\Delta_i)/\sum_k f_k(\Delta_k)$, $f_i(x) = 1 + \max(x - \Delta_i^0, 0)^2$ with watermark Δ_i^0 at the truthful marginal, is C^1, normalized, and responsive, yet its Jacobian vanishes at the deterministic truthful point: in the mean-statistic evaluation, $a_i = 0$ for every agent. So $\chi_i > 0$ alone is not enough: nondegeneracy $\Pi_{\mathcal{T}}a_i \neq 0$ is an *assumption*, discharged per weight-map family; the construction returns as remedy R2 (Section 5).

3.2 Theorem 1: dynamic noncomposition

Fix the everywhere-truthful profile and a deviating agent i with deviation $h \in \mathcal{T}$. Assume:

- **(S1)** *(Regularity.)* $\bar{p}$, every $\bar{b}_{-j}$, and every b_i lie in the relative interior of Δ^{K-1}; the last ensures a full tangent-space neighborhood of the truthful report is feasible (else the feasible tangent cone replaces $\mathcal{T}$; not pursued); $\phi \in C^2$ near these points with $\nabla^2\phi(\bar{p}) \succ 0$ on $\mathcal{T}$ (log score: additionally a compact-interior condition); $A = \alpha\,\mathbb{E}[\beta^{\,t+D_t}] \times$ per-epoch settled mass (the expected number of reports settled per serving epoch) is the discounted settlement scale;
- **(S2)** the rent functional (5) admits the expansion $\mathcal{R}_i(h) = \langle a_i, h\rangle + o(\|h\|)$ (sufficient: a directionally differentiable φ at $\bar{\Delta}^0$ plus the kernel structure (6); Lemma K′, Section B);

(S3) $\Pi_{\mathcal{T}} a_i \neq 0$ (*rent-gradient nondegeneracy*, per Section 3.1);
(S4) $\chi_i > 0$.

For parts (c) and (d) we additionally assume the *history-conditional curvature sandwich*:

(S1′) conditionally on the full history $\mathcal{H}_{k-1}$, a per-stage deviation h shifts the settled-marginal mean by at most its quadratic form: $\mathbb{E}[Z_{i,k} \mid \mathcal{H}_{k-1}] \geq \bar{\Delta}_i^0(\mathcal{H}_{k-1}) - \frac{1}{2} h^\top Q(\mathcal{H}_{k-1})\, h$, with $\bar{\Delta}_i^0(\cdot) > 0$, $Q(\cdot) \succ 0$ the conditional truthful marginal and curvature (exact for Brier under i.i.d. labels; the conditional form is what drift requires).

A stopping rule T_{det}, measurable in the own settled-marginal stream $(Z_{i,k})_k \subset [-B, B]$, is δ-*safe* if $\sup_{P \in H_0} P(T_{\mathrm{det}} < \infty) \leq \delta$ over the composite null $H_0 = \{P : \mathbb{E}_P[Z_{i,k} \mid \mathcal{F}^Z_{k-1}] \geq 0 \text{ a.s. } \forall k\}$.

Theorem 1 (Dynamic noncomposition)**.** *Under (S1)–(S2):*

(a) ***(First-order dichotomy.)*** *The one-sided directional derivative of U_i at truthful along h equals $\langle a_i, h\rangle$; hence* first-order local truthfulness holds if and only if $\Pi_{\mathcal{T}} a_i = 0$. *Under (S3)–(S4), for every h with $\langle a_i, h\rangle > 0$ there is $\bar{t} > 0$ such that $U_i(th) > U_i(0)$ for all $t \in (0, \bar{t})$: truthfulness is not a best response.*

(b) ***(Weak-externality optimum.)*** *Let $\mathcal{R}_{i,\varepsilon} = \varepsilon\, \widetilde{\mathcal{R}}_i$ with $\widetilde{\mathcal{R}}_i$ C^2 near 0, second derivative uniformly bounded (weight geometry fixed), $\tilde{a}_i := \nabla \widetilde{\mathcal{R}}_i(0)$, so $a_{i,\varepsilon} = \varepsilon \tilde{a}_i$. As $\varepsilon \to 0$, the gain $\Delta U_i(h) := U_i(h) - U_i(0)$ satisfies*

$$h^\star_{i,\varepsilon} = \tfrac{\varepsilon}{A}\, Q^{-1} \Pi_{\mathcal{T}} \tilde{a}_i + o(\varepsilon),$$
$$\max_{\|h\| \leq \rho} \Delta U_i = \tfrac{\varepsilon^2}{2A}\, \|\Pi_{\mathcal{T}} \tilde{a}_i\|^2_{Q^{-1}} + o(\varepsilon^2),$$

for any fixed radius $\rho > 0$ inside (S1)'s expansion neighborhood: a local *maximum in general, and the global one when the settlement term is globally quadratic (Brier) or its Hessian admits a global negative-definite bound (global strong properness alone, a value inequality, does not supply one; Section C, where the necessity of the ε-family is also explained).*

(c) ***(The undetectable budget is an ellipsoid.)*** *Under (S1′), the own-stream conditional mean stays nonnegative whenever the per-stage deviation obeys $h^\top Q(\mathcal{H}_{k-1})\, h \leq 2\bar{\Delta}_i^0(\mathcal{H}_{k-1})$: the* cushion ellipsoid $\mathfrak{E}_i$. *Binary Brier: $|c| \leq \varrho_i = \sqrt{\bar{\Delta}_i^0/\mu_i}$, $\mu_i = 1/n^2$ the scalar curvature; at fixed curvature, high-marginal agents have the largest budgets, boundary agents none.*

(d) ***(Statistical invisibility.)*** *Assume history-stationary labels (conditionally on the full history, the next label is an independent draw from $\bar{p}$) and (S1′). Then* any adapted deviation policy *staying within the history-conditional cushion ellipsoid induces a law in H_0; hence,* for every such policy, *every δ-safe detector misses it with probability $\geq 1 - \delta$.*

Proof kernel. The proper-score settlement regret is *second* order in the deviation ((3)), the dynamic rent *first* order ((S2)): a nonzero tangent projection yields a profitable direction, and (b)'s ε-family makes the concave trade-off exact. For (c)–(d), the sandwich (S1′) keeps the conditional mean nonnegative inside the ellipsoid, and the tower property over the coarser own-stream filtration places every cushioned adapted policy's law in H_0: composite-null invisibility. Full proofs: Section C. □

Part (d) does *not* claim distributional indistinguishability (the mean dips by $\Theta(\|h\|^2)$): no detector can fire *while remaining δ-safe over H_0*, whose members include exactly such benign dips; a point-null-safe detector can fire but forfeits drift robustness (Section 4).

A worked instance, exactly. Two periods, two agents, components $b = (0.3, 0.7)$, affine weight map, externality billed on period-2 serving only: every quantity is an exact polynomial in s, the profitable direction is toward the pooled posterior, and the gain at the optimum $s^* \approx 0.123$ is strictly positive (unit parameters). Closed forms, the weak-externality linearization, and the price-of-honesty bound: Section A, machine-verified (Section J).

Population corollary (proportional weights, binary Brier). In the *stationary mean-statistic (population) normalization*, proportional weights on the all-active region have $\psi_i^{\mathrm{pop}} = 2\bar{\Delta}_i^0(p -$

$b_i)/((n-1)^2 S_\Delta^2) \neq 0$ ($S_\Delta = \sum_k \bar{\Delta}_k^0$) *throughout*: every agent has an active farming direction $\operatorname{sign}(p - b_i)$ and (S3) is discharged, within this normalization only; finite random windows carry no such guarantee, and no multiclass or softmax discharge is claimed — for softmax the condition stays a per-operating-point check (Section D) — (proof: Section C).

4 DETECTION CHANGES THE INCENTIVES

Theorem 1(c) bounds what safe monitoring sees below the cushion; beyond it anytime-valid detection works — Theorem 2's territory. For each agent, run on the settled-marginal stream $(Z_k) \subset [-B, B]$ a restart-mixture Hoeffding e-detector (dyadic restart windows, geometric λ-grid, summable instance prior), alarming when the mixture crosses n/δ (Section E; throughout, $\log^+ x := \max(1, \log x)$).

Lemma (Safety and delay)**.** *(i) Under H_0 the family-wise false-alarm probability over all agents and times is at most δ. (ii) Under a conditional-mean change to $\leq -g$ at changepoint ν, with probability $\geq 1-\delta''$ the alarm fires within $N \leq C\,\frac{B^2}{g^2}\big(\log\frac{n}{\delta\,\delta''} + \log\log\frac{2B}{g} + \log\log^+ N + \log^+\frac{\nu}{N}\big)$ post-change samples, C absolute. (Proof: Ville plus a dyadic covering; Section E.)*

The $\log^+(\nu/N)$ term is what *our* prior pays for anytime validity (no lower bound claimed; an average-run-length formulation is ν-free, Section E); it enters Theorem 2's accounting.

4.1 THEOREM 2: DETECTION–ACTION AMPLIFICATION

Couple the detector to *automatic reallocation*: an excluded agent's serving weight transfers to the survivors. The baseline is *alert-only* coupling (alarm raised, no transfer, continuous weighting retained), coupled sample-for-sample (identical reports, labels, delays, hence alarm behavior). In the worked instance (Section 3.2) a *framing* deviation s keeps the deviator's stream mean $\mu_1(s) = 0.04 - s^2/4 \geq 0$ while driving the honest peer's $\mu_2(s) = 0.04 - 0.4s + 0.75s^2 < 0$ on $s \in (2/15,\, 2/5)$.

Objects. $\widehat{U}_1$: the realized value of (2)'s two sums ($U_1 = \mathbb{E}\widehat{U}_1$); r_t: the exact per-epoch weight-value differential of reallocation over alert-only ($|r_t| \leq \chi_1$); T_{act}: the first serving epoch with post-alert weights (a distinct clock from settlement); $G := \sum_{t=T_{\mathrm{act}}}^{H} \beta^t r_t$ ($G := 0$ on no-alarm paths); $C_{\mathrm{fail}} := \chi_1 \sum_{t=0}^{H} \beta^t \geq |G|$. *Hypotheses.* (i) *Clean settled-stream changepoint*: from settlement index ν on, every settled peer record has conditional mean $\leq -g(s)$, $g(s) := -\mu_2(s) > 0$; automatic after a finite flush, while unbounded out-of-order delays make it a genuine restriction (Section F). (ii) A constant $r > 0$ with $r_t \geq r$ on $[T_{\mathrm{act}}, H]$ on the capture event. $N_\star$: smallest integer meeting the delay bound at $g(s)$; $\bar{T}_{\mathrm{act}}$: the deterministic serving epoch for index $\nu + N_\star$ plus publication lag; $G_{\mathrm{succ}} := r \sum_{t=\bar{T}_{\mathrm{act}}}^{H} \beta^t$.

Theorem 2 (Detection–action amplification)**.** *Fix s in the framing region and $\delta'' \in (0,1)$; let $\mathcal{E}$ (the* capture event*, probability $\geq 1 - \delta - \delta''$) be the event that the farmer is never downweighted (its stream lies in H_0; Theorem 1(c) plus detector safety) while the honest peer is downweighted by settlement index $\nu + N_\star$ and its weight transfers (so $T_{\mathrm{act}} \leq \bar{T}_{\mathrm{act}}$); the detector behaves* correctly*, H_0 being a property of the report profile rather than of any single agent. Then: on* every *sample path $\Delta\widehat{U}_1 - \Delta\widehat{U}_1^{\mathrm{alert}} = G$; on $\mathcal{E}$, $G \geq G_{\mathrm{succ}} \geq 0$, strict iff $H \geq \bar{T}_{\mathrm{act}}$; and, ex ante,*

$$\Delta U_1 - \Delta U_1^{\mathrm{alert}} = \mathbb{E}G \;\geq\; (1 - \delta - \delta'')\, G_{\mathrm{succ}} - (\delta + \delta'')\, C_{\mathrm{fail}},$$

strictly positive whenever $\delta + \delta'' < G_{\mathrm{succ}}/(G_{\mathrm{succ}} + C_{\mathrm{fail}})$ (which presupposes $H \geq \bar{T}_{\mathrm{act}}$; otherwise $G_{\mathrm{succ}} = 0$, the bound is trivial, and no ex-ante amplification is claimed). $\Delta U_1^{\mathrm{alert}}$ may take either sign: amplification is a coupled difference and carries no claim that alert-only play is unprofitable.
Absolute profitability, ex ante: *with $C_{\mathrm{farm}} = \sum_{k \in \mathcal{D}} A_k \frac{s^2}{4}$ the deviation's expected settlement cost over the settled deviation epochs $\mathcal{D}$ ($|\mathcal{D}| \leq N_\star + N_{\mathrm{dev}}$; N_{dev} counts post-capture deviation epochs, zero if the farmer reverts), profitability is equivalent to $\Delta U_1^{\mathrm{alert}} + \mathbb{E}G > 0$; under the* direct *hypothesis $R_{\mathrm{cont}}(s) \geq 0$ (the deviation's continuous-channel rent relative to truthful play; monotonicity alone is insufficient; the worked instance verifies it by the exact polynomial $0.4s - s^2 > 0$, Section F), $\Delta U_1^{\mathrm{alert}} \geq -C_{\mathrm{farm}}$, and $(1 - \delta - \delta'')\, G_{\mathrm{succ}} - (\delta + \delta'')\, C_{\mathrm{fail}} > C_{\mathrm{farm}}$ suffices.*

Proof kernel. Sample-for-sample coupling cancels settlement, continuous weights, and alarm times *exactly*, leaving the transfer rent G, an identity on every path. Splitting $\mathbb{E}G$ over the capture event and its complement ($|G| \le C_{\text{fail}}$) gives the ex-ante bound; absolute profitability is the identity in expectation plus $R_{\text{cont}} \ge 0$. Full proof: Section F. □

Amplification is a *horizon* statement: the worked instance's framing-plus-capture gain approaches $\sim$20$\times$ the continuous-farming gain as $(H - \nu - N_\star)/N_\star \to \infty$, $\beta \to 1$; for $H < T_{\text{act}}$ detection is simply correct. The defect is the *composition*: a sound alarm coupled to a rent-transferring action becomes the deviator's instrument; mitigation must target the coupling (Section 5), and improving the alarm alone does not remove it.

4.2 Exclusion is not a generic fix: cascades at $n \ge 3$

One candidate repair, *reference exclusion*, removes a detected agent from the settlement coalition too; it works at $n = 2$ under a fixed empty-coalition anchor and provably does *not* generalize (Section H): under the unweighted coalition mean, every $n \ge 3$ admits an explicit profile with invisible framing *and* an attacker-free cascade (removing every informative agent for $n \ge 4$), at margins $\Theta(1/n^2)$. Structurally, exclusion breaks correct specification (the surviving coalition's mean no longer equals p), so exclusion dynamics are not closed under their own semantics; Theorem 2's reallocation is itself a cascade surface. Breadth at $n = 3$ (exhaustive, exact expectations): invisible framing in 9,060 pairs, 96% recursion closure, 269 attacker-free cascades.

5 Restricting adaptive influence, and its price

Since the failure is unpriced influence, the direct remedy is to meter it: constrain the *published* weights to a trust region around uniform, $\hat{s} \in \Delta^{n-1} \cap \{\|\hat{s} - u\|_1 \le 2\tau\}$ (projection via Dykstra, closed form at $n = 2$; Section G). The cap gives $\mathrm{TV}(\hat{p}_{\text{serve}}, \bar{p}_{\text{ref}}) \le \tau$ for *every* report matrix, deviant or not.

Theorem 3 (Influence trust region)**.** *Assume the utility* (2) *under the counterfactual coupling of* (5) *(the deviation reaches the rent term only through the published weights). With $\mathcal{A}_t$ the affected epochs of Section 3.1, define the* lifetime exposure $\Lambda_{i,t} := \mathbb{E}\big[\sum_{u \in \mathcal{A}_t} \beta^u\big] < \infty$. *Assume* global *strong properness of the settlement in the geometry* (3)*: for every admissible h, $\mathbb{E}[\bar{\Delta}_i(0)] - \mathbb{E}[\bar{\Delta}_i(h)] \ge \frac{1}{2} h^\top Q\, h$, satisfied* with equality *by the Brier instance of* (1). *Then every deviation h that an optimizing agent weakly prefers to truthful satisfies*

$$\frac{A}{2}\, h^\top Q\, h \;\le\; 2\chi_i \tau\, \Lambda_{i,t},$$

$$\textit{equivalently} \qquad \|h\|_Q \;\le\; 2\sqrt{\chi_i \tau \Lambda_{i,t}/A},$$

a global and exact bound (no asymptotics in τ or h); binary Brier: $|c| \le \sqrt{2\chi_i \tau \Lambda_{i,t}/(A\mu_i)}$, $\mu_i = 1/n^2$. For non-quadratic scores, replace global strong properness by α_ϕ-strong convexity of the potential (log score: Pinsker, on the compact interior) and the bound holds in that norm.

Proof kernel. In $\mathcal{W}_\tau$ the zero-sum coordinate bound gives $|\hat{s}_i - \frac{1}{n}| \le \tau$ in both worlds, so the i-th weight differs by at most 2τ per affected epoch and exactly 0 elsewhere (coupling), capping the lifetime rent difference at $2\chi_i \tau \Lambda_{i,t}$; compared with the global settlement regret $\frac{A}{2} h^\top Q h$ gives the display. Full proof: Section G. □

For an L_i-Lipschitz served-output utility the same argument replaces χ_i by L_i (served-output corollary, Section G; no global equivalence invoked); the population sweep checks that form (81/81, Section J), E3 the χ-form (Section 6). $\Lambda_{i,t}$ is a *worst-case range exposure* rather than the differential kernel ω_i (they coincide only for one affected epoch and $m = 1$; Section B). The bound covers *every* weakly-improving deviation and interpolates Section 2.3's poles ($\tau = 0$ static uniform; $\tau \to \infty$ recovers Theorem 1(a)); between them the *report distortion* scales as $O(\sqrt{\tau})$ while the *incentive gap* closes as $O(\tau)$. The trust region caps exactly the adaptivity that tracks drifting agents (the price curve of Section 6 makes the cost explicit) and constrains only continuous movement; detector-triggered exclusion lies outside it.

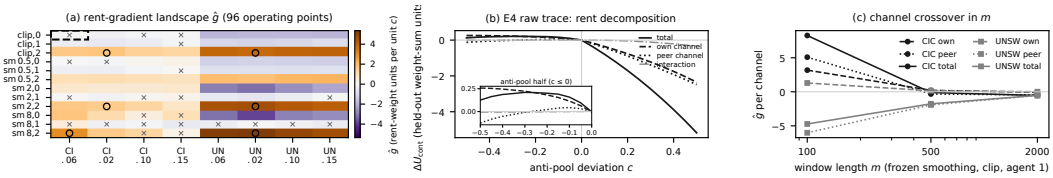

Figure 2: (a) R4 landscape: $\hat{g}$ at 96 operating points (rows: map (clip; sm $\tilde{\eta}$ = softmax at scaled temperature $\tilde{\eta}$) $\times$ deviating agent 0–2; columns: trace (CI = CICIDS2017, UN = UNSW-NB15) $\times$ report smoothing (A, B), $\tilde{r} = A + Br$; $\times$: fails the screen; rings: confirmed; dashed box: frozen E1 cell). (b) E4 rent decomposition; inset: the anti-pool half on its own scale. (c) Own/peer crossover in window length m. The claim is in (a); (c) explains the E1 cell's cancellation. E1/E3 panels: Section J.

A second design point: the dead-zone map of Section 3.1 returns as *watermark weighting* (R2), locally farming-proof in its population form (statement, proof, scope, and costs: Section G).

