Algorithmic Advice, Human Review, and Shared Liability

Wenxiao Yang
University of California, Berkeley

March 23, 2026

Abstract

Algorithmic advice is often used not only to guide decisions but also to determine whether cases are escalated for costly review. I study advice design at deployment for an algorithm provider that observes a calibrated internal score and sends advice to a human decision maker who retains final authority, can pay to open review, and shares misclassification losses with the provider. A startup cost of review makes advice part of the escalation rule: optimal advice design targets the posterior beliefs at which review switches on or off, rather than preserving fine distinctions across the score distribution. This endogenous coarsening can pool cases from opposite sides of the immediate-action cutoff and distort which cases receive review. Consequently, the combined human–algorithm system can underperform the better of the human-only and algorithm-only benchmarks, even without any behavioral frictions. Shifting liability toward the decision maker mitigates these deployment-stage distortions.

I Introduction

Consider a breast-imaging workflow. A model may compute a rich internal risk score, but the clinician typically sees a coarse interface: a flagged exam, a priority ranking, or a standardized assessment category. BI-RADS itself is organized around assessment categories tied to management recommendations, and FDA-cleared mammography AI products already provide prioritization outputs while leaving final interpretation to physicians (American College of Radiology, 2025; U.S. Food and Drug Administration, 2024). In such settings, the interface does not merely summarize information. It helps determine which exams are escalated for further review.

The same logic extends well beyond breast imaging. In many organizations, the output shown to the user helps determine whether a case is cleared immediately, denied outright, or sent for review. In these settings, advice is part of the escalation rule rather than just a forecast. The design problem is therefore how to communicate information when the interface itself helps determine whether costly review is opened. Hospitals, fraud teams, auditors, and compliance systems all face variants of this problem.[1]

To study this problem, I consider an algorithm provider that maps a continuous calibrated score into advice for a decision maker who retains final authority. After observing the advice, the decision maker either acts immediately or opens review before taking the final action. Review combines a startup cost with a marginal information-processing cost. Mistakes generate losses shared between the provider and the decision maker under an exogenous liability rule, so the provider internalizes only part of the downstream consequences of advice design.[2]

A natural intuition is that fuller disclosure should help. If extra detail is unhelpful, the decision maker can ignore it. That intuition is correct when review has no startup cost. With a startup cost, however, advice affects whether review is opened at all. Small changes in posterior beliefs can then switch review on or off, so the provider gains from moving cases to the posterior thresholds at which review starts or stops rather than from preserving fine distinctions across the score

[1] Other domains, including partially automated driving, also combine human oversight with formal reporting requirements and differentiated responsibilities across manufacturers, providers, deployers, and operators (National Highway Traffic Safety Administration, 2025a,b; EU AI Act, 2024; EU Product Liability Directive, 2024).

[2] Recent regulation and enforcement increasingly separate provider and deployer responsibilities. See the EU AI Act, the revised EU Product Liability Directive, and the Texas Attorney General's 2024 settlement with Pieces Technologies (EU AI Act, 2024; EU Product Liability Directive, 2024; Texas Office of the Attorney General, 2024; Mello and Guha, 2024; Price and Cohen, 2024).

distribution. A startup cost therefore makes gate-targeting disclosure optimal, and coarse advice arises endogenously even without any exogenous communication constraint.

The same mechanism also has a welfare implication. Conditional on a given posterior, opening review is weakly socially beneficial, so any welfare loss must arise earlier, at the advice stage. By pooling cases across the immediate-action cutoff, gate-targeting can mis-sort which cases are sent to review and thereby make the combined system perform worse than the algorithm-only benchmark. I refer to this distortion as a review trap.

Relative to the better standalone alternative, underperformance can arise in two distinct ways. At low decision-maker liability, the combined system may induce no review even though human-only review would still be worthwhile, because the decision maker internalizes only part of the loss in the combined system but all of it in the human-only benchmark. Once review becomes inducible, endogenous coarsening can still lower welfare through cross-cutoff pooling, so the combined system can also underperform algorithm-only deployment. Underperformance therefore does not require distrust, overreliance, or other behavioral frictions.

Holding score informativeness fixed, shifting liability toward the decision maker mitigates these deployment-stage distortions. Section VI then shows that this deployment-stage benchmark can fail once score informativeness is endogenous, because the same liability shift can weaken incentives to improve the score.