6 Validation and Transport

Theorem 1 is conditional, so these rounds test *existence* of a nonzero rent gradient, not universality. Preregistered on two deterministic network-intrusion trace pipelines (CICIDS2017 and UNSW-NB15, hereafter CICIDS and UNSW; offline, no deployment claimed): the *in-model bridge* E1–E3 (+S1/S2), *raw-trace diagnostics* (Proposition 1, E4), and the *operating-point round* R4 (Figure 2; E1's frozen criteria P1–P5 — envelope, population-direction rent, quadratic residuals, cushion silence — in Section J).

In-model: the envelope holds; the population prediction fails, and the failure is a measured channel cancellation. E1 smooths reports into the simplex interior and draws labels from the pooled posterior; the envelope (own-marginal stationarity at the truthful report) holds by construction (settlement quadratic to residuals $\le 2 \times 10^{-11}$). The identified population direction ($\psi^{\mathrm{tr}}_{\mathrm{pop}} = -0.129$; tr = training half) *does not transport*: rent is significantly negative in *both* directions ($\approx -730c^2$) and P1–P4 **fail**. The grid's minimum detectable slope 19.1 far exceeds the predicted 2.6; the label-conditional Rao–Blackwell estimator, 37$\times$ more precise, resolves it into opposing own $+0.36$ $[0.08, 0.65]$ and peer -0.31 $[-0.47, -0.14]$ channels that nearly cancel (paired total $+0.06$ $[-0.31, +0.42]$, excluding the population-scale -2.58; frozen window at the crossover, Figure 2c); within the cushion the tested detector stays silent (P5 **passes**; run-level, no undetectability proof).

The cancellation is knife-edge: nondegeneracy is confirmed at selected operating points (R4). R4 screens 96 operating points, 73 passing; all six top-ranked cells — including the *deployed* clip map on both traces — confirm on disjoint seeds at Bonferroni 99.17% CIs (UNSW clip $+3.82$ $[3.75, 3.88]$; CICIDS $+1.79$ $[1.66, 1.92]$; full grid: Figure 2a): (S3) is confirmed where it was tested at this pool size, not established grid-wide; E1 is where the channels cancel. All six are the *same* deviating agent; at n=5 the confirmed UNSW clip cell's CI covers zero, CICIDS stays positive (Section J). At the preregistered sizes no cell is profitable at $\kappa = 1$ (κ weights rent in $\Delta U_{\mathrm{total}} = \Delta U_{\mathrm{settle}} + \kappa \Delta U_{\mathrm{rent}}$): settlement's quadratic cost leaves the UNSW totals significantly negative (breakeven κ^* 1.3–4.4) and the CICIDS totals within noise of zero. A micro-round frozen after that negative result (design dependence disclosed) at the total-curvature optimum $c^* = \hat{g}/(2\hat{K})$, on fresh seeds, is *profitable* at $\kappa = 1$ on all three primary cells (clip: $\Delta U_{\mathrm{total}} = +0.0014$, 98.33% CI $[+0.0007, +0.0020]$, $c^* = 7 \times 10^{-4}$), the detector silent on every stream (0/600): the profitable branch, in-model (dead-zone control: machine zero).

Framing: null on the symmetric world; realized end-to-end on an engineered one. E2's frozen search finds *no framing region*; one engineered asymmetric world exhibits it with causal controls, that deviation's drift (0.0038) below detector power at 20k. On a *second* such world, selected for in-power drift (0.0052) and run as a preregistered extended-horizon replay (tiled reports, fresh labels; T=16), the loop closes: the detector fires on the honest victim on every path (median delay 126k), never on a truthful stream or the deviator's own (0/25); the deviator inherits the excluded

weight ($G = +0.133$ $[0.127, 0.139]$; Section J). E3's frontier is flat (distortion ≤ 0.005; bound 9/9; failed remedy gate: Section J).

Adapted baselines, matched protocol. Matched worlds, labels, deviation family and κ=1. Porting changes belief and utility, so these measure *transfer*, not the algorithms' guarantees (Section J): original WSU reduces but does not remove the channel; ELF-X's secant exceeds clip's in five of six cells, resolved on UNSW (paired -4.4 $[-5.0, -3.7]$), while serving better; Krum (n=5 only) gives a degenerate zero — this deviator never wins — though a post-hoc probe on a winner peaks at truthful.

Raw trace: the model's own assumption fails, not the threat.

Proposition 1 (Transport moment certificate)**.** *For any reports-measurable direction $(d_{it})_t$, $r_{it}(c) = p_{it} + c\, d_{it}$, settlement* (1) *gives $\frac{d\,\pi_i(c)}{dc}\Big|_{c=0} = -\frac{2\alpha}{n}\sum_t (\bar{p}_t - y_t)\, d_{it}$, and correct specification implies $\mathbb{E}[(\bar{p}_t - Y_t)d_{it}] = 0$; a nonzero estimate whose CI excludes zero is a* transport-failure certificate *(proof: only $v_t(N)$ depends on c; Section I).*

Endpoint-locked: -15.48 (95% epoch-block CI $[-18.95, -12.39]$; unmasked: Section I); UNSW-NB15: -78.8 $[-84.2, -73.3]$; all 12 preregistered segments agree (Section J). Transport fails on both traces, convention-independently: the assumption is what breaks.

E4: outside that assumption, a detector-silent, settlement-led, welfare-neutral deviation. The anti-pool deviation on the frozen grid ($c \in [-0.5, 0.5]$, both maskings) raises *both* ledgers, settlement first: at $c = -0.2$ (frozen rule) settlement rises $+2.07$, more than ten times the rent $\Delta U_{\text{cont}} = +0.200$ $[0.03, 0.35]$; served quality is unchanged (ΔBrier $\leq 3 \times 10^{-4}$) and the tested detectors never alarm (1.26M agent-samples; $B = 0.1$ too). The single-intervention decomposition is own-led and an affine replay collapses the own term $125\times$ (the deployed map's Jacobian response, not equated with the in-model covariance channel). The envelope fails on this trace, so $d_i = 0$ is not in force (zero-point rent gradient $+4.8$ $[4.0, 5.5]$); the own-led rent is an out-of-model *deployment vulnerability*, not a peer-channel validation (prespecified, outside P1–P5; Section J).

7 Related work, limitations, and conclusion

Related work. Proper scoring makes truth-telling a stage optimum (Gneiting & Raftery, 2007; Abernethy & Frongillo, 2012); a non-affine current-decision approval payoff defeats every strictly proper score (Lovén & Tarkoma, 2026), a channel we exclude (Section 2.2). We add the *delayed allocation state*: a mean first-order peer channel inside every own-stream composite null (Theorem 1c,d) and a detector–action composition that amplifies and cascades (Theorem 2); unlike market-scoring manipulation (Hanson, 2007; Chen et al., 2007), the rent needs no trade. Elicitation without verification (Miller et al., 2005; Dasgupta & Ghosh, 2013), incentive-compatible learning (Dekel et al., 2010; Braverman et al., 2017) and wagering (Lambert et al., 2008; Luo et al., 2026) price the report; dynamic pivot and Myersonian mechanisms (Bergemann & Välimäki, 2010; Pavan et al., 2014; Chambers & Lambert, 2021; Landolt et al., 2026) price future information rents under full protocol control; ours is the complement, a stagewise rule leaving it unpriced. Performative prediction (Perdomo et al., 2020; Boeken et al., 2025) moves the data distribution; our labels are frozen. IC online expert learning (Roughgarden & Schrijvers, 2017; Freeman et al., 2020; Frongillo et al., 2021; Sadeghi & Fazel, 2023; Komiyama et al., 2025) and Byzantine-robust aggregation (Blanchard et al., 2017) target their own objectives; we measure transfer, not those guarantees (Section 6). Reputation gaming is documented (Christiano, 2016; Pejó et al., 2026; Jaramillo-Velez et al., 2026); our detector is an e-process (Ville, 1939; Howard et al., 2021; Waudby-Smith & Ramdas, 2024; Shin et al., 2024; Lorden, 1971).

Limitations. Open: sound exclusion semantics (Section 4.2); strategic timing or label tampering; a certificate naming the failing hypothesis; incentivized evaluation streams; pricing that survives pool growth.

Conclusion. Delayed settlement turns adaptive allocation into unpriced intertemporal influence: stagewise truthful payments need not compose (Theorem 1), coupled detection can miss or amplify it (Theorem 2), restriction bounds the best response (Theorem 3). Pricing it remains open.

AI USE STATEMENT

Large language models were used substantially and interactively throughout this work, under continuous human direction and review; per the ICLR policy, we disclose by task. *Developing theoretical models and conceptual frameworks:* the formal model (exactly-once delayed settlement, the dynamic rent functional, the operator-sum kernel) was developed in AI-assisted sessions: the authors set the research question, the modeling choices, and the correctness rulings; the AI drafted formalizations that the authors repeatedly corrected — several load-bearing formulations drafted by the AI (an incorrect map-independence claim, a mis-scoped weak-externality expansion, a conflation of high-probability and expected utility, an overstated detector-delay necessity claim) were found and ordered fixed by the authors during review, and the corrected forms appear in the paper. *Formulating mathematical claims and proving them:* theorem statements and proofs were drafted by the AI from author-approved proof strategies and internal derivation notes, then audited by the authors line by line; every closed form carries an independent machine-verification script (Section J), and the interfaces between body statements and appendix proofs were cross-checked in dedicated audit passes. *Proposing hypotheses, designing methodology:* the τ-frontier experiment, the E1–E4 bridge and audit, the S1/S2 supplement, the R4 operating-point round (with its follow-up micro-round), the fifth follow-up series (end-to-end exclusion loop, decoupling analysis, segment certificates), and the defense-baseline round (selection maps, allocation-side monitor, collusion) were AI-drafted and author-frozen as preregistrations before execution, with one pre-outcome amendment and two post-result presentation decisions, all author-ruled and committed (Section J). *Implementation:* verification and experiment scripts were AI-written and author-reviewed; all numerical checks are deterministic given pinned seeds. *Interpreting results:* joint, with the authors making all final calls, including the decisions to report the failed remedy gate, the E1 criteria failures, and the transport-failure certificate as negative results. *Writing:* the text was AI-drafted and human-edited through multiple adversarial review rounds. *Figures:* every figure that reports a measurement is generated by the deterministic scripts above from the frozen result files. The schematic Figure 1 is the exception: its three icons are decorative line art produced by an image-generation model from an author-written prompt, then cropped and composited by an author-reviewed script. Every word, formula, rule, and arrow in that figure is vector, written by the authors and added afterwards; no text in it came from the model. The icons encode no measured quantity, and the prompts, crops, and model are recorded in the artifact. *Qualitative and thematic analysis:* none is performed on data. The only qualitative judgments in the paper are the characterizations of prior work in Section 7; those were AI-drafted and checked by the authors against the cited sources. *Data:* the datasets are the public CICIDS2017 and UNSW-NB15 traces, processed by deterministic code; E1's semi-synthetic labels are drawn by a frozen, seed-pinned Bernoulli sampler from the pooled posterior, and that sampling code, like the other experiment code, was AI-written and author-reviewed. Beyond this, AI was not used to generate data, to translate, or to clean data. The authors take full responsibility for all content.

REPRODUCIBILITY STATEMENT

All numbered theorems, propositions, and the lemmas they use carry complete proofs in the appendix (one unnumbered average-run-length aside is marked as a heuristic there), with every assumption stated in the main text. Closed forms are machine-verified by the scripts inventoried in Section J (exact-rational and symbolic-n certificates, finite-difference checks to stated tolerances, 16-path exact enumeration for the dynamic interfaces). All experiment rounds — the τ-frontier, the E1–E4 bridge, the S1/S2 supplement, the R4 operating-point round with its follow-up, the fifth follow-up series, and the defense-baseline round — were preregistered before execution (designs, grids, metrics, stop rules, the remedy gate, and the P1–P5 criteria, with one pre-outcome amendment, two post-result presentation decisions, and one follow-up micro-round whose design responds to a preceding negative result, that dependence disclosed; all committed; Section J); the trace arms use the public CICIDS2017 and UNSW-NB15 datasets through a deterministic, seed-pinned pipeline whose sample protocol is spelled out in Section J, and the transport certificates are reproduced by a closed form and an independent replay-derivative ladder that agree to 3×10^{-6} relative (Section I). An anonymized artifact with code, seeds, and environment pins accompanies the submission.

REFERENCES

Jacob D. Abernethy and Rafael M. Frongillo. A characterization of scoring rules for linear properties. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23 of *Proceedings of Machine Learning Research*, pp. 27.1–27.13, 2012.

Dirk Bergemann and Juuso Välimäki. The dynamic pivot mechanism. *Econometrica*, 78(2):771–789, 2010.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 119–129, 2017.

Philip Boeken, Onno Zoeter, and Joris M. Mooij. Conditional forecasts and proper scoring rules for reliable and accurate performative predictions. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

Mark Braverman, Jieming Mao, Jon Schneider, and S. Matthew Weinberg. Multi-armed bandit problems with strategic arms. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pp. 383–416, 2017.

Christopher P. Chambers and Nicolas S. Lambert. Dynamic belief elicitation. *Econometrica*, 89(1): 375–414, 2021.

Yiling Chen, Daniel M. Reeves, David M. Pennock, Robin D. Hanson, Lance Fortnow, and Rica Gonen. Bluffing and strategic reticence in prediction markets. In *Internet and Network Economics (WINE)*, pp. 70–81, 2007.

Paul Christiano. Provably manipulation-resistant reputation systems. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 670–697. PMLR, 2016.

Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pp. 319–330, 2013.

Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

Rupert Freeman, David M. Pennock, Chara Podimata, and Jennifer Wortman Vaughan. No-regret and incentive-compatible online learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3270–3279, 2020.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Rafael Frongillo, Robert Gomez, Anish Thilagar, and Bo Waggoner. Efficient competitions and online learning with strategic forecasters. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 479–496, 2021. doi: 10.1145/3465456.3467635.

Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 2242–2251, 2019.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Robin Hanson. Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets*, 1(1):1–15, 2007.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

Delio Jaramillo-Velez, Gergely Biczók, Alexandre Graell i Amat, and Balázs Pejó. Beyond leave-one-out: Private and robust contribution evaluation in federated learning. *Intelligent Systems with Applications*, 32:200725, 2026. doi: 10.1016/j.iswa.2026.200725.

Junpei Komiyama, Nishant A. Mehta, and Ali Mortazavi. No-regret incentive-compatible online learning under exact truthfulness with non-myopic experts. In *Proceedings of the 26th ACM Conference on Economics and Computation*, 2025. URL https://arxiv.org/abs/2502.11483.

Nicolas S. Lambert, John Langford, Jennifer Wortman Vaughan, Yiling Chen, Daniel M. Reeves, Yoav Shoham, and David M. Pennock. Self-financed wagering mechanisms for forecasting. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC)*, pp. 170–179, 2008.

Leo Landolt, Anna Maria Maddux, Andreas Schlaginhaufen, Saurabh Vaishampayan, and Maryam Kamgarpour. Eliciting truthful feedback for preference-based learning via the VCG mechanism. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5212–5220, 2026.

Gary Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, 1971.

Lauri Lovén and Sasu Tarkoma. The endogeneity of miscalibration: Impossibility and escape in scored reporting, 2026. arXiv:2605.07671, v2 (September 2026).

Yuhong Luo, David M. Pennock, and Xintong Wang. Decentralized aggregation of LLM predictions via wagering mechanisms. arXiv:2607.04389, 2026.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2): 583–601, 2002.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

Alessandro Pavan, Ilya Segal, and Juuso Toikka. Dynamic mechanism design: A Myersonian approach. *Econometrica*, 82(2):601–653, 2014.

Balázs Pejó, Marcell Frank, Krisztián Varga, Péter Veliczky, and Gergely Biczók. On the fragility of contribution evaluation in federated learning. In *2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 171–187, 2026. doi: 10.1109/satml68715.2026.00019.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609, 2020.

Tim Roughgarden and Okke Schrijvers. Online prediction with selfish experts. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Omid Sadeghi and Maryam Fazel. No-regret online prediction with strategic experts. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/ab9f9cfe97da3665e08f50ade9f8c4d6-Abstract-Conference.html.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games, Volume II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, 1953.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: A nonparametric framework for sequential change detection. *The New England Journal of Statistics in Data Science*, 2(2): 229–260, 2024.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Zhaohui Geoffrey Wang. When individually calibrated models become collectively miscalibrated. In Siddharth Swaroop, David Rügamer, and Agustinus Kristiadi (eds.), *Proceedings of The 1st Symposium on Probabilistic Machine Learning*, volume 327 of *Proceedings of Machine Learning Research*, pp. 214–255. PMLR, 2026. URL https://proceedings.mlr.press/v327/wang26a.html.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

APPENDIX ROADMAP

Every appendix section below is either a proof the body states without proving, or the frozen design and full output of an experiment the body summarizes. Nothing here introduces a claim that the body does not make. Reviewers checking one body result can go straight to its row.

§	Contents	Supports
A	Separability lemma, worked instance, price of honesty	the no-stake half of Theorem 1
B	Influence–externality units; operator-sum kernel K' and its stationary scalar corollary	the rent gradient of Theorem 1; Theorem 3's exposure Λ
C	Proof of Theorem 1; proportional corollary; cushion construction	Theorem 1(a)–(d)
D	Softmax weight response and its nondegeneracy condition	the softmax rows of Figure 2a
E	Restart-mixture e-detector: safety, grid, delay	the detector used throughout Sections 4 and 6
F	Proof of Theorem 2	framing, exclusion, amplification
G	Trust-region bound; watermark construction	Theorem 3
H	$n = 2$ proposition; all-n cascade family; $n = 3$ sweep	the cascade claim at every $n \geq 3$
I	Endpoint-locked deployed report map	the transport certificates of Proposition 1
J	Preregistrations, per-round results, the claim–evidence–scope table, machine-verification inventory	all of Section 6

Code, seeds, environment pins, and the preregistrations verbatim as frozen are in the supplementary artifact; its `README.md` maps each script to the result file it produces.

A SEPARABILITY LEMMA AND THE WORKED INSTANCE: FULL PROOFS

Proof of the Separability lemma. Step 1 (discount factorization). A-delay makes the settlement delay D_t independent of all reports *and* all labels, so for each t, $\mathbb{E}[\beta^{t+D_t}\pi_{i,t}] = \mathbb{E}[\beta^{t+D_t}] \cdot \alpha\,\mathbb{E}[\bar{\Delta}_i(r_{\cdot,t}, y_t)]$: the discount factors off the settled marginal.

Step 2 (exactly-once separability). By EO2, each transfer $\pi_{i,t}$ is a function of the frozen tuple $(r_{\cdot,t}, y_t)$ alone (later reports, weights, and serving decisions never enter it) and by EO1 it is credited exactly once. By EO3 the weight trajectory touches only serving decisions, and serving decisions enter U_i only through the allocation stake, at scale χ_i; with $\chi_i = 0$ that term vanishes identically, so $U_i = \alpha \sum_{t \leq H} \mathbb{E}[\beta^{t+D_t}]\,\mathbb{E}[\bar{\Delta}_i(r_{\cdot,t}, y_t)]$, a sum whose t-th term depends on agent i's strategy only through $r_{i,t}$.

Step 3 (termwise maximization). For $\beta < 1$ with uniformly bounded per-epoch settled mass, or for $H < \infty$, the sum converges absolutely, so a behavioral strategy maximizes it iff it maximizes every term on every reachable history. Conditional on any history, with peers truthful, the t-th term is strictly maximized at $r_{i,t} = b_{i,t}$; the stagewise-truthfulness hypothesis (truthful component reporting completes the pooled posterior, Section 2.3).