This paper relates to three strands of work. First, it contributes to costly verification, auditing, and information acquisition. The review technology nests posterior-separable acquisition when the startup cost is zero and introduces a distinct activation margin when the startup cost is positive (Sims, 2003; Caplin and Dean, 2015; Matejka and McKay, 2015; Mackowiak et al., 2023; Townsend, 1979; Mookherjee and Png, 1989; Halac and Yared, 2020; Byrne and de Roos, 2022).

Second, it relates to delegation, responsibility, and human–algorithm systems. The liability parameter formalizes how provider and decision-maker responsibilities are split, in the spirit of Holmström (1982) and Bartling and Fischbacher (2012). The model also provides a rational benchmark for why a combined human–algorithm system can underperform the better standalone alternative even without behavioral bias (Dietvorst et al., 2015; Buçinca et al., 2021; Vaccaro et al., 2024).

Third, and most closely, it relates to Matyskova and Montes (2023), who study persuasion

when the receiver can acquire additional information after disclosure. My setting differs in two economically important ways that drive both the disclosure result and the welfare result. First, review has a startup cost, so disclosure affects not only what is learned conditional on review but whether review is opened at all. That activation margin is what makes gate-targeting optimal and generates the mis-sorting distortion studied here. Second, the provider remaps a deployed calibrated score rather than designing an arbitrary Bayes-plausible signal. This keeps the focus on advice design at the point of use once a predictive technology is already in place, and makes implementability depend on the geometry of score coarsening. The paper also draws on Bayesian persuasion, especially work on optimal coarsening and convex-analytic methods (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019; Dworczak and Martini, 2019; Kolotilin, 2018).

The rest of the paper proceeds as follows. Section II presents the model. Section III characterizes the decision maker's review problem. Section IV studies optimal advice design with startup costs. Section V studies the resulting welfare distortions under shared liability. Section VI extends the analysis to endogenous score informativeness. Longer proofs are in the appendix.

II The Model

The analysis focuses on deployment, taking the predictive technology as given. The question is not how to build the score, but how to communicate it once the score is used to determine whether cases receive review. The provider observes the internal score and sends advice; the decision maker (DM) sees only that advice and then decides whether to act immediately or open review. Because the DM's review decision depends only on the posterior induced by the recommendation, I index the provider's policy by induced posterior beliefs.

II.1 Environment

The state is binary, $\theta \in \{\text{Good}, \text{Defective}\}$. Let $\mu_0 := \Pr(\theta = \text{Good})$ denote the prior probability that the state is Good. The provider privately observes a one-dimensional calibrated score $\omega \in [0,1]$, distributed according to F on $[0,1]$, where F admits a continuous density f with full support on $[0,1]$. Calibration means that

$$\Pr(\theta = \text{Good} \mid \omega) = \omega \qquad \text{for all } \omega \in [0,1].$$

Thus, ω is the provider's posterior probability that the state is Good conditional on the model's internal information. The DM does not observe ω directly; he observes only the recommendation sent by the provider. By the law of iterated expectations, $\mathbb{E}_F[\omega] = \mu_0$.

The DM ultimately chooses a binary action $a \in \{\text{Accept}, \text{Reject}\}$. Correct decisions yield normalized payoff zero for both players. Mistakes generate losses $D_{FP} > 0$ (accepting a Defective object) and $D_{FN} > 0$ (rejecting a Good object). The DM bears a fraction $\alpha \in [0,1]$ of any realized loss, and the provider bears the remaining fraction $1 - \alpha$. The analysis focuses on $D_{FP} \leq D_{FN}$ without loss of generality.[3]

II.2 Timing

The game unfolds in three stages. The separation between disclosure and review isolates the two margins of the decision maker's problem: the startup cost governs whether review is activated, while the marginal information cost governs learning conditional on review.

1. **Recommendation policy.** The provider chooses how to map scores into recommendations. Because recommendations matter only through the posterior beliefs they induce, the policy is indexed directly by those posteriors. For each score ω, let $\Gamma(\cdot \mid \omega) \in \Delta([0,1])$ denote the distribution of posterior beliefs delivered to the DM. Bayes plausibility requires

$$\int_0^1 \mu \, d\Gamma(\mu \mid \omega) = \omega \qquad \text{for every } \omega \in [0,1].$$

Aggregating over scores then induces a marginal distribution G of posterior beliefs:

$$G(B) := \int_0^1 \Gamma(B \mid \omega) \, dF(\omega) \qquad \text{for every Borel set } B \subseteq [0,1].$$

Throughout, I use F and G to denote both the corresponding distributions and their cumulative distribution functions when no confusion can arise.

This is a coarsening problem rather than an unrestricted persuasion problem: the provider can only remap an existing calibrated score, so not every Bayes-plausible posterior distribution is feasible. Let $\text{MPC}(F)$ denote the set of mean-preserving contractions of F, that is, the set of marginal posterior distributions induced by garbling the score distribution F. Feasible

[3] Relabeling the two states and the two actions maps a case with $D_{FP} > D_{FN}$ into one with $D_{FP} \leq D_{FN}$.

posterior distributions are therefore restricted to $G \in \mathrm{MPC}(F)$.[4]

2. **Score realization and review.** Nature draws $\omega \sim F$, which is observed privately by the provider. The provider then sends a recommendation according to the committed policy, inducing a posterior belief $\mu \sim \Gamma(\cdot \mid \omega)$ for the DM. After observing μ, the DM either acts immediately or opens review. If he opens review, he chooses any Bayes-plausible experiment $\tau \in \Delta([0,1])$ with mean μ and pays

$$\mathcal{C}(\tau;\mu) = \mathbf{1}\{I(\tau;\mu) > 0\}\, K + \lambda I(\tau;\mu), \tag{1}$$

where $I(\tau;\mu) := \phi(\mu) - \mathbb{E}_{\nu\sim\tau}[\phi(\nu)]$ is a posterior-separable information cost, $\phi : [0,1] \to \mathbb{R}$ is a strictly concave C^2 potential, and $\lambda > 0$ is the marginal processing-cost parameter. The parameter $K \geq 0$ is the startup cost of activating review; $\lambda I(\tau;\mu)$ captures marginal processing costs once review is open.[5] This nests posterior-separable rational inattention when $K = 0$ and a fixed-cost review benchmark when $\lambda = 0$.

3. **Final action.** After optional review, the DM chooses $a \in \{\text{Accept}, \text{Reject}\}$ based on his final posterior belief, and payoffs are realized.[6]

Reduced-form pricing and outside option. The DM's outside option is not to use the system. In that case, he starts from the prior μ_0, receives no recommendation, and solves the same review problem studied in Section III. Let $V(\mu;\alpha)$ denote the DM's continuation value when he enters that problem from posterior belief μ under liability share α; in particular, $V(\mu_0;\alpha)$ is the outside-option value.

A recommendation policy affects pricing only through the ex ante distribution G of posterior beliefs that it induces. Under the reduced-form pricing assumption, the provider can charge an ex ante transfer equal to a fraction $\beta \in [0,1]$ of the DM's expected continuation-surplus gain relative to the outside option:

$$T(G;\alpha) = \beta\left(\mathbb{E}_G[V(\mu;\alpha)] - V(\mu_0;\alpha)\right). \tag{2}$$

[4] Formally, $G \in \mathrm{MPC}(F)$ if and only if $\int_0^x G(t)\,dt \leq \int_0^x F(t)\,dt$ for all $x \in [0,1]$, and $\mathbb{E}_{\mu\sim G}[\mu] = \mathbb{E}_{\omega\sim F}[\omega] = \mu_0$.

[5] Examples include opening a case file, escalating to a second review, or launching an audit; K should be interpreted as a reduced-form startup friction rather than a literal technological fee.

[6] When indifferent, the DM reviews; at the final action stage, ties are broken in the provider's favor.

Equivalently, once policy G is fixed and before the score is realized, the provider captures a fraction β of the expected value that the system creates for the DM relative to facing the prior directly. The transfer is therefore ex ante and policy-dependent, not a case-by-case payment contingent on the realized posterior.[7]

In the baseline analysis, the score distribution F is exogenous. The next section solves the DM's review problem and derives the two review thresholds that discipline the disclosure problem.

III The Decision Maker's Review Problem

For a given posterior belief, the DM chooses whether to open review and, conditional on review, how much information to acquire. The key separation is simple: the marginal processing cost governs the learning problem once review is open, while the startup cost governs whether review is activated at all.