Step 4 (strictness qualification). Any *report-altering* deviation — one whose reports differ from truthful on some reachable (i, t) with positive probability — strictly loses that term conditional on

reaching it while changing no other term, hence strictly loses in U_i; deviations that never alter a report (e.g., payoff-irrelevant randomizations) are payoff-equivalent to truthful, which is the precise sense of "strict." History-dependent strategies are covered because the argument is conditional-on-history, term by term. Note what is *not* used: no restriction on the weight map at all; an agent with no allocation stake cannot be tempted by any weight dynamics. □

The worked instance: exact closed forms. The instance of Section 3.2: $n = 2$, two decision periods, components $b = (0.3, 0.7)$, $y_t \sim \mathrm{Bern}(p)$ i.i.d. with $p = \frac{1}{2} = \bar{p}$ (the correctly-specified pooled posterior), label delay one period (period-1 settlement sets the weights serving period 2), $\beta = 1$ in the per-report normalization. Affine weight map $\hat{s}_1^{(2)} = \frac{1}{2} + \eta(\hat{\bar{\Delta}}_1 - \hat{\bar{\Delta}}_2)$ on expected marginals; because the map is *affine*, the decoupled evaluation of Section B is exact and no finite-m caveat arises in the rent computation. Externality billed on period-2 serving only ($g_1 = -\lambda \hat{p}_2$; period-1 decisions carry no utility, isolating the dynamic channel). Agent 2 truthful throughout, agent 1 truthful in period 2; the deviation is $r_1 = b_1 + s$ in period 1.

With $\mathbb{E}\, v(S) = -(\bar{r}_S - p)^2 - p(1-p)$:

- *Stage settlement.* $\mathbb{E}\bar{\Delta}_1(s) = \mathbb{E}v(\{1,2\}) - \mathbb{E}v(\{2\}) = 0.04 - \frac{s^2}{4}$: derivative zero and second derivative $-\frac{1}{2} < 0$ at $s = 0$, so truthful is a strict interior stage optimum with $\mu_1 = \frac{1}{4}$ ($= 1/n^2$), stage loss $\alpha s^2/4$ — the envelope fact, exactly.
- *Weight shift.* $\hat{\bar{\Delta}}_1 - \hat{\bar{\Delta}}_2 = \mathbb{E}v(\{1\}) - \mathbb{E}v(\{2\}) = 0.4s - s^2$, so $\hat{s}_1^{(2)} = \frac{1}{2} + \eta(0.4s - s^2)$: the profitable direction $s > 0$ moves the report from $b_1 = 0.3$ *toward* the pooled posterior $p = \frac{1}{2}$.
- *Served output.* With both truthful in period 2, $\hat{p}_2 = \hat{s}_1 b_1 + (1 - \hat{s}_1) b_2 = \frac{1}{2} - 0.4\eta(0.4s - s^2)$: a higher own weight drags the served aggregate toward b_1, which is what the externality rewards.
- *Total gain.* Every quantity is an exact polynomial in s:

$$U_1(s) - U_1(0) \;=\; -\tfrac{\alpha}{4}s^2 \;+\; 0.4\,\lambda\eta\,(0.4\,s - s^2), \qquad s^* = \frac{0.08\,\lambda\eta}{\frac{\alpha}{4} + 0.4\lambda\eta},$$

  with $U_1'(0) = 0.16\lambda\eta > 0$ and strictly positive gain $(\frac{\alpha}{4} + 0.4\lambda\eta)(s^*)^2$ at the optimum (numerically $s^* \approx 0.1231$, gain $\approx 9.85 \times 10^{-3}$ at $\alpha = \eta = \lambda = 1$, against a stage loss $\approx 3.8 \times 10^{-3}$ at that shade).

In the weak-externality regime the optimum linearizes, $s^* = \chi_1\psi_1/(2A\mu_1) \cdot (1 + O(\lambda\eta/\alpha))$ with $\chi_1 = 0.4\lambda$, $\psi_1 = 0.4\eta$, $\mu_1 = \frac{1}{4}$, $A = \alpha$ — the ε-family of Theorem 1(b) realized exactly, with $\varepsilon = \lambda$ the rent scale and the $O(\lambda\eta/\alpha)$ factor the instance's $o(\varepsilon)/\varepsilon$ term. Read in reverse, the instance prices honesty: enforcing $|s^*| \le \varepsilon'$ requires $\alpha \gtrsim \chi_1\psi_1/(2\mu_1\varepsilon')$, so the settlement budget is the price of ε'-truthfulness. Genericity: the first-order term needs only $b_1 \neq b_2$ and a nonvanishing map derivative at the operating point; the degenerate cancellations are measure-zero in (b, η). All closed forms machine-verified (`verify_counterexample.py`, Section J).

B Lemma K: influence–externality units, and the influence kernel

Spaces and notation. Deviations live on the tangent space $\mathcal{T} = \{h \in \mathbb{R}^K : \mathbf{1}^\top h = 0\}$. For a fixed deviation epoch t: $\Delta^{(t)}(h) \in \mathbb{R}^n$ is the vector of sample-t settled marginals when agent i reports $b_i + h$ and everyone else is truthful; its response matrix $\nabla_h \Delta^{(t)}(h) \in \mathbb{R}^{n \times K}$ has rows $\nabla_h \Delta_j$, with means at $h = 0$ the deterministic $d_j \in \mathcal{T}$ of (4) and $d_i = 0$ (envelope). $\hat{\bar{\Delta}}_e^W \in \mathbb{R}^n$ is the realized m-window statistic feeding publication e; the weight map $\varphi_i : \mathbb{R}^n \to [0,1]$ has row Jacobian $D\varphi_i(\cdot) \in \mathbb{R}^{1 \times n}$, so $D\varphi_i(\cdot)\, \nabla_h \Delta^{(t)} \in \mathbb{R}^{1 \times K}$, transposed into $\mathcal{T}$ after projection. $\mathcal{A}_t = \{u > t : t \in W(e(u))\}$ are the affected epochs.

Symbol-reuse conventions. A few letters carry two standard meanings, disambiguated by their arguments and context; we state the convention once. μ_i without an argument is the scalar settlement curvature ($1/n^2$ for the Brier instance); $\mu_i(s)$ with a strength argument is agent i's mean settled marginal under a deviation of strength s (Section 4). A (roman, in (S1)) is the discounted settlement scale, A_k its per-epoch version; "A-delay" names the delay assumption; $\mathcal{A}_t$ (calligraphic) the

affected epochs. $r_{i,t}$ with indices is a report; r_t in Theorem 2 is the per-epoch rent differential, and r alone its lower bound. B is the detector range parameter; in bootstrap contexts ($B = 2000$ resamples, Section J) it is the resample count. α is the settlement price scale; the strong-convexity modulus of Theorem 3 is always written α_ϕ. λ_j with a grid index is a detector grid point; λ alone in the worked instance is the externality scale. H is the horizon; H_0, H_1 are hypotheses; H_{k-1} a history σ-field. m is the settlement window length; detector windows are powers 2^ℓ; the symbolic cascade offset is $n = 3 + m'$ with a primed letter.

Lemma (K: influence–externality units)**.** *Let the primitive externality of agent i be a per-epoch business value $g_i : \Delta^{K-1} \to \mathbb{R}$ of the served distribution, differentiable at the truthful operating point $\hat{p}^0$ and L_i-Lipschitz in total variation. Fix an* allocation direction $v \in \mathbb{R}^n$*: own weight up at unit rate ($v_i = 1$), peers shifted by v_j, $\sum_j v_j = 0$. Then the served output moves at rate $\sum_j v_j b_j$ and the per-unit-own-weight rent coefficient* along v *is the directional derivative*

$$\chi_i(v) \;=\; Dg_i(\hat{p}^0)\Big[\textstyle\sum_j v_j b_j\Big], \qquad |\chi_i(v)| \;\le\; L_i \,\big\|\textstyle\sum_j v_j b_j\big\|_{\mathrm{TV}},$$

with equality in the bound iff $Dg_i(\hat{p}^0)$ attains its Lipschitz constant along that direction (in particular for linear, direction-aligned g_i); Lipschitzness alone does not imply saturation. For a pure own-vs-peers transfer with debit profile $\tilde{w}_j \ge 0$, $\sum_{j\neq i} \tilde{w}_j = 1$, this reads $\chi_i = Dg_i(\hat{p}^0)[b_i - \bar{b}^w_{-i}]$, $\bar{b}^w_{-i} = \sum_{j\neq i} \tilde{w}_j b_j$. The coefficient is direction-dependent*, and the direction is induced by the weight map: at $n \ge 3$ different maps debit different peers and generally produce different $\chi_i(v)$. It is map-independent exactly when the direction is forced — at $n = 2$, where the unique peer takes the full debit, or when the mechanism fixes the peer-debit proportions $\tilde{w}_j$ across maps. In* (2)*, χ_i is the reduced-form primitive evaluated at the mechanism's own induced direction.*

Proof. The served output is $\hat{p} = \sum_j \hat{s}_j r_j$; a zero-sum shift $\delta w = cv$ gives $\delta\hat{p} = c\sum_j v_j b_j$ at the truthful profile. The rent rate per unit of own weight is $Dg_i(\hat{p}^0)$ applied to that direction; the bound is TV–Lipschitz duality. Direction dependence: a C^1 map with per-coordinate response $\beta^{(i)}_j = \frac{d}{dc}\varphi_j(\bar{\Delta}(c))|_0$ (so $\sum_j \beta^{(i)}_j = 0$ by normalization) induces the direction $v_j = \beta^{(i)}_j/\beta^{(i)}_i$, and matching the reduced form $\chi_i\beta^{(i)}_i$ gives $\chi_i = Dg_i[\sum_j \beta^{(i)}_j b_j]/\beta^{(i)}_i = \chi_i(v)$ — which varies with the map through v whenever $n \ge 3$ leaves the debit split free. At $n = 2$ the direction is forced, $v = (1, -1)$ up to indexing, and χ_i is the same for every normalized map; the worked instance's check is exactly this $n = 2$ statement (binary, λ-scaled linear g): affine map $\beta^{(1)} = (0.4, -0.4)$, rent rate 0.16λ, $\chi_1 = 0.4\lambda$; proportional map $\beta^{(1)} = (2.5, -2.5)$, rent rate 1.0λ, $\chi_1 = 1.0\lambda/2.5 = 0.4\lambda$ — same χ because $n = 2$, *not* evidence of general map independence (Section J). □

Lemma (K′: exact operator form of the rent gradient)**.** *Assume, in addition to (S1):*

- ***(K1)*** (Finite exposure.) $\mathbb{E}\big[\sum_{u\in\mathcal{A}_t} \beta^{\,u}\big] < \infty$ *(sufficient: $\beta < 1$ with boundedly many affected epochs per sample, or $H < \infty$; we do* not *assume a sample enters boundedly many windows in general).*
- ***(K2)*** (Regular map.) *$\varphi_i \in C^1$ on a neighborhood of the a.s. range of the truthful window statistics, with* $\sup \|D\varphi_i\| \le C_\varphi < \infty$ *there.*
- ***(K3)*** (Dominated response.) *Reports and labels range over compact sets and each $h \mapsto \Delta^{(t)}_j(h)$ is C^1 with $\|\nabla_h \Delta^{(t)}\| \le C_\Delta$ uniformly (immediate for the Brier instance: polynomial on a compact domain).*

Then $\mathcal{R}_i$ of (5) *is differentiable at* 0 *and*

$$a_i = \nabla_h \mathcal{R}_i(0) = \chi_i\, \mathbb{E}\Big[\textstyle\sum_{u\in\mathcal{A}_t} \frac{\beta^{\,u}}{m}\, D\varphi_i\big(\hat{\bar{\Delta}}^{W,0}_{e(u)}\big)\, \nabla_h \Delta^{(t)}(0)\Big]^\top\Big|_{\mathcal{T}},$$

the expectation joint over the window contents and sample t's randomness — exactly (6)*, with $J_{\varphi_i,u}$ read as the realized-statistic Jacobian.*

Proof. $\hat{s}^{(e(u))}_i(h) = \varphi_i(\hat{\bar{\Delta}}^W_{e(u)}(h))$, and $\hat{\bar{\Delta}}^W_{e(u)}$ depends on h only through sample t's marginals, with window-average coefficient $\frac{1}{m}\mathbf{1}\{t \in W(e(u))\}$; the chain rule gives, path by path, $\nabla_h \hat{s}^{(e(u))}_i(0) =$

$\frac{1}{m}\mathbf{1}\{t \in W(e(u))\}\, D\varphi_i(\hat{\bar{\Delta}}^{W,0}_{e(u)})\, \nabla_h \Delta^{(t)}(0)$. By (K2)–(K3) each term is bounded by $C_\varphi C_\Delta/m$, so the series $\sum_u \beta^u(\hat{s}_i(h) - \hat{s}_i(0))$ has the summable envelope $2C_\varphi C_\Delta \sum_u \beta^u \mathbf{1}\{u \in \mathcal{A}_t\}$, finite in expectation by (K1); dominated convergence licenses both exchanges (∇_h with $\mathbb{E}$, and with the a.s.-finite sum). Multiply by χ_i. □

Lemma (Decoupled evaluation)**.** *The two random factors above share sample t and are dependent; the factorized form $\mathbb{E}[\cdots] = \sum_u \mathbb{E}[\frac{\beta^u}{m}\mathbf{1}\{t \in W\}]\, J^{\text{mean}}_{\varphi_i,u}\, d^{(t)}$ with $J^{\text{mean}}_{\varphi_i,u} = D\varphi_i(\bar{\bar{\Delta}}^0_{e(u)})$ evaluated at the* mean *statistic and $d^{(t)}$ the mean response* (4) *is: (i)* exact *when $D\varphi_i$ is constant on the statistics' range (affine maps; the worked instance and the population arm); (ii) in general off by the remainder $\sum_u \mathbb{E}\big[\frac{\beta^u}{m}\mathbf{1}\{t \in W\}(D\varphi_i(\hat{\bar{\Delta}}^{W,0}) - D\varphi_i(\bar{\bar{\Delta}}^0))\nabla_h\Delta^{(t)}\big]$, which collects* two *effects: an evaluation (Jensen) term $\mathbb{E}[D\varphi_i(\hat{\bar{\Delta}}^{W,0})] - D\varphi_i(\bar{\bar{\Delta}}^0)$ from the nonlinear map at a random statistic, and a covariance term between the realized Jacobian and the sample-t response (they share sample t). Only under explicit concentration hypotheses ($D\varphi_i$ Lipschitz, uniformly bounded responses, and window entries that are i.i.d. (or suitably mixing) settled samples) is the remainder $O(m^{-1/2})$* in absolute (norm) terms *(window noise $O(m^{-1/2})$ per publication, uniformly over the affected publications under (K1)); a relative rate would additionally require the exact value to stay bounded away from zero, which we do not assume, and absent such hypotheses we claim no rate. At small m the remainder can* dominate*: in the verification toy (softmax map, $m = 2$) the exact operator value is $+0.0122$, the factorized value $+0.0402$, absolute gap 0.0280 (229% of the exact value; a meaningful relative figure only because the exact value is nonzero here), while the operator form matches the exact derivative to 3×10^{-12}. The operator form is the primary object; the factorized form is a large-m or affine-map convenience (*`verify_dynamic_interfaces.py`*; an implementation check of this appendix's formulas rather than empirical evidence for the lemma).*

Corollary (K′-stationary: the scalar influence kernel)**.** *Two stationarity hypotheses must be kept distinct.* (i) Deterministic stationarity: *if $J_{\varphi_i,u} \equiv J^0_{\varphi_i} = (\partial\varphi_i/\partial\bar{\Delta}_j)_j$ is one* deterministic *operating-point Jacobian for every affected publication (the locally stationary truthful profile), it factors out of the joint expectation directly (no decoupling hypothesis is needed) and the sum factors through the scalar influence kernel*

$$a_i = \omega_i\, \chi_i \sum_{j \neq i} \frac{\partial \varphi_i}{\partial \bar{\Delta}_j}\, d_j, \qquad \omega_i = \mathbb{E}\Big[\tfrac{1}{m}\textstyle\sum_{u \in \mathcal{A}_t} \beta^u\Big] > 0, \tag{7}$$

so cadence and window length rescale a_i without rotating it. (ii) Shared random Jacobian: *if the affected publications merely share one* random *Jacobian, sharing alone does not license the exchange — factoring additionally requires the decoupling of the previous lemma (independence from the sample-t response, or a controlled covariance remainder). Without either form of the hypothesis, different publications contribute differently-oriented terms and no scalar kernel exists; the dichotomy of Theorem 1(a) is then judged on the total gradient. Moreover $\omega_i = \Lambda_{i,t}$ (the lifetime exposure of Theorem 3)* only *when a single epoch is affected and $m = 1$; for $m > 1$, ω_i carries the $1/m$ attenuation and $\Lambda_{i,t}$ does not.*

Timescale consistency. In the two-period normalization of the examples (one publication, window $m = 1$, rent billed on the next epoch), $\mathcal{A}_t = \{t+1\}$ and $\omega_i = \beta^{t+1}$: (7) is the bare peer Jacobian. On the settlement side the deviation report is costed exactly once, at scale $A = \alpha\,\mathbb{E}[\beta^{t+D_t}]\times$ per-epoch settled mass; both terms of Theorem 1's expansion are therefore per deviation report, and no discount appears twice. Numerical verification of the operator form, the decoupling error, the stationary factorization, and the $\omega \neq \Lambda$ distinction: `verify_dynamic_interfaces.py` (Section J).

C Proofs of Theorem 1 and the proportional corollary

Lemma (M1: marginal geometry)**.** *Under (S1) and the pooled-posterior model, for a deviation $h \in \mathcal{T}$ by agent i (others truthful; $w_j := b_j - \bar{p} \in \mathcal{T}$):*

(i) $\mathbb{E}\bar{\Delta}_i(h) = \bar{\Delta}^0_i - D_\phi(\bar{p}, \bar{p} + h/n)$, *with* $\bar{\Delta}^0_i = D_\phi(\bar{p}, \bar{b}_{-i}) \geq 0$;

(ii) $\mathbb{E}\bar{\Delta}_i(h) = \bar{\Delta}^0_i - \frac{1}{2}h^\top Q h + o(\|h\|^2)$ *with* $Q = \nabla^2\phi(\bar{p})|_{\mathcal{T}}/n^2$, *and* no remainder *for the Brier instance (globally quadratic);*

(iii) $d_j = \nabla_h \mathbb{E}\bar{\Delta}_j|_0 = -\nabla^2\phi(\bar{b}_{-j})|_{\mathcal{T}}\, w_j/(n-1)^2 \in \mathcal{T}$ *for* $j \neq i$, *and* $d_i = 0$.