III.1 Concavification

Given a posterior belief μ induced by the recommendation, the DM either acts immediately or acquires additional information before choosing between Accept and Reject. If he acts immediately, his payoff is $-\alpha\ell(\mu)$, where

$$\ell(\mu) := \min\{\mu D_{FN}, (1-\mu)D_{FP}\}$$

is the expected misclassification loss under the optimal myopic action. The loss function ℓ is concave and piecewise linear, with a peak at the myopic cutoff $\overline{\mu} := \frac{D_{FP}}{D_{FP}+D_{FN}} \leq \frac{1}{2}$. Hence, the immediate-action utility $-\alpha\ell(\mu)$ is convex, with a kink at $\overline{\mu}$.

The DM's continuation value solves

$$\begin{aligned} V(\mu;\alpha) &= \max_{\tau:\, \mathbb{E}_{\nu\sim\tau}[\nu]=\mu} \Big(\mathbb{E}_{\nu\sim\tau}[-\alpha\ell(\nu)] - \mathcal{C}(\tau;\mu)\Big) \\ &= \max_{\tau:\, \mathbb{E}_{\nu\sim\tau}[\nu]=\mu} \Big(\mathbb{E}_{\nu\sim\tau}[-\alpha\ell(\nu)+\lambda\phi(\nu)] - \lambda\phi(\mu) - \mathbf{1}\{I(\tau;\mu)>0\}\,K\Big). \end{aligned}$$

Define the net utility function $N(\mu) := -\alpha\ell(\mu) + \lambda\phi(\mu)$.

[7] This reduced-form transfer rule can be interpreted as a licensing fee, access fee, or negotiated revenue-sharing rule. The parameter β captures bargaining power, platform pricing power, competitive pressure, or a pass-through constraint. What matters for the analysis is that β governs the provider's disclosure incentives and, once training is endogenized in Section VI, her incentive to invest in score informativeness.

Conditional on review, the DM solves a standard information-acquisition problem; the startup cost governs only whether that problem is activated.

Lemma 1 (Value Function Characterization)**.** *Define the* concave envelope *of N as*

$$(\operatorname{cav} N)(\mu) := \sup\left\{\sum_i w_i N(\mu_i) : w_i \geq 0, \sum_i w_i = 1, \sum_i w_i \mu_i = \mu\right\}.$$

The DM's value function satisfies

$$V(\mu;\alpha) + \lambda\phi(\mu) = \max\{N(\mu),\ (\operatorname{cav} N)(\mu) - K\}.$$

Proof. See Appendix A.1. □

III.2 Conditional on review: invariant stopping points

Lemma 2 (Invariant stopping points conditional on review)**.** *Suppose $\lambda > 0$. Then there exist unique beliefs*

$$0 < \mu_L^* < \overline{\mu} < \mu_H^* < 1$$

such that:

1. *If $\mu \notin (\mu_L^*, \mu_H^*)$, the optimal experiment is degenerate.*
2. *If $\mu \in (\mu_L^*, \mu_H^*)$, every optimal experiment is binary and supported on $\{\mu_L^*, \mu_H^*\}$.*

The pair (μ_L^*, μ_H^*) is uniquely characterized by the common-tangent conditions

$$\frac{N(\mu_H^*) - N(\mu_L^*)}{\mu_H^* - \mu_L^*} = -\alpha D_{FN} + \lambda\phi'(\mu_L^*), \tag{3a}$$

$$\frac{N(\mu_H^*) - N(\mu_L^*)}{\mu_H^* - \mu_L^*} = \alpha D_{FP} + \lambda\phi'(\mu_H^*). \tag{3b}$$

Proof. See Appendix A.2. □

Lemma 2 characterizes the intensive margin. Conditional on review, the optimal experiment always splits beliefs to the same stopping points (μ_L^*, μ_H^*). These depend on (α, λ, ϕ) but not on K, so startup frictions affect whether review occurs, not how learning proceeds once review is open.

For any prior $\mu \in (\mu_L^*, \mu_H^*)$, Bayes plausibility pins down the probability of reaching the upper

stopping point, $\pi(\mu) := \frac{\mu - \mu_L^*}{\mu_H^* - \mu_L^*}$.

III.3 Activation decision and the three-region policy

Because the stopping points are already pinned down by the conditional-on-review problem, the fixed cost K enters only through the activation decision. Define the net benefit of review by $\Delta(\mu) := (\operatorname{cav} N)(\mu) - N(\mu)$, which measures the gain from information acquisition before paying the startup cost. The DM reviews if and only if $\Delta(\mu) \geq K$.

Proposition 1 (Three-region review rule)**.** *Suppose $\lambda > 0$.*

1. ***Inactive regime.*** *If $K > \Delta(\overline{\mu})$, the DM never reviews and acts myopically for all $\mu \in [0, 1]$.*
2. ***Active regime.*** *If $0 \leq K \leq \Delta(\overline{\mu})$, there exist unique thresholds μ^L and μ^H satisfying*

$$\mu_L^* \leq \mu^L \leq \overline{\mu} \leq \mu^H \leq \mu_H^*,$$

such that the optimal policy is:

- ***Immediate reject:*** *for $\mu \in [0, \mu^L)$, the DM rejects without review;*
- ***Review and learn:*** *for $\mu \in [\mu^L, \mu^H]$, the DM pays K and runs the optimal binary experiment supported on $\{\mu_L^*, \mu_H^*\}$;*
- ***Immediate accept:*** *for $\mu \in (\mu^H, 1]$, the DM accepts without review.*

The thresholds are the unique solutions to

$$\Delta(\mu^L) = \Delta(\mu^H) = K, \qquad \mu^L \leq \mu^H.$$

In the boundary cases, if $K = 0$ then $(\mu^L, \mu^H) = (\mu_L^*, \mu_H^*)$, and if $K = \Delta(\overline{\mu})$ then $\mu^L = \mu^H = \overline{\mu}$.

Proof. See Appendix A.3. □

Proposition 1 delivers the review policy used throughout the paper. With a positive startup cost, review is selective and bounded by two thresholds, μ^L and μ^H: below μ^L the DM rejects immediately, above μ^H he accepts immediately, and inside $[\mu^L, \mu^H]$ he pays K and runs the binary experiment from Lemma 2. I refer to μ^L and μ^H as the review gates, and to (μ_L^*, μ_H^*) as the stopping points reached conditional on review. Figure 1 summarizes these objects.[8]

[8] This cutoff structure echoes a familiar feature of costly auditing and verification models: when review is costly, optimal policies are selective rather than universal (Townsend, 1979; Mookherjee and Png, 1989).

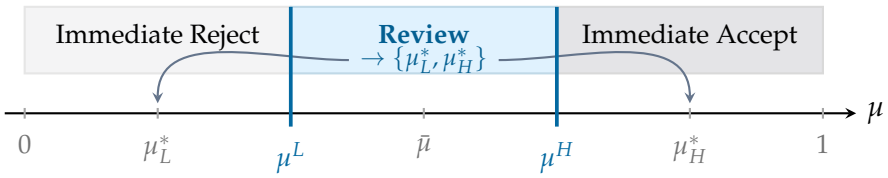


Figure 1: Review gates and stopping points. Outside the interval $[\mu^L, \mu^H]$, the decision maker acts immediately: reject for $\mu < \mu^L$ and accept for $\mu > \mu^H$. Inside that interval, the decision maker opens review; conditional on review, the optimal binary experiment moves beliefs to the stopping points μ_L^* and μ_H^*. The myopic cutoff $\bar{\mu}$ lies between the two review gates.

$$V(\mu;\alpha) = \begin{cases} -\mu\alpha D_{FN}, & \mu \in [0, \mu^L), \\ (\operatorname{cav} N)(\mu) - \lambda\phi(\mu) - K, & \mu \in [\mu^L, \mu^H], \\ -(1-\mu)\alpha D_{FP}, & \mu \in (\mu^H, 1]. \end{cases} \tag{4}$$

On the review region, the concavification is the chord through $(\mu_L^*, N(\mu_L^*))$ and $(\mu_H^*, N(\mu_H^*))$. Hence, for $\mu \in [\mu^L, \mu^H]$,

$$V(\mu;\alpha) = -\alpha[(1-\pi(\mu))\mu_L^* D_{FN} + \pi(\mu)(1-\mu_H^*)D_{FP}] - K - \lambda I^*(\mu),$$

where $I^*(\mu) := \phi(\mu) - \left[(1-\pi(\mu))\phi(\mu_L^*) + \pi(\mu)\phi(\mu_H^*)\right]$ is the information cost of the optimal binary experiment.

These review gates are the key state variables for the provider's disclosure problem. Once review switches on and off discretely at μ^L and μ^H, moving a posterior to either gate changes the DM's continuation behavior discretely. That discrete change, rather than any generic preference for opacity, is what makes optimal disclosure gate-targeting.