Proof. (i) By the Savage representation, $\mathbb{E}\, v(S'; r, Y) = \phi(\bar{p}) - D_\phi(\bar{p}, \bar{r}_{S'})$ for any coalition S'; subtract the two coalitions of (1). With others truthful, $\bar{r}_{N\setminus i} = \bar{b}_{-i}$ carries no h while $\bar{r}_N = \bar{p} + h/n$. (ii) Taylor-expand $u \mapsto D_\phi(\bar{p}, \bar{p}+u)$ at $u = 0$: value 0, gradient $\nabla_u D_\phi(\bar{p}, u) = -\nabla^2\phi(u)(\bar{p} - u) \to 0$ at $u = \bar{p}$, Hessian $\nabla^2\phi(\bar{p})$; set $u = h/n$. For Brier ($\phi_{\mathrm{B}}(q) = \frac{1}{2}(\|q\|^2 - 1)$, Bregman divergence $\frac{1}{2}\|q - r\|^2$) the expansion is the exact value. (iii) Chain rule: $\bar{r}_{N\setminus j} = \bar{b}_{-j} + h/(n-1)$, and $\nabla_u D_\phi(\bar{p}, u)|_{\bar{b}_{-j}} = -\nabla^2\phi(\bar{b}_{-j})(\bar{p} - \bar{b}_{-j}) = -\nabla^2\phi(\bar{b}_{-j})\, w_j/(n-1)$ (since $\bar{p} - \bar{b}_{-j} = w_j/(n-1)$); the chain rule multiplies by the inner derivative $1/(n-1)$, giving (iii). The full-coalition term contributes zero gradient at $h = 0$, which also gives $d_i = 0$: the envelope fact. All statements restricted to $\mathcal{T}$ after projection. *Binary reduction* (scalar parametrization, $\mathcal{S}_{\mathrm{B}}$ of Definition 1): with the $\frac{1}{2}$-normalization the vector and scalar parametrizations agree *exactly* — $\bar{\Delta}_i^0 = (b_i - p)^2/(n-1)^2$, $\frac{1}{2}h^\top Q h = \mu_i c^2$ with $\mu_i = 1/n^2$, and the scalar directional derivative of (iii) is $2(p - b_j)/(n-1)^2$ — machine-verified (`verify_multiclass.py`, C1–C4, Brier residuals $\leq 10^{-12}$; an implementation check of these formulas rather than evidence for the lemma). □

Proof of Theorem 1. (a) First-order dichotomy. By Lemma M1(ii) the settlement side of $U_i(th) - U_i(0)$ is $-A\frac{t^2}{2}h^\top Q h + o(t^2)$ (per deviation report, scale A from (S1)); by (S2) the rent side is $\mathcal{R}_i(th) = t\langle a_i, h\rangle + o(t)$. Divide by $t > 0$ and let $t \to 0$: the one-sided directional derivative is $\langle a_i, h\rangle$. Since $\sup_{h\in\mathcal{T}, \|h\|=1}\langle a_i, h\rangle > 0$ iff $\Pi_{\mathcal{T}} a_i \neq 0$, first-order local truthfulness holds iff $\Pi_{\mathcal{T}} a_i = 0$. Under (S3)–(S4), pick h with $\langle a_i, h\rangle > 0$: the first-order term dominates for small t, so $U_i(th) > U_i(0)$ on $(0, \bar{t})$ — existence only; the optimum needs (b). Normalization of the weight map contributes only $\mathbf{1}^\top J_\varphi = 0$, which does not force $\Pi_{\mathcal{T}} a_i \neq 0$ (the dead-zone map of Section 3.1 satisfies it with a zero Jacobian) — (S3) is a genuine assumption.

(b) Weak-externality optimum. In the ε-family $\mathcal{R}_{i,\varepsilon} = \varepsilon\widetilde{\mathcal{R}}_i$ the rent gradient is $\varepsilon\nabla\widetilde{\mathcal{R}}_i(h) = \varepsilon\tilde{a}_i + O(\varepsilon\|h\|)$ (uniformly bounded second derivative of $\widetilde{\mathcal{R}}_i$), so the objective $\Delta U_i(h) = \varepsilon\widetilde{\mathcal{R}}_i(h) - \frac{A}{2}h^\top Q h + o(\|h\|^2)$ has Hessian $-AQ + O(\varepsilon)$ on $\mathcal{T}$: once $\varepsilon < A\lambda_{\min}(Q|_{\mathcal{T}})/\sup\|\nabla^2\widetilde{\mathcal{R}}_i\|$ it is strongly concave near 0 and its local maximizer is the unique stationary point there. Stationarity reads $AQh = \varepsilon\Pi_{\mathcal{T}}\tilde{a}_i + O(\varepsilon\|h\|) + o(\|h\|)$; since $Q \succ 0$ on $\mathcal{T}$ this forces $\|h^\star_\varepsilon\| = O(\varepsilon)$, whence $h^\star_\varepsilon = \frac{\varepsilon}{A}Q^{-1}\Pi_{\mathcal{T}}\tilde{a}_i + o(\varepsilon)$ (Q^{-1} the inverse on $\mathcal{T}$; $O(\varepsilon^2)$ for the remainder-free Brier instance), and substituting back, $\varepsilon\langle\tilde{a}_i, h^\star_\varepsilon\rangle - \frac{A}{2}h^{\star\top}_\varepsilon Q h^\star_\varepsilon = \frac{\varepsilon^2}{2A}\|\Pi_{\mathcal{T}}\tilde{a}_i\|^2_{Q^{-1}} + o(\varepsilon^2)$. *Scope of the max.* On the sphere $\|h\| = \rho$ (any fixed ρ inside (S1)'s expansion neighborhood), $\Delta U_i \leq \varepsilon\, O(1) - \frac{A}{2}\lambda_{\min}(Q|_{\mathcal{T}})\rho^2 + o(\rho^2) < 0$ for ε small, while $\Delta U_i(h^\star_\varepsilon) = \Theta(\varepsilon^2) > 0$; hence the maximum over the closed ρ-ball is interior and equals the stationary value (the display is a *local* maximum in general. When the settlement term is globally quadratic (the Brier instance)) or, more generally, when its Hessian in h is bounded above by a fixed negative-definite $-AQ_0$ on the whole admissible tangent set; global strong properness alone does *not* supply this, being a value inequality on the Bregman divergence, which need not be convex in its second argument; and $\sup\|\nabla^2\widetilde{\mathcal{R}}_i\| < \infty$ there, the objective's Hessian is $\varepsilon\nabla^2\widetilde{\mathcal{R}}_i - AQ \prec 0$ everywhere for ε small, so the unique stationary point is the global maximum and the same display holds with $\max_h$ unrestricted. The ε-family is load-bearing: a rent Hessian merely *bounded* at a fixed externality scale licenses only $AQh = \Pi_{\mathcal{T}}a_i + O(\|h\|)$, whose solution is governed by $(AQ - \nabla^2\mathcal{R}_i(0))^{-1}$ and need not equal $A^{-1}Q^{-1}\Pi_{\mathcal{T}}a_i$ to leading order; it is the scaling $\nabla^2\mathcal{R}_{i,\varepsilon} = O(\varepsilon)$ that collapses the two. (Verified numerically: the relative error of the closed form against a gradient-ascent optimum of the full utility scales linearly in the rent scale χ (the ε of the family in the verification) 1.5% at $\chi = 10^{-2}$, 0.15% at 10^{-3}; `verify_multiclass.py` C3.)

(c) Cushion ellipsoid. Immediate from (S1′): $\mathbb{E}[Z_{i,k} \mid \mathcal{H}_{k-1}] \geq \bar{\Delta}_i^0(\mathcal{H}_{k-1}) - \frac{1}{2}h^\top Q(\mathcal{H}_{k-1})h \geq 0$ *whenever* $h^\top Q(\mathcal{H}_{k-1})h \leq 2\bar{\Delta}_i^0(\mathcal{H}_{k-1})$; a sufficient condition, since (S1′) is an inequality; it is "exactly when" only in the Brier instance, where (S1′) holds with equality. Binary: $2\mu_i c^2 \leq 2\bar{\Delta}_i^0$, i.e., $|c| \leq \sqrt{\bar{\Delta}_i^0/\mu_i}$.

(d) Statistical invisibility. Fix any adapted deviation policy whose per-stage deviation stays in the history-conditional ellipsoid, and any δ-safe detector T_{det}. Condition on the full history $\mathcal{H}_{k-1}$ (reports, policy randomness, settled labels): by adaptedness the stage deviation h_k is determined,

and by history-stationarity (which is exactly where that assumption is spent) the next label is an independent draw from $\bar{p}$, so (S1′) plus the ellipsoid constraint give $\mathbb{E}[Z_{i,k} \mid \mathcal{H}_{k-1}] \geq 0$. The tower property over $\mathcal{F}^Z_{k-1} \subseteq \mathcal{H}_{k-1}$ preserves the inequality, so the induced law of the own stream lies in H_0, and δ-safety gives $P(T_{\mathrm{det}} < \infty) \leq \delta$: the detector misses *that policy* with probability $\geq 1 - \delta$; a per-policy statement, quantified over every policy in the class; no single event is claimed to cover the (uncountable) class at once. The argument operates on the *scalar* settled stream and never uses the deviation's dimension; it is verbatim across K. A detector safe only for the *point* null (the exact truthful law) can fire on the $\Theta(\|h\|^2)$ mean dip, but forfeits the uniform drift-robust false-alarm control over H_0 that production monitoring requires; the impossibility applies to every detector meeting the composite safety standard. □

Corollary 1 (Proportional weights, binary Brier, population normalization)**.** *Let welfare be the coalition-mean score* (1) *(Brier, $K = 2$, under the pooled-posterior model of Section 3.1) and weights proportional, $\varphi_i = \bar{\Delta}_i / \sum_k \bar{\Delta}_k$, on the all-active region, in the* stationary mean-statistic (population) normalization*; the map evaluated at deterministic expected marginals, so $\psi_i = \omega_i \psi_i^{\mathrm{pop}}$ (Corollary K′-stationary). At the truthful point,*

$$\bar{\Delta}_i^0 = \frac{(b_i - p)^2}{(n-1)^2}, \qquad \mu_i = \frac{1}{n^2}, \qquad d_j = \frac{2(p - b_j)}{(n-1)^2},$$

$$\psi_i^{\mathrm{pop}} = \frac{2\,\bar{\Delta}_i^0 (p - b_i)}{(n-1)^2\, S_\Delta^2}, \qquad S_\Delta = \sum\nolimits_k \bar{\Delta}_k^0,$$

and the cushion is $\varrho_i = n|b_i - p|/(n-1)$. Every active agent has $b_i \neq p$ by definition of the region, so $\psi_i^{\mathrm{pop}} \neq 0$ throughout*: within this normalization every agent has an active farming direction, $\mathrm{sign}(p - b_i)$, and (S3) is discharged by the model's own structure (ψ_i^{pop} vanishes only at the excluded, non-differentiable boundary $b_i = p$; no claim there). For finite random windows the operator form* (6) *governs and the Jensen-plus-covariance remainder can move the channel; no window-free discharge is claimed.*

Proof of Corollary 1. The normalization premise does real work and is where the corollary's scope lives: in the stationary mean-statistic evaluation the map consumes the deterministic expected marginals, Corollary K′-stationary applies with the operating-point Jacobian below, and $\psi_i = \omega_i \psi_i^{\mathrm{pop}}$; the sign and vanishing of ψ_i are those of ψ_i^{pop} since $\omega_i > 0$. For finite random windows the operator form (6) governs instead, and the Jensen-plus-covariance remainder (nonlinear map at a random statistic, shared sample t) can move the channel; no window-free discharge is claimed.

On the all-active region every $\bar{\Delta}_k^0 > 0$, the positive-part clip is inactive, and $\varphi_i = \bar{\Delta}_i / \sum_k \bar{\Delta}_k$ is real-analytic; boundary agents ($b_k = p$, a measure-zero clip case) are excluded; and with them the clip's non-differentiability. Specializing Lemma M1 to scalar binary Brier ($\phi'' = 2$ in the scalar parametrization): $\bar{\Delta}_i^0 = (b_i - p)^2/(n-1)^2$ (from $\bar{b}_{-i} - p = (p - b_i)/(n-1)$), $\mu_i = 1/n^2$, and $d_j = 2(p - b_j)/(n-1)^2$ (coalition-level envelope $\partial \mathbb{E} v(N)/\partial r_i = 0$ at truthful; only the leave-j-out reference moves). Quotient rule at the truthful point: $\partial \varphi_i / \partial \bar{\Delta}_j = -\bar{\Delta}_i^0 / S_\Delta^2$ for $j \neq i$, $S_\Delta = \sum_k \bar{\Delta}_k^0$, so

$$\psi_i^{\mathrm{pop}} = \sum_{j \neq i} \frac{\partial \varphi_i}{\partial \bar{\Delta}_j} d_j = -\frac{\bar{\Delta}_i^0}{S_\Delta^2} \cdot \frac{2}{(n-1)^2} \sum_{j \neq i} (p - b_j) = \frac{2\,\bar{\Delta}_i^0\,(p - b_i)}{(n-1)^2\, S_\Delta^2},$$

the last step by the telescoping $\sum_{j \neq i} (p - b_j) = b_i - p$ (from $\sum_k b_k = np$). Hence $\psi_i^{\mathrm{pop}} \neq 0$ throughout the all-active region (where $b_i \neq p$ for every active agent), vanishing only in the excluded boundary limit $b_i \to p$ — where the map is not differentiable and no claim is made; the profitable direction is $\mathrm{sign}(p - b_i)$, and the cushion is $\varrho_i = \sqrt{\bar{\Delta}_i^0 / \mu_i} = n|b_i - p|/(n-1)$. All closed forms verified numerically at $n \in \{2, 3, 4\}$ with asymmetric components (Section J). The characterization is about the report–posterior gap rather than informativeness: an informative agent may still sit at $b_i = p$ — outside the all-active region, with zero truthful marginal. □

D SOFTMAX WEIGHT RESPONSE

For $\varphi_i = e^{\eta \bar{\Delta}_i} / \sum_k e^{\eta \bar{\Delta}_k}$ the Jacobian is $\partial \varphi_i / \partial \bar{\Delta}_j = \eta\, \varphi_i (\delta_{ij} - \varphi_j)$, so with the envelope fact $d_i = 0$ the (one-dimensional, binary) weight response; in the same stationary mean-statistic normalization

as Corollary 1; softmax is nonlinear, so at finite m the operator form and its remainder govern instead (Section B) — is

$$\psi_i^{\mathrm{pop}} \;=\; -\,\eta\,\varphi_i^0 \sum_{j\neq i} \varphi_j^0\, d_j.$$

Nondegeneracy enters as an assumption, and we do not prove it: the condition is $\sum_{j\neq i}\varphi_j^0 d_j \neq 0$, a φ-weighted sum of peer marginal gradients at the operating point. Only at the *equal-weight* operating point does it reduce — via $d_j = 2(p-b_j)/(n-1)^2$ and the telescoping $\sum_{j\neq i}(p-b_j) = b_i - p$ — to the proportional-map characterization $b_i \neq p$; at a non-uniform φ^0 it is a per-operating-point check and cancellations are possible. The same caveat holds for softmax at every K; no assumption-free multiclass discharge is claimed (Corollary 1 stays binary). Checks against finite differences: $n=2, \eta=4$: $\psi_1^{\mathrm{pop}} = 0.400$; $n=3, \eta=7$, asymmetric components, non-uniform φ^0: $\psi_1^{\mathrm{pop}} = 0.1219$ — both to 10^{-5} (Section J).

E THE RESTART-MIXTURE E-DETECTOR: FULL PROOFS

Construction. Per agent, on the settled stream $(Z_k) \subset [-B, B]$ adapted to $\mathcal{F}_k$, with null H_0: $\mathbb{E}[Z_k \mid \mathcal{F}_{k-1}] \geq 0$ and alternative $H_1(\nu, g)$: $\mathbb{E}[Z_k \mid \mathcal{F}_{k-1}] \leq -g$ for $k > \nu$. The λ-grid is $\lambda_j = 2^{-j}/B$, $j \geq 0$; instance (ℓ, q) activates at $s_{\ell q} = q\, 2^{\max(\ell-1,0)}$ and runs 2^ℓ samples (dyadic covering: any window of length N contains a full instance of scale $\ell^* = \lfloor \log_2 N \rfloor - 1$ starting within $N/2$ of its start). The mixture is $\bar{M}_t = \sum_{\ell,q,j} \pi_{\ell q j} M_t^{(\ell,q,j)}$ with

$$M_t^{(\ell,q,j)} = \exp\Big(-\lambda_j S_{s_{\ell q}:t\wedge(s_{\ell q}+2^\ell)} - \lambda_j^2 \tfrac{B^2}{2}\big(t \wedge (s_{\ell q}+2^\ell) - s_{\ell q}\big)\Big),$$

$M \equiv 1$ before activation, prior $\pi_{\ell q j} = c\,(\ell+1)^{-2}(q+1)^{-2}(j+1)^{-2}$ with $\sum \pi \leq 1$; alarm at the first t with $\bar{M}_t \geq n/\delta$.

Lemma (Detector safety)**.** *Under H_0, $\Pr[\exists t : \bar{M}_t \geq n/\delta] \leq \delta/n$; a union bound over the n agents gives the family-wise budget δ.*

Proof. For $Z_k \in [-B, B]$ with $\mathbb{E}[Z_k \mid \mathcal{F}_{k-1}] \geq 0$, Hoeffding's lemma (range $2B$) gives $\mathbb{E}[e^{-\lambda Z_k} \mid \mathcal{F}_{k-1}] \leq e^{-\lambda\mathbb{E}[Z_k|\mathcal{F}_{k-1}] + \lambda^2 B^2/2} \leq e^{\lambda^2 B^2/2}$, so each $M^{(\ell,q,j)}$ is a nonnegative supermartingale with initial value 1 from its activation (constant 1 before). A convex combination with $\sum \pi \leq 1$ is again a nonnegative supermartingale with $\bar{M}_0 \leq 1$; Ville's inequality gives the crossing bound. □

Lemma (Grid adequacy)**.** *For any $g \in (0, B]$ there is j^* with $\lambda_{j^*} \in [g/2B^2,\, g/B^2]$ (the grid is geometric with ratio 2 and $g/B^2 \leq \lambda_0$), and every λ in that bracket has drift $\lambda g - \lambda^2 B^2/2 \geq 3g^2/8B^2$; moreover $j^* + 1 \leq \log_2(2B/g) + 1$.*

Proof. The concave quadratic $\lambda \mapsto \lambda g - \lambda^2 B^2/2$ attains its bracket minimum at the endpoint $\lambda = g/2B^2$: $g^2/2B^2 - g^2/8B^2 = 3g^2/8B^2$. □

Proof of the delay bound (Safety-and-delay lemma, (ii)). Let N be a candidate delay. Take the covering instance: scale $\ell^* = \lfloor \log_2 N \rfloor - 1$, so the activation spacing is $2^{\ell^*-1} \in [N/8, N/4]$, and the first start $s \geq \nu$ obeys $s - \nu \leq 2^{\ell^*-1} \leq N/4$; the window $[s, s+2^{\ell^*}]$, of length $m' = 2^{\ell^*} \geq N/4$, lies entirely post-change and within the first N post-change samples ($N/4 + N/2 \leq N$). Its start index is $q = s/2^{\ell^*-1} \leq \nu/2^{\ell^*-1} + 1 \leq 8\nu/N + 1$, whence $\log(q+1) \leq \log^+(\nu/N) + O(1)$. Take j^* from the grid-adequacy lemma. Over the window the increments have conditional mean $\leq -g$; Azuma–Hoeffding on the martingale part (increments bounded by $2B$) gives, with probability $\geq 1 - \delta''$, $S_{s:s+m'} \leq -gm' + 2B\sqrt{2m' \log(1/\delta'')}$. Hence

$$\log M_{s+m'}^{(\ell^*,q,j^*)} \;\geq\; \tfrac{3g^2}{8B^2}m' - \tfrac{g}{B^2}\cdot 2B\sqrt{2m'\log\tfrac{1}{\delta''}} \;\geq\; \tfrac{3g^2}{16B^2}m' - \tfrac{32}{3}\log\tfrac{1}{\delta''},$$

the last step by $2ab \leq \epsilon a^2 + b^2/\epsilon$. The alarm fires once this exceeds $\log(n/\delta) + 2\log(\ell^*+1) + 2\log(q+1) + 2\log(j^*+1) + O(1)$, which by the index bounds is $\log(n/\delta) + O(\log\log^+ N) + O(\log^+(\nu/N)) + O(\log\log(2B/g))$. Solving $\frac{3g^2}{16B^2}\cdot\frac{N}{4} \geq (\text{threshold}) + \frac{32}{3}\log(1/\delta'')$ for N gives the stated bound (the inequality is self-referential only through iterated logarithms; solving it changes constants only). The body's $\log\frac{n}{\delta\delta''}$ merges $\log\frac{n}{\delta} + \log\frac{1}{\delta''}$ — identical. □

What pays the $\log^+(\nu/N)$ term; variants. The term is the price of the restart-position union in *our* summable-prior construction: the prior mass at start index $q \approx 8\nu/N$ is $\Theta((q+1)^{-2})$, and its logarithm enters the threshold. We do *not* prove a matching lower bound, and make no claim that every procedure must pay a growing-in-ν term; under an infinite-horizon family-wise false-alarm budget such a price is consistent with the classical false-alarm-probability vs. average-run-length distinction, and we state the construction-specific form because the ν-dependence enters Theorem 2's accounting. Under an ARL formulation instead, e-CUSUM/restart-at-alarm detectors achieve $N = O\big(\frac{B^2}{g^2}\log\frac{n}{\delta_{\mathrm{rate}}}\big)$ uniformly in ν (Lorden-type shape); the $(B^2/g^2)\log(1/\delta)$ shape matches the standard two-point lower-bound heuristic (we do not prove optimality here). A variance-adaptive version (empirical-Bernstein sub-ψ processes) replaces B^2 by $O(\sigma^2+Bg)$ throughout; grid and covering arguments unchanged. Scope: this detector bounds reaction to H_1-visible deviations only — extractive deviations beyond the cushion of Theorem 1(c); it says nothing about farming (null side), and it is the reallocation *coupling* of this sound alarm that Theorem 2 weaponizes.