IV Advice as Escalation Design

IV.1 Provider Payoff Geometry

When $K = 0$, disclosure affects only how much the DM learns conditional on review, so posterior dispersion is valuable and full revelation is optimal. When $K > 0$, disclosure also affects whether review occurs. Advice therefore becomes part of the escalation rule governing whether review is opened.

The provider's payoff combines two components: her share of misclassification losses and

the ex ante transfer defined in (2). Thus, her objective combines her expected share of realized misclassification losses with the fraction β of the DM's ex ante continuation-surplus gain created by the recommendation policy relative to the outside option. To study disclosure, it is useful to substitute the DM's continuation value from Section III directly into this objective and express payoffs as a function of the posterior belief μ induced by a recommendation.

This substitution yields a reduced-form provider payoff with three regions. Outside the review region, the DM acts immediately, so the provider's payoff is just her effective misclassification loss. Inside the review region, the provider values a posterior through the continuation surplus generated by review, net of the review costs she internalizes through the transfer rule. Specifically, the provider directly bears a $(1-\alpha)$-share of realized misclassification losses, and through the transfer she internalizes an additional $\beta\alpha$-share of misclassification losses and a β-share of review costs. It is therefore convenient to define the effective loss weight $\gamma := (1-\alpha)+\beta\alpha$, together with the effective error costs $C_{FN} := \gamma D_{FN}$ and $C_{FP} := \gamma D_{FP}$. Normalizing by the constant term $\beta V(\mu_0;\alpha)$, the provider's reduced-form payoff at posterior μ can be written as

$$\widetilde{u}(\mu) = \begin{cases} -\mu C_{FN}, & \mu < \mu^L, \\ -\big((1-\pi(\mu))\,\mu_L^*\,C_{FN} + \pi(\mu)\,(1-\mu_H^*)\,C_{FP}\big) - \beta K - \beta\lambda\, I^*(\mu), & \mu \in [\mu^L, \mu^H], \\ -(1-\mu)C_{FP}, & \mu > \mu^H, \end{cases} \tag{5}$$

where $\pi(\mu) := (\mu - \mu_L^*)/(\mu_H^* - \mu_L^*)$ is the mixing weight on the high posterior μ_H^* in the optimal binary review experiment, and $I^*(\mu)$ is the associated information cost, as defined after Lemma 2 and Proposition 1 in Section III.

The provider therefore chooses a recommendation policy Γ to maximize expected reduced-form payoff. Because her objective depends on recommendations only through the induced marginal distribution of posterior beliefs, the problem can equivalently be written as the choice of $G \in \mathrm{MPC}(F)$ to maximize $\mathbb{E}_G[\widetilde{u}(\mu)]$ (Gentzkow and Kamenica, 2016; Kolotilin, 2018; Arieli et al., 2023).

Equation (5) can be read as follows. For $\mu < \mu^L$ and $\mu > \mu^H$, the DM acts immediately, so the provider faces affine myopic-loss tradeoffs. For $\mu \in [\mu^L, \mu^H]$, the DM opens review, so the provider's payoff reflects the continuation value of the optimal review experiment through μ_L^*, μ_H^*, $\pi(\mu)$, and $I^*(\mu)$. When $K > 0$, moving a posterior to either review gate changes whether