F Proof of Theorem 2

Realized utility, and the coupling. Throughout, $\widehat{U}_1$ is the *realized* (pathwise) value of the two sums in (2) (the discounted settled transfers plus χ_1 times the discounted realized weight trajectory) so $U_1 = \mathbb{E}[\widehat{U}_1]$; pathwise statements are about $\widehat{U}_1$, and (2)'s U_1 itself is already an expectation. Both worlds — *reallocation* and *alert-only* — are constructed on one probability space carrying the same reports, labels, and delays. Under the mechanism's serve-only exclusion semantics, settlement is computed from the full report coalition in *both* worlds, so every settled marginal, every transfer $\pi_{i,t}$, and hence every detector statistic is *identical path by path for all t*: on each path the alarm fires at the same time in both worlds, or in neither (on the capture event $\mathcal{E}$, by settlement index $\nu + N_\star$). The worlds differ in exactly one act: the reallocation world executes the weight transfer, the alert-only world does not, both retaining the continuous adaptive weight map. Let T_{act} be the first serving epoch whose weights are computed from a post-alert publication; before T_{act} the two worlds' weight trajectories coincide, and on paths where no alarm ever fires ($T_{\mathrm{act}} = \infty$) they coincide throughout. On the capture event $\mathcal{E}$ the alarm has fired by settlement index $\nu + N_\star$, so $T_{\mathrm{act}} \le \bar{T}_{\mathrm{act}}$, the *deterministic* serving epoch corresponding to $\nu + N_\star$ plus the lag to the next weight publication.

Per-epoch differential. For serving epoch $t \ge T_{\mathrm{act}}$ define the *exact* (realized, undiscounted) rent differential

$$r_t \;:=\; \chi_1\big(\hat{s}_{1,t}^{\mathrm{realloc}} - \hat{s}_{1,t}^{\mathrm{alert}}\big),$$

the value to the farmer of the weight the transfer moved at epoch t. The discount β^t is applied *once*, in the utility sum (2) — r_t itself is undiscounted, so no discount appears twice. Since serving weights lie in $[0,1]$, $|r_t| \le \chi_1$ on every path; write $G := \sum_{t=T_{\mathrm{act}}}^{H} \beta^t r_t$ for the realized coupled difference ($G := 0$ when $T_{\mathrm{act}} = \infty$), so that $|G| \le C_{\mathrm{fail}} := \chi_1 \sum_{t=0}^{H} \beta^t$ always. The theorem's hypothesis, stated as part of the theorem: a constant $r > 0$ with $r_t \ge r$ for every effective epoch $t \in [T_{\mathrm{act}}, H]$ on the capture event $\mathcal{E}$, i.e., the transfer's weight advantage does not reverse while it is in force.

Proof of Theorem 2. Parts (a)–(b) below establish the two components of the capture event $\mathcal{E}$; (c)–(d) the amplification and profitability statements. Work on the intersection of the events of (a) and (b), of probability $\ge 1 - \delta - \delta''$ by a union bound.

(a) Farmer never downweighted. The framing deviation leaves the farmer's own settled stream with conditional mean $\mu_1(s) = 0.04 - s^2/4 \ge 0$ for $s \le 0.4$; by Theorem 1(c)–(d) its law lies in the composite null H_0, and the safety half of the detector lemma (family-wise budget δ) implies no alarm on agent 1 at any time.

(b) Honest peer captured. By the settled-stream changepoint hypothesis of the theorem, agent 2's settled stream has conditional mean $\le -g(s)$, $g(s) = -\mu_2(s) > 0$, from settlement index ν on; the hypothesis is where the timeline is glued: a deviation launched at a *decision* epoch produces a clean settlement-index changepoint only when no pre-deviation truthful record settles after ν (deviation from the first record, or a finite flush); with unbounded out-of-order delays the interleaving never

ends and the hypothesis is a genuine restriction. The delay half of the detector lemma (budget δ'') fires by settlement index $\nu + N_\star$ with $N_\star$ the smallest integer satisfying the delay bound at gap $g(s)$ — including its $\log^+(\nu/N_\star)$ term, so a late onset pays a longer delay. The alarm is *correct*: H_0 is a property of the report profile rather than of any single agent. The transferred weights first serve at epoch T_{act}.

(c) Pathwise identity, success bound, ex-ante bound. Subtract the two worlds' realized utilities $\widehat{U}_1$ term by term, *on each sample path*. The settlement sums are identical path by path (serve-only semantics), and so are the weight trajectories before T_{act} (throughout, when no alarm fires); both cancel *exactly* — in particular the deviation's settlement cost is a common term of both worlds and does not appear in the difference. What remains is the realized coupled difference G: the identity holds on *every* path, alarm or not, and no expectation has been taken. *Success bound:* on $\mathcal{E}$ the hypothesis $r_t \geq r > 0$ applies on $[T_{\mathrm{act}}, H]$ and $T_{\mathrm{act}} \leq \bar{T}_{\mathrm{act}}$, so

$$G = \sum_{t=T_{\mathrm{act}}}^{H} \beta^t r_t \;\geq\; r \sum_{t=T_{\mathrm{act}}}^{H} \beta^t \;\geq\; r \sum_{t=\bar{T}_{\mathrm{act}}}^{H} \beta^t = G_{\mathrm{succ}} \;\geq\; 0,$$

strictly positive iff $H \geq \bar{T}_{\mathrm{act}}$; for $H < T_{\mathrm{act}}$ the sum is empty and the two worlds coincide. *Ex-ante bound:* the pathwise positive gain on $\mathcal{E}$ does *not* by itself sign the unconditional expectation; the complement must be controlled. Split:

$$\mathbb{E}G = \mathbb{E}[G;\mathcal{E}] + \mathbb{E}[G;\mathcal{E}^c] \;\geq\; \Pr(\mathcal{E})\, G_{\mathrm{succ}} - \Pr(\mathcal{E}^c)\, C_{\mathrm{fail}} \;\geq\; (1 - \delta - \delta'')\, G_{\mathrm{succ}} - (\delta + \delta'')\, C_{\mathrm{fail}},$$

using $G \geq G_{\mathrm{succ}} \geq 0$ on $\mathcal{E}$, $|G| \leq C_{\mathrm{fail}}$ always (a false alarm on agent 1, e.g., makes r_t negative; the crude bound covers every failure mode), and $\Pr(\mathcal{E}) \geq 1 - \delta - \delta''$. The right side is positive iff $\delta + \delta'' < G_{\mathrm{succ}}/(G_{\mathrm{succ}} + C_{\mathrm{fail}})$ (when $\Pr(\mathcal{E}^c) \leq \delta + \delta''$ is used at its cap) — meaningful only for $H \geq \bar{T}_{\mathrm{act}}$; when $H < \bar{T}_{\mathrm{act}}$, $G_{\mathrm{succ}} = 0$ by the empty-sum convention, the bound degenerates to the trivial $\mathbb{E}G \geq -(\delta + \delta'')C_{\mathrm{fail}}$, and no ex-ante amplification is claimed (on $\mathcal{E}$ the pathwise sum may still be nonempty when $T_{\mathrm{act}} \leq H < \bar{T}_{\mathrm{act}}$, but we do not price it).

(d) Absolute profitability, ex ante. Take the pathwise identity $\Delta\widehat{U}_1 = \Delta\widehat{U}_1^{\mathrm{alert}} + G$ in expectation: $\Delta U_1 = \Delta U_1^{\mathrm{alert}} + \mathbb{E}G$, so ex-ante profitability is equivalent to $\Delta U_1^{\mathrm{alert}} + \mathbb{E}G > 0$ — an equivalence licensed by the identity rather than by the lower bound. The deviation's expected settlement cost, which enters $\Delta U_1^{\mathrm{alert}}$ (and only there in this decomposition), is $C_{\mathrm{farm}} = \sum_{k \in \mathcal{D}} A_k\, s^2/4 \leq \bar{A}\, \frac{s^2}{4} |\mathcal{D}|$, summing the per-epoch discounted settlement scales $A_k \leq \bar{A}$ over the settled deviation epochs $\mathcal{D}$, $|\mathcal{D}| \leq N_\star + N_{\mathrm{dev}}$; if the farmer reverts to truthful after the capture, $N_{\mathrm{dev}} = 0$ and its stream statistics return to the truthful law; still in H_0. Assume, as a direct hypothesis, that the framing deviation's continuous-channel rent relative to truthful play is nonnegative in expectation, $R_{\mathrm{cont}}(s) \geq 0$. Peer-channel monotonicity ($\partial\varphi_1/\partial\bar{\Delta}_2 \leq 0$) alone does *not* imply this: framing also lowers the farmer's own measured marginal ($\mu_1(s) = 0.04 - s^2/4 < \mu_1(0)$), and the own-channel loss could in principle dominate the peer-channel gain. In the affine worked instance the hypothesis holds by exact polynomials: the framing deviation shifts the weight-map input by $\bar{\bar{\Delta}}_1(s) - \bar{\bar{\Delta}}_2(s) - (\bar{\bar{\Delta}}_1(0) - \bar{\bar{\Delta}}_2(0)) = 0.4s - s^2 > 0$ on $(0, \frac{2}{5}) \supset (\frac{2}{15}, \frac{2}{5})$, so the farmer's continuous-channel weight, and with $\chi_1 > 0$ its rent, weakly rises. Under $R_{\mathrm{cont}}(s) \geq 0$, $\Delta U_1^{\mathrm{alert}} \geq -C_{\mathrm{farm}}$, and by (c)'s ex-ante bound, $(1 - \delta - \delta'')\, G_{\mathrm{succ}} - (\delta + \delta'')\, C_{\mathrm{fail}} > C_{\mathrm{farm}}$ suffices (undiscounted, in the small-budget reading $\delta + \delta'' \to 0$: $(H - \bar{T}_{\mathrm{act}} + 1)_+\, r > C_{\mathrm{farm}}$). □

Worked-instance constants. In the two-agent instance (Section 3.2): framing region $s \in (2/15,\, 2/5)$ (the exact roots of μ_2), $\mu_2(s) = 0.04 - 0.4s + 0.75s^2$; post-capture rent differential $r = \chi_1 \Delta\hat{s}_1$, evaluated *at the operating point*: the alert-only world keeps the continuous map, so at $s = 0.135$ (with $\eta = 1$) $\hat{s}_1^{\mathrm{alert}} = \frac{1}{2} + \eta(0.4s - s^2) = 0.535775$ and $\Delta\hat{s}_1 = 0.464225$, giving $r = 0.4\lambda \cdot 0.464225 = 0.1857\lambda$ — the constant the theorem's hypothesis requires, and the value the coupled difference below reports. (Using $\Delta\hat{s}_1 = \frac{1}{2}$ would be the $s = 0$ weight, which overstates r and is *not* a valid lower bound at the operating point.) At $s = 0.135$ the framing-plus-capture gain is 0.1954 per decision against 0.0098 for continuous farming at s^* (the alert-only world's channel), coupled difference 0.186 — success-path per-decision figures; the $\sim 20\times$ ratio is the long-horizon limit $(H - \nu - N_\star)/N_\star \to \infty$, $\beta \to 1$ only. The framing region's upper endpoint coinciding with the cushion boundary at 0.4 is an artifact of $b_2 - b_1 = 0.4$ in this instance rather than structural.

G TRUST-REGION BOUND AND WATERMARK CONSTRUCTION

Lemma (G.1: global settlement regret)**.** *Under (S1) and the pooled-posterior model, a per-report deviation $h \in \mathcal{T}$ costs, in expected settled marginal, $\mathbb{E}[\bar{\Delta}_i(0)] - \mathbb{E}[\bar{\Delta}_i(h)] = D_\phi(\bar{p}, \bar{p} + h/n)$; global strong properness (the hypothesis of Theorem 3) is the requirement $D_\phi(\bar{p}, \bar{p} + h/n) \geq \frac{1}{2} h^\top Q h$ for every admissible h, and the Brier instance satisfies it* with equality*: its Bregman divergence is $\frac{1}{2}\|q - r\|_2^2$, so the drop is $\frac{1}{2n^2}\|h\|^2 = \frac{1}{2} h^\top Q h$ exactly. Binary scalar form: $\mu_i c^2$ with $\mu_i = 1/n^2$.*

Proof. The identity is Section 3.1's exact marginal geometry (proof in Section C); only the reference coalition is r_i-free, so the whole h-dependence sits in $D_\phi(\bar{p}, \bar{r}_N)$ with $\bar{r}_N = \bar{p} + h/n$. For $\phi_{\mathrm{B}}(q) = \frac{1}{2}(\|q\|^2 - 1)$ the Bregman divergence is exactly $\frac{1}{2}\|q - r\|^2$. For α_ϕ-strongly convex potentials (log score on the compact interior, via Pinsker) the inequality holds with Q replaced by $(\alpha/n^2) I|_{\mathcal{T}}$. □

Lemma (G.2: served-output trust bound)**.** *Project the published weights onto $\mathcal{W}_\tau := \Delta^{n-1} \cap \{\|\hat{s} - u\|_1 \leq 2\tau\}$ (u uniform). Then for* every *report matrix r — deviant or not —* $\mathrm{TV}(\hat{p}, \bar{p}_{\mathrm{ref}}) \leq \tau$*, where $\hat{p} = \sum_j \hat{s}_j r_j$ and $\bar{p}_{\mathrm{ref}} = \frac{1}{n}\sum_j r_j$ is the uniform-weight reference on the same reports.*

Proof. $\hat{p} - \bar{p}_{\mathrm{ref}} = \sum_j (\hat{s}_j - \frac{1}{n}) r_j$ with zero-sum coefficients. Split the coefficients into positive and negative parts, each of total mass $M = \frac{1}{2}\|\hat{s} - u\|_1 \leq \tau$; the sum is $M(q^+ - q^-)$ with $q^\pm \in \Delta^{K-1}$ convex combinations of the reports, so its TV norm is $M \cdot \mathrm{TV}(q^+, q^-) \leq M \leq \tau$. □

Proof of Theorem 3. Fix the deviation report h at epoch t and the counterfactual coupling of (5): both worlds carry the same realizations of all other reports, all labels, and all delays, so the deviation reaches the rent term of (2) only through the published weights. The proof works directly on that rent term — χ_i times the discounted own-weight trajectory — and never converts weights into a served-output value: Lemma K links χ_i to a business-value derivative only locally and directionally, and no global equivalence between the two utility forms is available or used.

Step 1 (per-epoch weight range bound). The published weights of each world lie in $\mathcal{W}_\tau$, and $\hat{s} - u$ is a zero-sum vector, so its positive and negative parts each carry mass $\frac{1}{2}\|\hat{s} - u\|_1 \leq \tau$, whence every coordinate obeys $|\hat{s}_i - \frac{1}{n}| \leq \tau$. At any epoch $u \in \mathcal{A}_t$, therefore, $|\hat{s}^h_{i,u} - \hat{s}^0_{i,u}| \leq 2\tau$; for every report matrix, deviant or not. At $u \notin \mathcal{A}_t$ the two worlds' weights coincide (the coupling: all other inputs to the weight map are identical) and the epoch contributes zero.

Step 2 (discounted lifetime sum). The rent difference is therefore

$$\Big|\chi_i \,\mathbb{E}\Big[\sum_u \beta^{\,u}\big(\hat{s}^h_{i,u} - \hat{s}^0_{i,u}\big)\Big]\Big| \;\leq\; 2\chi_i\tau\,\mathbb{E}\Big[\sum_{u \in \mathcal{A}_t} \beta^{\,u}\Big] \;=\; 2\chi_i\tau\,\Lambda_{i,t}.$$

This is a worst-case *range* argument: no factor $1/m$, no weight-map Jacobian, and in particular the differential kernel ω_i of (7) plays no role here.

Step 3 (settlement side). The deviation report is costed once, at settlement, and Lemma G.1 gives $\mathbb{E}[\text{settlement}(h)] - \mathbb{E}[\text{settlement}(0)] \leq -\frac{A}{2} h^\top Q h$, where A is the full discounted settlement scale of the single report (S1) — the curvature side of the ledger never rescales with the exposure. Weak preference for h requires the sum of the two sides to be ≥ 0: $\frac{A}{2} h^\top Q h \leq 2\chi_i\tau\Lambda_{i,t}$, i.e., $\|h\|_Q \leq 2\sqrt{\chi_i\tau\Lambda_{i,t}/A}$. Binary scalar parametrization: $h^\top Q h = 2\mu_i c^2$ gives $|c| \leq \sqrt{2\chi_i\tau\Lambda_{i,t}/(A\mu_i)}$. □

Corollary (Served-output form of Theorem 3)**.** *If the rent is instead a per-serving-epoch L_i-Lipschitz business value of the served distribution (Lemma K's primitive), the same bounds hold with χ_i replaced by L_i; the population arm's utility model.*

Proof of the served-output corollary. Under the alternative utility whose rent is $\mathbb{E}[\sum_u \beta^{\,u} g_{i,u}(\hat{p}_u)]$ with each $g_{i,u}$ L_i-Lipschitz in total variation: at any $u \in \mathcal{A}_t$, Lemma G.2 applies in each world against the *same* uniform-weight reference $\bar{p}_{\mathrm{ref},u}$ (the coupling makes the report matrices identical), so $\mathrm{TV}(\hat{p}^h_u, \bar{p}_{\mathrm{ref},u}) \leq \tau$ and $\mathrm{TV}(\hat{p}^0_u, \bar{p}_{\mathrm{ref},u}) \leq \tau$, whence $\mathrm{TV}(\hat{p}^h_u, \hat{p}^0_u) \leq 2\tau$ and $|g_{i,u}(\hat{p}^h_u) - g_{i,u}(\hat{p}^0_u)| \leq 2L_i\tau$; unaffected epochs contribute zero. (Without the coupling the two references could differ and the triangle inequality would pick up $\|\bar{p}^{\,h} - \bar{p}^{\,0}\|$; the corollary prices allocation influence rather than a report's direct effect on the forecast.) Steps 2–3 then run verbatim

with L_i in place of χ_i. This is the utility model of the population arm, which checks this corollary's bound in the two-period normalization ($\Lambda_{i,t} = \beta^{t+1}$; undiscounted grids, $\Lambda = 1$; Section J). □

Proposition (R2: watermark weighting, population form)**.** *For the* mean-statistic (population) mechanism *(weights evaluated at the expected marginals, the deterministic operating point that is also the $m \to \infty$ limit of the window statistic) let $\varphi_i(\Delta) = f_i(\Delta_i)/\sum_k f_k(\Delta_k)$ with $f_i(x) = 1 + \max(x - \Delta_i^0, 0)^2$, the agent-specific watermark Δ_i^0 set at agent i's truthful expected marginal. Then the map is C^1, normalized, and responsive, yet locally farming-proof* in that form*: own-weight gain would require the own expected marginal to exceed its own truthful maximum, and the peer channel moves weights only* against *a deviator.*

Proof of the watermark proposition (R2, population form). Throughout, the map is evaluated at *expected* marginals — the mean-statistic mechanism of the proposition. $f_i(x) = 1 + \max(x - \bar{\Delta}_i^0, 0)^2$ is C^1 with $f_i'(\bar{\Delta}_i^0) = 0$, so the full Jacobian of $\varphi_i = f_i/\sum_k f_k$ vanishes at the deterministic truthful operating point: $a_i = 0$ for every agent in this evaluation, and the first-order channel of Theorem 1(a) is closed. The stronger, non-local claim: (i) by the envelope fact (the truthful report is the strict argmax of the own expected marginal), no deviation lifts $\mathbb{E}\bar{\Delta}_i$ above its truthful watermark, so $f_i(\mathbb{E}\bar{\Delta}_i)$ cannot exceed its truthful value; (ii) peers' expected marginals move by $\langle d_j, h\rangle$ in either sign, but their f_j-scores sit at the floor $f_j = 1$ (each watermark is at that agent's truthful maximum) and can only rise; hence the ratio φ_i can only fall (the peer channel moves weight *against* a deviator. Verified numerically for both deviation signs up to $|c| = 0.15$ (population arm) mean statistics; Section J). Honest caveats, part of the statement's scope: the watermark $\bar{\Delta}_i^0$ is endogenous (a trailing quantile in practice); the dead zone removes adaptivity exactly near equilibrium (the same adaptivity-for-incentives price as the trust region); and agents legitimately operating *above* their watermarks put f_j off its floor and reopen a peer channel; R2 pays for this protection with the τ-frontier adaptivity cost recorded above.

Adjacency note. The map belongs to the same threshold-knot family as the kinked escapes of Lovén & Tarkoma (2026), though C^1 — knotted in curvature where their escapes kink in slope; its content here is that the escape survives the *peer* channel; the knot construction plays no role.

Scope: population form only. The argument maximizes the *expected* marginal, but at finite m the map consumes the random window statistic $\hat{\bar{\Delta}}$. Even when a deviation lowers $\mathbb{E}\hat{\bar{\Delta}}_i$, the event $\hat{\bar{\Delta}}_i > \bar{\Delta}_i^0$ has positive probability, and the convex f_i gives $\mathbb{E}f_i(\hat{\bar{\Delta}}_i) \geq f_i(\mathbb{E}\hat{\bar{\Delta}}_i)$ (Jensen), with the window variance itself deviation-dependent; so "no deviation raises own weight" is *not* claimed at finite m: the construction closes the first-order Jacobian channel at the deterministic operating point, and no closure guarantee extends to the random-window channel — finite-window noise *may* reopen it; whether it does, and in which direction, is map- and instance-dependent, and we claim nothing either way. The proposition is proved for the mean-statistic mechanism (equivalently the $m \to \infty$ deterministic operating point), and nothing stronger is claimed. □

The projection. Serving weights are published as the Euclidean projection onto $\mathcal{W}_\tau = \Delta^{n-1} \cap \{\|\hat{s} - u\|_1 \leq 2\tau\}$, an intersection of two convex sets with standard $O(n \log n)$ projections (sorting-based simplex projection; ℓ_1-ball via soft-thresholding). The population arm implements it by Dykstra's alternating projections, which converge to the exact projection for intersections of closed convex sets; at $n = 2$ the projection is available in closed form and the implementation is cross-checked against it (Section J).

H Exclusion semantics: the $n = 2$ proposition, the all-n cascade family, and the $n = 3$ sweep

Semantics and the key identity. *Reference exclusion* removes a detected agent from serving, settlement, and payment coalitions simultaneously; marginals are leave-one-out within the current active coalition; the empty coalition is anchored at the prior score, $v(\emptyset) = -(p_0^{\text{prior}} - y)^2$ (a convention, and deliberately flagged as one below). All computations in this section are exact expectations under binary Brier and truthful reports off the deviator: for an active coalition S and $j \in S$,

$$\mathbb{E}\,\bar{\Delta}_j^S = (\bar{b}_{S\setminus j} - p)^2 - (\bar{b}_S - p)^2, \tag{8}$$

since $\mathbb{E}(\bar{b}_{S'} - y)^2 = (\bar{b}_{S'} - p)^2 + p(1-p)$ for any S' and the variance term cancels in the difference; each mean offset is (the sum of $w_k = b_k - p$ over the coalition, plus any deviation) divided by the coalition size.

Proposition (X1: reference exclusion at $n = 2$, fixed anchor)**.** *In the two-agent worked instance (Section 3.2; $p_0^{\text{prior}} = p = \frac{1}{2}$), suppose the peer has been excluded and the remaining agent reports the constant r. Under reference exclusion with the fixed anchor, its settled-marginal stream has mean $v(\{1\}) - v(\emptyset)$ in expectation $= -(r-p)^2 \le 0$. Hence the captured position is either* detectable *or* worthless*: harvesting ($r = 0.3$, serving its own component) has stream mean -0.04, a gap $g = 0.04$ that the detector of Section 4 flags within its delay bound; the capture is time-boxed; hiding ($r = p$) zeroes the stream mean but serves the truthful pooled baseline, forfeiting the entire harvest. Under the paper's serve-only semantics, by contrast, the post-capture stream is that of a truthful agent in the intact settlement coalition (mean $+0.04$): the capture is permanent and invisible.*

Proof. By (8) with $S = \{1\}$ and the anchor in place of the leave-one-out term: $\mathbb{E}[v(\{1\})] = -(r-p)^2 - p(1-p)$ and $\mathbb{E}[v(\emptyset)] = -(p_0^{\text{prior}} - p)^2 - p(1-p) = -p(1-p)$, so the marginal mean is $-(r-p)^2$. The two branches are $r = 0.3$ and $r = p$; the serve-only comparison is the coupling of Section F, where settlement is computed from the full report coalition in every world. □

Convention fragility (why X1 is stated at $n = 2$ with a fixed anchor, and no more). The endgame leans on the anchor: a *lone truthful survivor* with component $b_0 \neq p$ also scores $-(b_0 - p)^2 < 0$ against the prior anchor — through no fault of its own. The "hide-or-be-caught" dichotomy is real in the two-agent instance but is a property of the constant-component toy structure plus the p_0^{prior} convention rather than of exclusion semantics in general; we therefore state X1 as a two-agent, fixed-anchor proposition only, and the failure of the generalization is the content of the rest of this section.

The family E_n. For $n \ge 3$, take the explicit component profile

$$E_n: \quad b = \big(\tfrac{9}{20},\ \tfrac{3}{20},\ \tfrac{9}{10},\ \underbrace{\tfrac{1}{2}, \ldots, \tfrac{1}{2}}_{n-3}\big), \qquad p = \tfrac{1}{2}, \qquad w = b - p = \big(-\tfrac{1}{20},\ -\tfrac{7}{20},\ \tfrac{2}{5},\ 0, \ldots, 0\big).$$

The $n-3$ dummies sit exactly at the pooled posterior: appending them preserves p (the embedding is exact rather than perturbative; the unweighted coalition mean gives every agent influence exactly $1/n$, so no "vanishing-influence auxiliary agents" exist in this mechanism), their truthful marginals are exactly 0 (the null boundary), and they are honest; no attacker is needed for Proposition E2. Every inequality below is a rational function of n with a two-line hand proof and a symbolic positive-coefficient certificate (`verify_cascade_all_n.py`: substitute $n \mapsto 3 + m'$ and check all numerator coefficients positive; exact-rational numerics at $n = 3, \ldots, 12$).

Proposition (Cascades at every $n \ge 3$; quoted in Section 4.2)**.** *For every $n \ge 3$ there is an explicit profile of forecast components on which (i)* invisible framing exists*: a deviator drives an honest peer's conditional marginal mean strictly negative while every other stream, its own included, stays inside the composite null; and (ii)* no attacker is needed*: after one agent is excluded for any reason, an honest survivor's truthful conditional mean is strictly negative (a consistent detector coupled to exclusion removes it) and for $n \ge 4$ the cascade continues until only uninformative agents remain. Margins are $\Theta(1/n^2)$: cascades slow with n and persist at every n.*

The profile is the family E_n above; (i) is Proposition E1 and (ii) is Proposition E2 below.

Proposition (E1: invisible framing at every $n \ge 3$)**.** *In E_n, let agent 3 ($w_3 = \frac{2}{5}$) deviate by $s = -\frac{1}{10}$. In the full coalition, by* (8)*:*

- *the framed agent 1 has $\mu_1(s) = \big(\frac{s - w_1}{n-1}\big)^2 - \big(\frac{s}{n}\big)^2 = \frac{1}{400(n-1)^2} - \frac{1}{100n^2} = -\frac{(n-2)(3n-2)}{400(n-1)^2 n^2} < 0$ for all $n \ge 3$;*
- *the deviator stays inside its own null: $\mu_3(s) = \frac{(2/5)^2}{(n-1)^2} - \frac{1}{100n^2} > 0$, and $|s| = \frac{1}{10}$ is within its cushion $n \cdot \frac{2}{5}/(n-1)$ (Theorem 1(c));*

- *the bystander and the dummies stay positive:* $\mu_2(s) = \frac{(1/4)^2}{(n-1)^2} - \frac{1}{100n^2} > 0$ *and* $\mu_{\text{aux}}(s) = \frac{s^2}{(n-1)^2} - \frac{s^2}{n^2} > 0$.

Hence invisible framing; the framed peer's conditional mean strictly negative while every other stream, the deviator's included, lies inside the composite null — exists at every $n \geq 3$.

Proof. All four displays are (8) with the stated coalitions and offsets: for $j \neq 3$ the leave-j-out coalition retains the deviator, so $\bar{b}_{S\setminus j} - p = (s - w_j)/(n-1)$ (using $\sum_k w_k = 0$); for the deviator, $\bar{b}_{S\setminus 3} - p = -w_3/(n-1)$; and $\bar{b}_S - p = s/n$. The framed sign in one line: $|s - w_1| = \frac{1}{20}$, so $\mu_1 < 0 \iff \frac{1/20}{n-1} < \frac{1/10}{n} \iff n < 2(n-1) \iff n > 2$. The certificate factors the numerator as $-(n-2)(3n-2)$. The remaining signs are immediate ($\frac{2/5}{n-1} > \frac{1/10}{n}$ and $\frac{1/4}{n-1} > \frac{1}{10n}$ for all $n \geq 2$; the dummy display is positive because $n - 1 < n$). □

Proposition (E2: attacker-free cascade at every $n \geq 3$; termination at $n \geq 4$)**.** *All truthful full-coalition marginals in* E_n *are* ≥ 0 *(squares). Suppose agent 2 (*$b_2 = \frac{3}{20}$*) is excluded* for any reason*; a false alarm within the* δ *budget, operator action, timeout. In the survivor coalition* $S = \{1, 3\} \cup \{$*dummies*$\}$*, all truthful,*

$$\mathbb{E}\,\bar{\Delta}_3^S = \left(\tfrac{1/20}{n-2}\right)^2 - \left(\tfrac{7/20}{n-1}\right)^2 = -\frac{(6n-13)(8n-15)}{400\,(n-1)^2(n-2)^2} \;<\; 0 \quad \textit{for all } n \geq 3 :$$

a consistent detector, coupled to exclusion, removes honest agent 3 within the delay bound at gap $g_n = \Theta(1/n^2)$*. Then, for* $n \geq 4$*, in* $S' = \{1\} \cup \{$*dummies*$\}$*,* $\mathbb{E}\,\bar{\Delta}_1^{S'} = 0 - \left(\frac{1/20}{n-2}\right)^2 < 0$ *(the dummy-only coalition sits exactly at* p*), so honest agent 1 is excluded next and the terminal coalition is all-uninformative: the cascade removes every informative agent. For* $n = 3$ *the second step reaches the lone-survivor anchor of X1 and is not claimed.*

Proof. After excluding agent 2, $\sum_{k \in S} w_k = -\frac{1}{20} + \frac{2}{5} = \frac{7}{20}$, so $\bar{b}_S - p = \frac{7/20}{n-1}$ and $\bar{b}_{S\setminus 3} - p = \frac{-1/20}{n-2}$; apply (8). The sign in one line: $\frac{1/20}{n-2} < \frac{7/20}{n-1} \iff n - 1 < 7(n-2) \iff 6n > 13$, true for $n \geq 3$. The certificate: the numerator factors as $(n-1)^2 - 49(n-2)^2 = \big[(n-1) - 7(n-2)\big]\big[(n-1) + 7(n-2)\big] = -(6n-13)(8n-15)$, both factors positive from $n \geq 3$. For the second step, $\bar{b}_{S'} - p = \frac{-1/20}{n-2}$ while the leave-1-out coalition is dummies only, offset exactly 0. Margins: both gaps are $\Theta(1/n^2)$; the same dilution as every marginal of the unweighted coalition mean, so *relative* to typical truthful marginals they are constants; the detection delay at gap $\Theta(1/n^2)$ grows like $n^4 \log(\cdot)$ — cascades slow with n but never disappear. □

Scope of E1–E2. This is an existence theorem (one explicit family per n) matching the body's claim "for every $n \geq 3$ there is an explicit profile"; it uses the unweighted coalition mean, reference-exclusion semantics, and the fixed anchor, exactly as in the sweep protocol below. Nothing is claimed for re-centered references: exclusion semantics that remain incentive-sound under their own dynamics are the paper's stated open design problem.

The $n = 3$ **exhaustive sweep (breadth evidence).** Protocol, fixed before the sweep: reference exclusion, leave-one-out marginals in the current active coalition, anchor as above; components $b \in \{0.05, 0.10, \ldots, 0.95\}^3$ (6,859 instances $\times$ 2 target peers), deviation grid $s \in [-0.5, 0.5]$ in steps of 0.01; "invisible framing of peer j" requires $\mu_j < 0$ with the deviator's and the third agent's means ≥ 0 in the full coalition. All quantities are exact expectations via (8); the sweep is an exhaustive evaluation over its finite grid rather than a simulation. Findings: (i) invisible framing exists in 9,060 (instance, target) pairs; (ii) in 96% of them (8,736/9,060) the framing recursion *closes*: after the framed peer's exclusion, the surviving honest agent can itself be invisibly framed in the reduced coalition, and the deviator reaches monopoly through the detector, exclusion by exclusion; (iii) 269 instances cascade with *no attacker at all* (e.g., $b = (0.45, 0.15, 0.9)$: after agent 1 is excluded for any reason, honest agent 2's truthful marginal in the survivor coalition is $-0.028 < 0$). The structural reason, in one sentence: exclusion breaks correct specification; the surviving truthful coalition's mean no longer equals p, so the subgame reopens with a mis-centered reference and live first-order channels. Recursion-to-monopoly beyond the first two steps remains $n = 3$-verified only (Section 4.2).

I ENDPOINT-LOCKED DEPLOYED REPORT MAP

Proof of Proposition 1. Only $v_t(N) = -(\bar{r}_t - y_t)^2$ depends on c (the reference coalition $V(N \setminus i)$ contains no term in r_i), through $\bar{r}_t = \bar{p}_t + \frac{c}{n} d_{it}$; differentiate at $c = 0$. The orthogonality follows by conditioning on reports: $\mathbb{E}[(\bar{p}_t - Y_t)d_{it}] = \mathbb{E}[(\bar{p}_t - \mathbb{E}[Y_t \mid \cdot])\, d_{it}] = 0$. □

The map, and why the mask is its true Jacobian. The trace arm's deployed deviation family is *not* the Euclidean clip $\Pi_{[0,1]}(p_0 + c\,d)$ (whose endpoint directional derivatives are one-sidedly nonzero). The deployed report map is

$$r_0(c) = \mathrm{clip}\big((1-c)\,p_0 + c\,\bar{p},\ \epsilon,\ 1-\epsilon\big), \qquad \epsilon = 10^{-4},$$

applied to raw scores p_0 that, for a large share of flows, sit at exactly 0 or 1 — *outside* the clamp interval. Such a flow is pinned at the clamp on a *two-sided* neighborhood of $c = 0$: an outward deviation pushes the pre-clip value further out, and an inward deviation unpins only at $c_{\mathrm{unpin}} = (\epsilon - p_0)/(\bar{p} - p_0) \geq \epsilon$ (measured minimum 3.2×10^{-4} on the trace). The two-sided directional derivative at $c = 0$ is therefore *exactly zero* for boundary flows; the "endpoint-locked deployed report map" — and the interior mask $\mathbf{1}\{p_{0t} \in (\epsilon, 1-\epsilon)\}$ used in the deployed direction $d_t = (\bar{p}_t - p_{0t})\,\mathbf{1}\{\cdot\}$ is the map's true local Jacobian, not a data-deletion choice. On the trace, 77.3% of flows are derivative-pinned (the RF/GBM agents emit hard $0/1$ scores — itself a transport observation). The mask is reports-measurable, so the moment restriction of Proposition 1 applies to the masked direction unchanged; the unmasked, convention-free direction is reported alongside it in Section 6, and the transport-failure verdict is the same (and stronger) without the lock.

Three-way verification of the closed form. (i) *Analytic:* the two-sided lock and the unpin threshold follow from the display above. (ii) *Replay-derivative ladder:* a 13-replay central-difference ladder through the deployed mechanism reproduces the closed-form certificate to 3×10^{-6} relative ($g_{\mathrm{pay}}(0) = -15.4828$ closed-form vs. -15.4828 replayed); the ladder's two-sided agreement at step $h \leq 10^{-4}$ and its one-sided asymmetry at $h = 10^{-2}$ (-28.19 toward-pool side vs. -15.04: the toward-pool side progressively unpins inward flows, the other side cannot) are both reproduced by the closed form with the clip in place (-28.1899 / -15.0392 / -15.4833, to four decimals). (iii) The lock-radius bound explains the onset of two-sided convergence at $h \approx c_{\mathrm{unpin}}$. The bootstrap intervals quoted in Section 6 (epoch-block bootstrap, $B = 2000$, 40 blocks, fixed seed) are computed on these certificates; the per-flow normalization of the masked certificate is $[-9.5, -6.2] \times 10^{-4}$. We use "endpoint-locked deployed report map" throughout and never bare "clip."

J PREREGISTRATION AND MACHINE VERIFICATION

J.1 PREREGISTRATIONS, AS FROZEN BEFORE EXECUTION

Preregistration of the τ-frontier. The frontier of Section 6 (the first of the paper's three preregistered experiment rounds; the E1–E4 round and the S1/S2 supplement follow below) was frozen before any run (design file committed; commit hash is the timestamp), with the following declared in advance. *Hypotheses:* (P) every population cell satisfies the best-response bound of Theorem 3; (T) some pre-declared $\tau^* < 1$ makes the trust region a practically useful remedy under the gate below — otherwise the paper reports only the continuous price curve. *Design:* projection onto $\mathcal{W}_\tau = \Delta^{n-1} \cap \{\frac{1}{2}\|\hat{s} - u\|_1 \leq \tau\}$ applied to published serving weights only (settlement untouched); τ grid $\{0, .005, .01, .02, .05, .1, .2, .5, 1\}$; externality scales $L \in \{0.1, 1, 10\}$, all reported. Population arm: the closed-form instances $b = (.3, .7)$, $(.2, .6, .7)$, $(.1, .35, .55, .8)$, exact best responses on a deviation grid of step 5×10^{-5}, with a declared *stop rule*: any cell exceeding the bound by more than one grid step aborts the experiment as a proof-or-code defect (the arm is exact; sampling noise is not an admissible explanation). Trace arm: the deterministic CICIDS2017 pipeline with all pins frozen — 60,000 flows drawn with a fixed seed; a seeded permutation assigns half (30,000) to train the three feature-partitioned agents (each sees 25% of the features) and holds the other half out as the replay stream, whose first $40 \times 500 = 20{,}000$ flows form the 40 settlement epochs of 500 (window 500, incremental rolling settlement, $\alpha = 1$); the deviation families are the *toward-pool* family $r_0 = (1-c)p_0 + c\bar{p}$, $c \in \{0, 0.05, \ldots, 1\}$, and a fixed-shading comparator. *Metrics* (all τ reported, no truncation): attacker utility gain, weight capture, benign served cost at the fixed operating threshold, served expected calibration error (ECE; 15 bins, plus a logistic calibration fit to defuse

binning disputes); paired bootstrap with fixed seeds and simultaneous sup-t bands across the τ grid. *Remedy gate* (declared in advance, no post-hoc edits): $\geq 50\%$ attack reduction at $\leq 1\%$ benign cost and $\leq +0.01$ ECE, robust at the unfavorable band edges; the gate **fails** on the trace, and the paper says so. The transport-moment certificate and the endpoint audit (Section I) are *post hoc diagnostics* triggered by the gate failure, labeled as such in Section 6.

The E1–E4 bridge and audit (second preregistration). The two-regime study of Section 6 was governed by a second frozen chain: a preregistration (design, estimands, grids, seeds, splits, criteria P1–P5, stop rules, and the utility decomposition $\Delta U = \Delta U_{\text{settle}} + \Delta U_{\text{cont}} + \Delta U_{\text{excl}}$) committed before any run; one amendment, committed after the prespecified smoke stage and before any held-out attack-utility outcome, replacing the statistically unidentified finite-window operator-FD direction estimator ($\hat{a}_1 = 0.39$, 95% CI $[-19.3, 20.0]$ — retained as a mandatory diagnostic) by the deterministic Corollary-1 population coefficient at the training-half population operating point; and a *post-result presentation decision* that moved E4's summary into the main text — changing presentation only: E4's symmetric grid, directions, decomposition, detector family, and interpretation rules were fixed before execution, E4 is not counted toward P1–P5, and all E1–E3 negative outcomes are retained in the main text. Scope diagnostics (frozen; never pass criteria): truthful all-active window frequency $3/20$ (train) and $10/40$ (all); windows with a clipped negative marginal $30/40$; truthful w_1 range $[0, 1]$ with quartiles $(0.03, 0.32, 0.59)$ and boundary share 27.5% — quantifying why Corollary 1 cannot transport unconditionally to finite windows.

E4's full masked grid (unmasked nearly identical; complete tables in the artifact). The own and peer columns are *single interventions* from the truthful baseline (deviated own marginals with truthful peers, and vice versa) through the nonlinear window map, so an interaction remainder $\text{rent}_{\text{int}} := \text{rent}_{\text{tot}} - \text{rent}_{\text{own}} - \text{rent}_{\text{peer}}$ remains; it stays within ± 0.006 on the anti-pool half (-0.003, CI $[-0.005, -0.001]$, at the reported $c = -0.2$) and grows to -0.340 at $c = +0.50$:

c	settle	rent_{tot}	rent_{own}	$\text{rent}_{\text{peer}}$	rent_{int}
-0.50	$+62.18$	$+0.126$	$+0.258$	-0.126	-0.006
-0.20	$+60.68$	$+0.200$	$+0.181$	$+0.022$	-0.003
-0.05	$+59.30$	$+0.096$	$+0.065$	$+0.032$	-0.000
0	$+58.61$	0	0	0	0
$+0.05$	$+57.15$	-0.383	-0.174	-0.207	-0.003
$+0.20$	$+52.55$	-1.701	-0.764	-0.889	-0.049
$+0.50$	$+42.43$	-5.175	-2.351	-2.484	-0.340

Detector alarms: zero at every grid point, both maskings (1.26M settled agent-samples processed per masking: 3 agents $\times 21$ grid points $\times 20{,}000$ samples, a cumulative processing count rather than independent trials; restart-mixture detector, $\delta = 0.05$, $B = 1$). The earlier raw-trace τ sweep's two-panel figure (population arm 81/81; price curve and failed remedy gate) is retained as Figure 3.

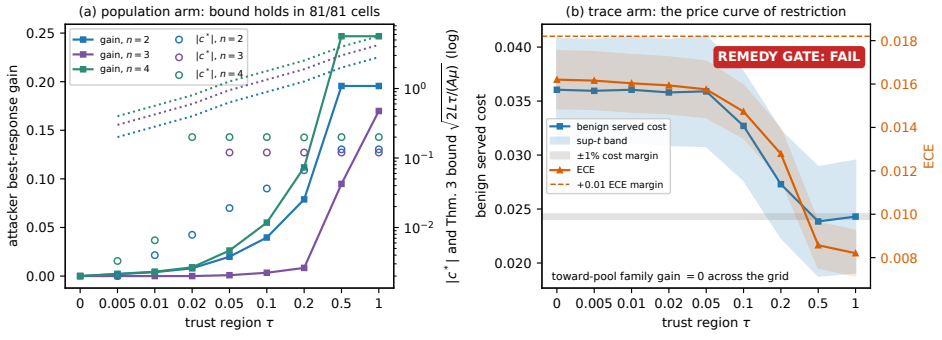

Figure 3: The original preregistered raw-trace τ-frontier (superseded in the main text by the two-regime study; retained for the record). (a) Population arm: the served-output-corollary bound holds in 81/81 cells. (b) Trace arm: benign cost of τ; the preregistered toward-pool family shows zero gain and the remedy gate fails.

Third preregistration: realized operator gradient (S1) and detector positive control (S2). A third frozen supplement (design, estimator, seeds, grids, and interpretation rules committed before execution) closes the gap between Theorem 1's finite-window object and E1's population surrogate. *S1* estimates the realized gradient $g = \mathbb{E}[\, dU_{\mathrm{rent}}/dc \,|\, c = 0]$ with $U_{\mathrm{rent}} = \sum_{e=20}^{39} w_{e,1}$ by label-conditional (Rao–Blackwell, RB) integration: within each window the summand integrates the *joint product* of the recomputed window statistic, the a.e. weight-map Jacobian row, and the single-sample response over $Y_t \in \{0, 1\}$ at its conditional probability (factor-wise expectations are exactly the decoupling error of Section 3), with E1's windows, map, cadence, and deviation parametrization $r_1 \mathrel{+}= c$ unchanged. Implementation checks (preregistered and not outcome-contingent): on an enumerable 2×6 mini-world the estimator's outer expectation equals the exhaustively enumerated realized operator-sum exactly (gap 0.0), and the realized operator-sum matches central differences at the $O(\delta^2)$ rate (max relative gap $1.9 \times 10^{-4} \to 1.9 \times 10^{-6}$ as δ shrinks $10\times$); on the full world, $\mathbb{E}_Y[D_1] = 0$ holds per sample to 6×10^{-17} (labels are drawn from the pooled posterior, so $\bar{r}_N = \bar{p}_t$ exactly at $c = 0$), which makes the own term equal its covariance form $\mathrm{Cov}(J_{11}, D_1)$ identically (verified to 1.4×10^{-14}); the exposure multipliers (window length, cadence) are deterministic here.

Differentiability at $c = 0$ and the interchange. The label space is finite and the label law does not depend on c (the sampler uses truthful $\bar{p}_t$), so the estimand is a *finite* sum with c-independent weights, and exchanging d/dc with the expectation requires only that each positive-probability path's $U_{\mathrm{rent}}(c)$ be differentiable at $c = 0$; no dominated-convergence argument is needed. A path fails this only if some window-mean marginal coordinate is *exactly* zero at $c = 0$ (a clip kink of the a.e. Jacobian) or the all-clipped uniform fallback triggers; discreteness means such an event would carry positive probability, so we audit it directly rather than appeal to "almost everywhere." Over all 301 paths entering the estimates (200 gradient, 100 utility, the frozen path) and every conditional label configuration the estimator touches, the closest coordinate-to-kink distance is 1.2×10^{-10} (four orders above the $\sim 10^{-14}$ floating-point summation scale: a well-signed nonzero value rather than a masked exact zero), the minimum positive-part normalizer is 1.05×10^{-3} (bounding every Jacobian denominator), and the fallback never triggers: every audited path derivative and every RB summand is defined. An exact kink at $c = 0$ would require an exact algebraic coincidence among the cached posterior reals; none occurs on the audited paths. The audit covers the configurations the estimator actually computes, not the entire label support, so the identification $g = \mathbb{E}[dU_{\mathrm{rent}}/dc]$ stays conditional on differentiability along every positive-probability path. Clip-state flips across conditional label values (3.1% of configurations) change the active set *between* configurations and leave the differentiability of every single configuration intact; central-difference stencils at $\delta \geq 0.002$ can straddle such kinks, which is why the finite-difference comparison is an expectation-level (enumeration) identity rather than a per-path equality.

Estimates. All intervals below are 95% normal-approximation intervals, mean $\pm 1.96\, s/\sqrt{K}$, with one label path (one seeded random-number stream) as the independent unit, conditional on the frozen prediction cache and truthful reports. Over $K = 200$ fresh label paths: own $+0.364$, CI $[+0.079, +0.650]$; peer -0.308 $[-0.472, -0.143]$; the *paired* total (own+peer summed within each path before averaging rather than a combination of the component intervals) $+0.056$ $[-0.306, +0.419]$. The components carry opposite signs and largely cancel; the total's sign is unresolved at this precision, and its interval excludes the population surrogate $20\, \psi^{\mathrm{tr}}_{\mathrm{pop}} = -2.58$ under the same deviation parametrization, horizon, and aggregation scale, conditional on the frozen prediction cache (the surrogate is computed on the same fixed data); the peer component is roughly $8\times$ smaller in magnitude than that population-scale prediction. The estimator's variance is $22\times$ below the naive realized operator-sum; clipping is active in $14.4/20$ windows on average, clip-state flips occur in 3.1% of conditional label configurations, and the uniform fallback never triggers. A disjoint-seed utility check ($K = 100$, a random-number stream disjoint from the gradient set's) is consistent: the odd (linear) part of $\Delta U_{\mathrm{rent}}(\pm c)$ matches the RB prediction within its intervals and excludes the population-scale slope for $c \geq 0.02$, while the even part $\approx -800c^2$ reproduces E1's symmetric-negative curvature (the held-out grid's weighted-LS fit of the same world reads $\approx -730c^2$: same object, different estimator and flow set). On the single frozen E1 label path the same estimator gives own $+0.76$, peer -0.74, total $+0.02$: the path underlying Figure 4(a) sits near the ensemble mean.

S2 is a positive control of the deployed restart-mixture detector (identical code path, $\delta = 0.05$, $B = 1$, threshold $n/\delta = 60$ as deployed) on streams $X_t = \mu_t + \mathrm{Unif}(\pm 0.9)$ with frozen changepoint positions $\nu \in \{0, 5000\}$, drift grid, and horizon 20,000.

Support condition. The safety lemma is stated for $Z_k \in [-B, B]$, but both proof steps use only the conditional *range*: Hoeffding's lemma with range $2B$ (safety) and Azuma increments bounded by $2B$ (delay), plus the conditional-mean sign. The S2 streams have conditional support $[\mu_t - 0.9, \mu_t + 0.9]$, of length $1.8 < 2B$, at every cell; for $g \in \{0.2, 0.5\}$ the post-change support leaves the symmetric interval $[-1, 1]$ while the range condition still holds. The implementation carries no step that uses $|X_t| \le B$ beyond the $\lambda^2 B^2/2$ compensator, and the stream is nowhere clipped (the only `clip` in the detector caps $\log M$ at ± 60 against floating-point overflow, perturbing e-values by $< 10^{-25}$ relative to the threshold 60), so the post-change conditional mean is exactly $-g$. The control therefore validates the implementation under the range condition its proofs actually use rather than under the verbatim symmetric-support phrasing; the lemma statement's $[-B, B]$ remains sufficient for the theory as written.

Cells. Null cells ($\mu \in \{0, +0.04\}$, 300 runs each): 0 alarms, a finite-horizon empirical false-alarm result; uniform anytime safety is supplied by the theorem rather than by this experiment. Change cells (100 runs each; delay quantiles are over *detected* runs only, measured from ν; censored = no alarm by the horizon; no pre-change alarms occurred in any cell):

g	ν	detected	censored	delay $q_{25}/q_{50}/q_{75}$
0.0133	0	0/100	100	–
0.0133	5000	0/100	100	–
0.02	0	0/100	100	–
0.02	5000	0/100	100	–
0.05	0	96/100	4	10184/11398/13193
0.05	5000	38/100	62	12587/13547/14366
0.1	0	100/100	0	2042/2612/3128
0.1	5000	100/100	0	3772/4277/4687
0.2	0	100/100	0	471/592/683
0.2	5000	100/100	0	838/983/1097
0.5	0	100/100	0	64/76/90
0.5	5000	100/100	0	135/152/176

Median delays at $\nu = 0$ (2612/592/76 for $g \in \{0.1, 0.2, 0.5\}$) decrease in g consistently with the $g^{-2} \log$ trend (we do not claim a verified scaling law); delays from the late changepoint are uniformly longer (dyadic restart-window alignment). $g = 0.05$ is transitional, its censoring driven by the remaining horizon. Drifts $g \in \{0.0133, 0.02\}$ (the smaller being the worked-instance drift scale) are censored in 100/100 runs at this horizon. Extrapolating the observed delay trend to $g = 0.0133$ suggests order 10^5 samples to detection; this is an extrapolation from the fitted trend rather than a measured detection time. The control validates the implementation and maps its power at this horizon on these test processes; it does not establish that any particular replay's silence (E1's P5, E2's, or E4's) is a power phenomenon, since those streams need not satisfy the corresponding negative-drift condition.

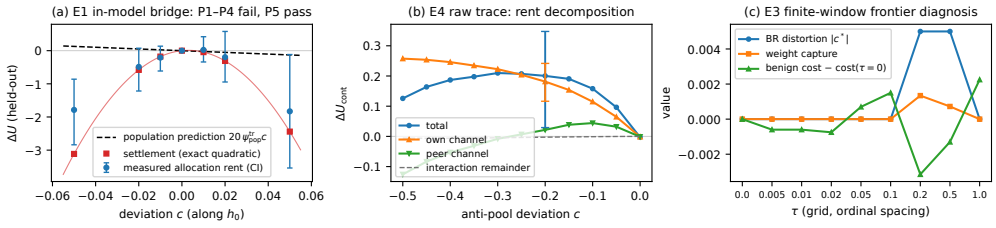

Figure 4: The E1–E4 bridge panels (previously the main figure; retained unchanged). (a) E1: population prediction $20\,\psi^{\mathrm{tr}}_{\mathrm{pop}} c$ (dashed) vs. measured allocation rent (CIs) and the exactly quadratic settlement; P1–P4 fail, P5 passes. (b) E4 rent decomposition. (c) E3 finite-window frontier diagnosis.

Fourth preregistration: the R4 operating-point round. A fourth frozen design (grid, seed blocks, two-stage selection, claim hierarchy, and interpretation rules committed before the first screening path; per-map implementation gates archived: exact enumeration identity $\le 2.8 \times 10^{-14}$

over 30 (map, agent, world) combinations, FD $O(\delta^2)$ ladders, kink/overflow audits, permutation and replay asserts, the settlement first-order stop rule re-run per (world, smoothing), all passing; the UNSW-NB15 world is built by the identical feature-partitioned pipeline and cached canonically, rebuilds agreeing to 3×10^{-16}). *Stage A* screens 96 cells at $K{=}800$ paths (shared seed block; screening quantities only; the full per-cell table and z-histogram ship in the artifact). Registered directional predictions all held: the UNSW world (clip-degeneracy removed) and the softmax family (non-homogeneous, breaking the Euler annihilation $J(\bar{m})\bar{m} = 0$ of the degree-zero clip map) de-cancel, and the fully pinned agent carries the cleanest own channel. *Stage B* confirms the six top-ranked cells on a disjoint seed block ($K{=}1600$, Bonferroni-6 99.17% t-CIs, sign agreement): all six confirm, including the deployed clip map on both traces (UNSW $+3.82$ $[3.75, 3.88]$; CICIDS $+1.79$ $[1.66, 1.92]$). *Stage C* (paired $\Delta U_{\text{total}} = \Delta U_{\text{settle}} + \kappa\,\Delta U_{\text{rent}}$, $\kappa = 1$, $K{=}1600$ with a variance-triggered extension, fresh seeds): at the preregistered sizes $c \in \{\frac{1}{4}, \frac{1}{2}, 1\}\, c_{\text{demo}}$, with c_{demo} capped by the *rent* curvature, no confirmed cell is profitable at $\kappa = 1$: no ΔU_{total} CI excludes zero from above at any tested size. The three UNSW totals are significantly negative (the settlement's second-order cost over all settled flows dominates; resolved breakeven κ^* 1.3–1.9 at the smallest size, 3.1–4.4 at the middle, rising with c), while the three CICIDS totals are statistically indistinguishable from zero at the smallest size, their component ratios too noise-dominated to support a breakeven estimate. The deployed detector is silent on deviant and truthful streams at every confirmed cell ($0/600$ per class); the beyond-cushion control at $1.5\hat{c}$ also raises no alarm; at these operating points even the beyond-cushion deviation induces subpower drift, so the cushion is not the detection boundary at this horizon (power is, consistent with S2); the dead-zone negative control returns a machine-zero gradient. A follow-up micro-round (frozen separately; its design dependence on Stage C's negative result is disclosed in the preregistration and here) tests profitability at the *total*-curvature optimum $c^* = \hat{g}/(2\hat{K}_{\text{tot}})$ on fresh seeds at $K{=}6400$, grid $\{\frac{1}{2}, 1, 2\}c^*$, Bonferroni-3 98.33% t-CIs on the three primary (UNSW) cells. All three pass at c^*: clip $\Delta U_{\text{total}} = +0.00136$ $[+0.00069, +0.00202]$ ($c^* = 7.0 \times 10^{-4}$; rent $+0.00251$, settlement -0.00115); softmax 2 $+0.00184$ $[+0.00106, +0.00262]$; softmax 8 $+0.00257$ $[+0.00159, +0.00354]$; at $2c^*$ every total falls back toward zero, the measured signature of first-order rent against second-order settlement cost. The secondary (CICIDS) cells behave as predicted: clip and softmax 2 are positive with CIs excluding zero at their (smaller) c^*, softmax 8 sits below resolution. The deployed detector at c^* raises no alarm on any deviant or truthful stream of the three passing cells ($0/600$ stream-runs): the demonstration is profitable and silent simultaneously. Magnitudes are small in absolute terms ($\sim 10^{-3}$ held-out weight-sum units per horizon at $\kappa = 1$); their economic weight scales linearly with the allocation stake χ, exactly the coefficient (2) leaves as a deployment primitive.

J.2 Round results reported here rather than in the body

κ-sensitivity and units. $\Delta U_{\text{total}} = \Delta U_{\text{settle}} + \kappa\,\Delta U_{\text{rent}}$ is linear in κ, so every reported cell carries an exact breakeven. At the total-curvature optimum c^* the three primary (UNSW) cells break even at $\kappa^* = 0.43$–0.46 (secondary CICIDS cells 0.30–0.56): profitability at c^* needs only $\kappa \gtrsim 0.5$, while at the larger preregistered sizes the breakeven rises to 1.3–4.4 (Stage C). All profits are stated in normalized units: one unit is one held-out weight-sum per horizon, priced by the deployment stake χ. Under the explicit assumption that a unit of served weight over the horizon is worth V (the serving revenue a contributor attributes to being served), the deployed-clip gain at c^* is $0.0014\,V$ per horizon at $\kappa = 1$; we measure no real-system χ or V and claim none.

Engineered framing instance and its power gap. On a constructed asymmetric world (victim agent's reports compressed to $(0.45, 0.10)$; labels from the pooled mean of the actual reports, so the envelope holds and the victim is viable at truthful), the echo deviation $r_1 \to (1 - s)r_1 + s\,r_2$ at $s = 0.5$ satisfies the frozen E2 framing rule (victim train mean marginal $-0.0038 < 0$, deviator $+0.0004 \geq 0$) on a bounded s-window, as Theorem 2's framing region prescribes. Causal controls: zero truthful alarms at $B = 1$ and at the range-matched $B = 0.3$ (25 paths each). The induced victim drift (0.0038) is 1–2 orders below the range-valid detector's in-horizon power (S2 map), so no exclusion fires; the weight-inheritance mechanics are therefore reported counterfactually: excluding the victim at the first held-out epoch raises the deviator's held-out rent by $+0.023$ $[0.014, 0.032]$. This is an existence demonstration on an engineered world; E2's null on the symmetric world stands.

Design-point and attribution diagnostics. *WSU update* (registered prediction: gradient ≈ 0, the $a_i = 0$ escape class): the original diagnostic inadvertently used normalized multiplicative loss instead of the centered WSU update of Freeman et al. (2020). A post-hoc code-identity audit applies the original update to epoch losses on the same label paths: its four finite-difference estimates remain significantly nonzero (-0.037 to -0.563; the earlier proxy gave -0.037 to -0.606). On the two R4 headline cells, matched-seed ± 0.01 differences for WSU at $\eta \in \{0.1, 0.5\}$ are UNSW $+0.276/+1.361$ and CICIDS $+0.049/+0.241$, versus clip $+3.942/+2.021$; all four paired clip-minus-WSU intervals exclude zero. Thus the original WSU update reduces but does not remove the measured rent response in this pooled-component, delayed-settlement model; its individual-belief IC guarantee does not transfer to this different utility. The registered prediction failed under both implementations. *Window-length sweep* (clip, frozen smoothing, agent 1, K=400): the own channel dies with m (CICIDS $+3.18 \to -0.08$; UNSW $+1.29 \to +0.01$) while the peer channel dominates at large m; the CICIDS total changes sign across $m \in \{100, 500, 2000\}$ ($+8.27/-0.11/-0.54$), placing the frozen cell at the crossover (Figure 2c). *E1 power*: weighted LS on the frozen held-out grid gives slope SE 6.8, minimum detectable slope 19.1 at 80% power: the population prediction 2.6 was below the grid's power, and the RB estimator's SE 0.185 is the 37× identification fix. *E4 ledger*: served decision quality under the anti-pool deviation is essentially unchanged (ΔBrier $+2 \times 10^{-5}$ at $c = -0.05$ to $+2.9 \times 10^{-4}$ at -0.5; accuracy shifts $< 10^{-3}$); the affine (constant-Jacobian) replay collapses the own rent channel from $+0.181$ to $+0.0014$ at $c = -0.2$; the deployed detector remains silent on the E4 grid at range-matched $B = 0.1$. *Second-trace transport certificate* (UNSW-NB15, identical pipeline): -78.8, epoch-block bootstrap 95% CI $[-84.2, -73.3]$ (unmasked -121.6): the pooled-posterior specification fails on both tested traces, and Proposition 1 functions as a falsification instrument across traces.

Fifth preregistration series (frozen 2026-09-19). Three follow-up designs were frozen together before execution (grids, selection rules, criteria, and outcome wording committed with the implementation; non-outcome gates archived): an end-to-end detector–exclusion–inheritance run on an engineered extended-horizon world, a naive weight-stream-decoupling analysis, and segment-level transport certificates. Results below.

End-to-end detector–exclusion–inheritance run. Frozen closed-form selection over the engineered family ($A_2 \times B_2 \times s$, 205 cells): two candidates satisfy the E2 framing rule with in-power predicted drift; the selected cell compresses the victim to a constant report (A_2=0.55, B_2=0; viable truthful, train marginal $+0.023$) with echo s=0.5, induced drift 0.0052, range-valid B=0.3, and horizon tiling T=16 (320k flows; probe guard: five disjoint paths, all negative, mean within factor two). Arms (paired CRN, K=100): the detector fires on the deviated victim stream on every path (delay quartiles 117k/126k/144k); exclusion from the alarm epoch's next weight index; the deviator's held-out rent gain over alert-only is $G = +0.133$, 95% CI $[+0.127, +0.139]$, and the excluded honest victim loses its held-out rent. Causal controls at the same B: 0/25 alarms on every truthful stream and on the deviator's own deviated stream. Serve-only ledger: settlement is identical across arms on the shared path, so $\Delta U^{\text{excl}} - \Delta U^{\text{alert}} = G$ exactly; no-alarm paths satisfy the arm identity machine-exactly. Disclosure: this is an existence demonstration on an engineered, horizon-extended world; the deployed-horizon power gap and E2's symmetric null stand.

Naive weight-stream decoupling. Computing serving weights from an evaluation stream statistically independent of the settled stream (same reports, independent labels), with the deviator free to deviate there, changes nothing structural in its favor being closed: the rent-gradient functional is identical in law (measured: evaluation-block RB gradient 3.80, 95% CI $[3.71, 3.89]$, overlapping Stage B's $[3.75, 3.88]$), while the settlement second-order counterweight that made the preregistered sizes unprofitable is exactly zero for an evaluation-only deviation — breakeven $\kappa^* = 0$. The deployed detector on the deviator's evaluation stream stays silent at every tested size $c \in \{1, 4, 16, 64\}\, c^*$ at $B \in \{1.0, 0.1\}$ (0/25 each; 36% of flows clip at $64c^*$): the settlement-free deviation is undetected through 64× the coupled optimum. Naive temporal decoupling relocates the incentive problem and removes its economic counterweight; whether an evaluation stream can itself be made incentive-compatible is the open design question.

Defense-baseline round (frozen 2026-09-19, second series). Three remaining defense candidates, measured in the frozen worlds under a committed preregistration with prior non-outcome

gates (FD pipeline anchored on the clip map within 3 SE of Stage B). *Robust selection maps* (top-1 and median-agent serving, simplified selection rules rather than implementations of Krum): the registered annihilation prediction FAILS at the preregistered ± 0.01 scale for top-1 — the finite-scale rent secant is nonzero on both confirmed worlds (UNSW $+5.4$, 95% CI $[+1.9, +8.9]$; CICIDS $+5.4$ $[+1.5, +9.3]$; comparison with clip's own magnitude unresolved at this K): a finite deviation crosses selection boundaries and flips the entire served weight, so winner-take-all serving is gameable at this scale. This secant is not a derivative at zero: on each sampled path the discrete selection is locally constant away from ties. A post-hoc scale audit finds 218 and 268 top-1 selection flips at ± 0.01 across the 400 paths in the two worlds, but only 1 and 0 at ± 0.0001. The median-agent map is unresolved at $K{=}400$ (CIs cover both zero and clip scale). Serving cost: top-1 and median pay $+0.0093$ and $+0.0086$ Brier against the pooled clip map, while the dead-zone map — which does annihilate the gradient — serves slightly better than clip (-0.0025), consistent with E3's flat adaptivity price curve on these traces. *Allocation-side monitor*: the constant-drift approximation for the deployed anytime family gives $0.64B$ over the 20-point held-band weight trajectory as a planning scale, not a necessary detection threshold. An oracle-calibrated batch test (empirical 5% FPR on its 400 calibration paths, disclosed as non-anytime) has $106/1200$ false alarms on independent truthful paths in a post-hoc calibration audit. Its alarms at every size $c \in \{1, 4, 16, 64\}\, c^*$ ($6/13/8/4$ of 100; effects ≤ 0.25 SD) do not establish useful power at these sizes; the calibration result rules out a nominal 5% false-alarm claim. *Two-agent collusion* (joint additive deviation at $\{1, 2\}c^*$, $K{=}1600$ paired): sub-additive at both sizes (additivity gap -0.0047 and -0.0187, paired CIs excluding zero, deepening with size — the colluders share the pool-shift penalty), and the joint total sits below the best solo deviation; detectors silent on all streams ($0/25$ each). This is a transferable coalition-sum metric: at c^* the two individual joint gains are $+0.078$ and -0.038, so the joint action is not individually profitable for both without a side transfer. Collusion does not amplify the summed channel at these sizes; coalitions beyond two agents remain open.

Cross-n extension and Byzantine-robust aggregation proper (frozen 2026-09-21, third series). The worlds above carry $n{=}3$, which *excludes* Krum (Blanchard et al., 2017): it requires $n \geq 2f + 3$, so $n \geq 5$ at a single Byzantine agent, and it aggregates vectors rather than scalar marginals. We rebuilt the identical pipeline at $n{=}5$; because the agent factory draws one feature subset per agent in index order, agents 0–2 are the frozen agents (rebuild agreement 3.3×10^{-16}, the value the determinism gate already records; nesting $\leq 5.6 \times 10^{-16}$, against $O(0.1)$ for a substituted model), so the world *extends* rather than replaces. Krum is then implemented exactly (Eq. 2 of Blanchard et al., 2017: score = summed squared distance to the $n-f-2$ nearest report vectors, argmin serves), gated against an independent reimplementation and asserted undefined at $n{=}3, 4$. This is Krum's *selection rule* on report vectors; its Byzantine-gradient theorem does not transfer to delayed settlement and no transfer is claimed. *Four of six registered predictions failed.* (i) Krum and Multi-Krum weights are label-free, so the served allocation ignores realized accuracy entirely and the rent secant is exact rather than estimated (registered in advance). (ii) That secant is 0 at every registered scale — but as a **degenerate** null, not a defense: the registered deviator never wins a held-band epoch even truthfully (rent 0.0; winners are other agents), and escalation to $c{=}1.0$, the entire report range, never gives it rent. A post-hoc probe on an agent that *does* win shows rent *maximized at truthful* and decaying monotonically in $|c|$ to zero by $c{=}0.2$ in *both* directions (UNSW $12 \to 0$, CICIDS $7 \to 0$): on this probe Krum suppresses the additive channel by penalizing departure from the pool, i.e., by rewarding conformity. (iii) The registered direction prediction (positive secant, toward-pool sign $+1$ frozen beforehand) and the Multi-Krum non-annihilation prediction both **failed**. (iv) Welfare ordering **failed**: Multi-Krum serves *better* than the deployed clip map on both traces (-0.0075, -0.0076 Brier, paired t_{24} CIs excluding zero), Krum is mixed ($+0.0031$, -0.0044), only top-1 is uniformly worse. (v) Three-agent collusion is sub-additive at both sizes (gaps -0.0037, -0.0148, CIs excluding zero) with *two of three* members individually worse off ($+0.044$, -0.014, -0.015): still a transferable coalition sum, never a no-transfer Pareto improvement. *Most consequentially*, the confirmed rent channel does not survive pool growth on one trace: under the identical finite-difference estimand at both sizes, CICIDS moves $+1.66$ $[0.88, 2.44] \to +0.73$ $[0.34, 1.12]$ (same sign, smaller, as predicted) but UNSW moves $+3.61$ $[3.21, 4.02] \to -0.07$ $[-0.35, +0.22]$, a CI covering zero. The $n{=}3$ anchor $+3.61$ agrees with Stage B's Rao–Blackwell $+3.82$ under a different estimator and disjoint seeds, so this is not an estimator artifact. In this model the pool *is* the label law, so n and the posterior move together and the comparison is between two pool sizes rather than a controlled ablation of n; that confound was declared before execution. Nondegeneracy (S3) is therefore established *at the preregistered* $n{=}3$ *family*, and one of two traces loses it at $n{=}5$.

ELF-X adapted baseline (frozen 2026-09-21, fourth series). The forward-looking counterpart to the WSU update above is ELF-X (Freeman et al., 2020), introduced in the same paper, which proves WSU itself is *not* forward-looking incentive-compatible. It is an *unnumbered* two-step construction in that paper's Section 5, and we invent no equation number for it: per completed event τ draw a winner from $q_{i\tau} = \frac{1}{K}(1 - \ell_{i\tau} + \frac{1}{K}\sum_j \ell_{j\tau})$, count wins $N_i(t)$, and serve a uniform tie-broken $\arg\max_i N_i(t)$; the weight is $\pi_{i,t} = \Pr[i \text{ is that } \arg\max]$, the law of the whole randomized count process — not $q_{i,t}$, not $N_i(t)/t$, and not a multiplicative update. Implementation gates assert each of those three distinctions, that ELF-X differs from the earlier peer-only-mean ELF, that the first served weight is uniform (causal timing), and that a vectorized path reproduces the reference exactly; an independent audit additionally enumerated all 3^3 winner sequences of a $T{=}3$, $K{=}3$ instance and matched the closed-form π to 9.6×10^{-4} against 2×10^5 Monte-Carlo repetitions. π has no closed form in general, so it is estimated by Monte Carlo ($n_{\mathrm{mc}}{=}120$) with its own seed block, disjoint from the label block. The $\pm h$ arms of a path share that seed: common random numbers, which *couples* the arms and reduces the variance of their difference but does *not* cancel the Monte-Carlo noise, since the arms have different q and therefore different inverse-CDF maps; the residual randomness is carried in the across-path sample variance behind every interval below. *Result.* On the two confirmed clip cells, matched worlds, label paths, deviation family and $\kappa{=}1$: the registered direction prediction holds (all six secants positive, matching the pre-registered $\mathrm{sign}(\bar{p} - r_{\mathrm{dev}}) = +1$), but the magnitude prediction **fails** — ELF-X's secant exceeds clip's in five of six cells, and on UNSW the paired clip$-$ELF-X difference is -5.86 $[-8.15, -3.56]$ at $h{=}10^{-3}$ and -4.35 $[-5.00, -3.71]$ at $h{=}10^{-2}$, both excluding zero. The welfare prediction also **fails**: ELF-X serves *better* than the deployed clip map on both traces (-0.0024 $[-0.0025, -0.0022]$ and -0.0014 $[-0.0016, -0.0013]$, paired t_{24}). *Scope, registered before execution.* This is **not** evidence against that paper's Theorem 5.1. Its hypothesis is full information with each expert's own product-Bernoulli belief, and its utility is future selection probability; here labels come from the pooled posterior, so a component's truthful report is not its belief about the label law, and the utility is delayed settlement plus allocation rent. Both objects change, so the experiment measures *transfer* and nothing more; the source paper proves no regret bound for ELF-X, and none is claimed or tested here.

Segment-level transport certificates. Per trace, four disjoint 10-epoch blocks and both 20-epoch halves, each with the identical masked certificate and within-segment epoch-block bootstrap CI; a segment counts as envelope-compatible only as a powered null (CI covers zero with half-width below half the full-trace mean). Result: $0/12$ compatible — every segment's CI lies strictly below zero on both traces (CICIDS per-flow means -7.0 to $-8.5{\times}10^{-4}$; UNSW -3.7 to $-4.2{\times}10^{-3}$): the envelope violation is present at segment scale throughout, and the in-model mechanism's empirical support is confined to the replay family.

J.3 What each claim rests on

Claim–evidence–scope table.

Claim	Evidence	Scope
Noncomposition dichotomy (Thm 1)	proofs; verify scripts	smooth strictly proper settlement, (S1′)–(S3)
Cushion invisibility (Thm 1c,d)	proof; S2 power control; P5 run-level	$O(\epsilon)$-optimal deviations, own-stream family
Detection–action amplification (Thm 2)	proof; end-to-end detector–exclusion–inheritance run ($G = +0.133$)	engineered extended-horizon world; no in-power exclusion at the deployed horizon
Trust-region price (Thm 3)	proof; E3 frontier 9/9; remedy gate FAIL	cap-not-price; global strong properness
Rent-gradient nondegeneracy	R4: 73/96 screen, 6/6 disjoint-seed confirm	semi-synthetic replay family, both traces
Profitable and silent at c^*	16c micro-round (post-hoc frozen, disclosed)	normalized units $\kappa = 1$; $\sim 10^{-3}$ magnitudes
Envelope fails on raw traces	transport certificates -15.5 / -78.8	both traces, convention-independent
E4 deployment vulnerability	prespecified audit; affine control	out-of-model; not peer-channel validation

J.4 MACHINE VERIFICATION

Machine-verification inventory. Each entry is an *implementation check* of a stated closed form, in its stated normalization: not empirical evidence for any theorem, and none appears in the contribution list. The artifact's `README.md` maps every script to the result file it writes and the command that reproduces it; this table records only what is checked and to what tolerance.

Check	Result, and the tolerance it holds to
Worked instance (Section A)	exact polynomials, s^*, and the price-of-honesty bound
Proportional corollary (Section C)	$\bar{\Delta}_i^0$, μ_i, d_j, ψ_i^{pop}, cushion at $n \in \{2,3,4\}$, asymmetric components
Multiclass reduction	Lemma M1 vector/scalar agreement under the $\frac{1}{2}$-Brier convention, residuals $\leq 10^{-12}$; Theorem 1(b) against gradient ascent, error linear in χ (1.5% at 10^{-2}, 0.15% at 10^{-3})
Dynamic interfaces (Sections B and G)	exact 16-path enumeration ($n = 3$, $m = 2$, softmax and affine): operator-sum matches finite differences to 3.2×10^{-12} and 5.4×10^{-12}; decoupling gap 0.0280 absolute (229% of the exact value) for softmax and exactly 0 for affine; stationary factorization exact; $\Lambda/\omega = m$; Theorem 3's bound holds in $36/36$ grid cells
Cascades (Section H)	symbolic-n positive-coefficient certificates for E1–E2 ($n \mapsto 3+m'$), exact-rational numerics at $n = 3, \ldots, 12$, and the exhaustive $n = 3$ sweep over the full finite grid
Frontier arms (Theorem 3)	population arm $81/81$ cells for the served-output corollary in the single-publication $m{=}1$ normalization; Dykstra projection cross-checked against the $n{=}2$ closed form
Watermark map (Section G)	checked numerically for both deviation signs to $\lvert c\rvert = 0.15$ in the mean-statistic arm
Transport certificate (Section I)	13-replay derivative ladder and epoch-block bootstrap agree to 3×10^{-6} relative
E1–E4 bridge and audit	direct window-map computation against the settlement daemon to 10^{-12}; all replays deterministic on cached agent outputs
S1/S2 supplement	enumeration identity, $O(\delta^2)$ finite-difference convergence, Rao–Blackwell estimator, and the differentiability audit over all 301 paths (closest kink approach, denominator bound, fallback count)
R4 round, defense baselines, and the fifth series	implementation gates, landscape, confirmation, profitability, controls, and diagnostics, each with its own pre-outcome gates; deterministic given the pinned per-path seeds

Environments are pinned in the artifact (interpreter and library versions, seeds); all checks are deterministic given the pins.