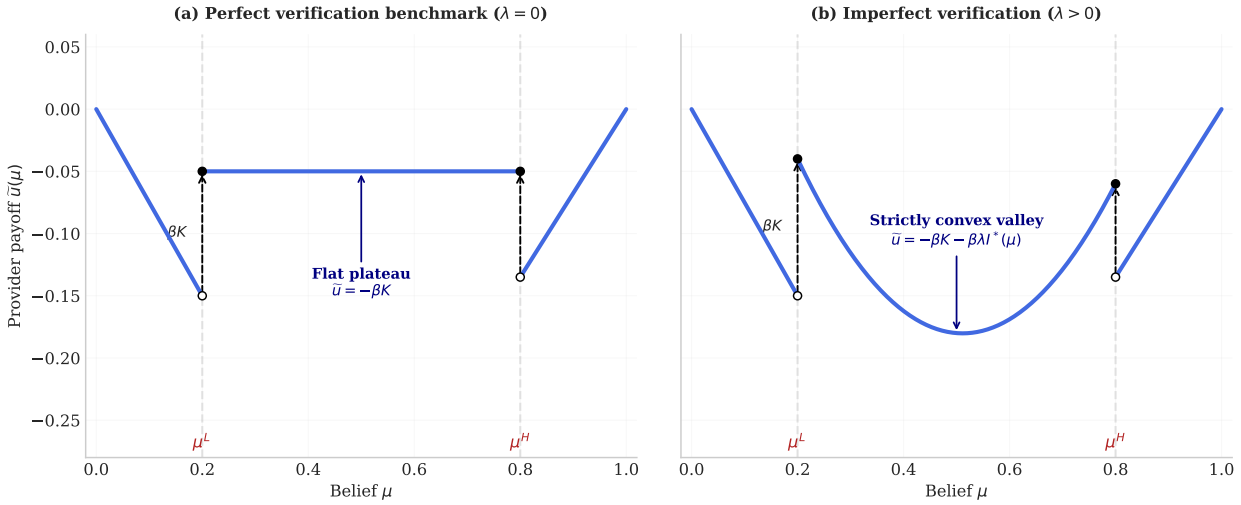


Figure 2: Provider payoff geometry. Outside the review region, the provider's payoff is affine because the DM acts myopically. Inside the review region, the payoff is flat when $\lambda = 0$ and strictly convex when $\lambda > 0$. With $K > 0$, crossing a review gate changes whether review occurs and creates upward payoff jumps at μ^L and μ^H. Those jumps make pooling at the gates more valuable than using finer recommendation categories away from them.

review is activated and therefore creates a discrete payoff gain at μ^L and μ^H. Those discrete gains are what make gate-targeting optimal.

IV.2 Benchmark without a startup friction: full revelation

The role of K is easiest to see from the benchmark in which review has no startup friction. When $K = 0$, the review region is $[\mu_L^*, \mu_H^*]$, but the payoff jumps at its boundaries disappear. As a result, the provider no longer gains from pushing posterior beliefs just across a review threshold.

What remains is a pure dispersion motive. When $K = 0$, $\widetilde{u}$ is affine on the two wings and weakly convex on the review region (strictly convex when $\lambda > 0$). Every feasible design $G \in \mathrm{MPC}(F)$ is a mean-preserving contraction of the score distribution F, so any coarsening weakly compresses posterior dispersion. Under a convex payoff, that compression weakly lowers the provider's expected payoff. The provider therefore prefers the most dispersed feasible design, namely full revelation.

Proposition 2 (No-startup benchmark: full revelation)**.** *Fix $\lambda > 0$ and $K = 0$. The optimal implementable posterior distribution is $G^\star = F$, i.e., full revelation.*[9]

Proof. When $K = 0$, the provider payoff $\widetilde{u}$ is convex on $[0, 1]$: it is affine on the two wings and convex on the review region, with slopes weakly increasing at the boundaries. Any feasible posterior

[9] Pooling within $[0, \mu_L^*]$ or $[\mu_H^*, 1]$ is payoff-equivalent to full revelation.

distribution $G \in \text{MPC}(F)$ is a mean-preserving contraction of F with the same mean. Therefore, for the convex function $\widetilde{u}$,

$$\mathbb{E}_G[\widetilde{u}(\mu)] \leq \mathbb{E}_F[\widetilde{u}(\omega)].$$

Hence the provider's expected payoff is maximized by the most dispersed feasible posterior distribution, namely full revelation $G^\star = F$. □

Proposition 2 isolates the role of startup costs. Costly information acquisition alone does not generate coarse recommendations here. Coarseness appears only when disclosure can move cases across the review gates.

IV.3 Startup Costs and Gate-Targeting Disclosure

Consider now $K > 0$ and $\lambda > 0$. With a startup cost, the natural target posteriors are the review gates μ^L and μ^H. Moving a case just across a gate switches review on or off and yields a discrete payoff gain; moving beliefs within the review region changes payoffs only smoothly. This discreteness is what makes gate-targeting disclosure optimal.

Theorem 1 (Gate-targeting disclosure under startup frictions)**.** *Suppose $K > 0$ and $\lambda > 0$. Any optimal implementable posterior distribution takes one of three forms:* single-gate pooling, two-sided gate pooling, *or* residual interior disclosure*. In every optimal design, no strategically pooled posterior lies in (μ^L, μ^H): every strategically pooled posterior is either μ^L or μ^H. In the residual case, Bayes plausibility prevents complete gate absorption and the remaining interior interval is fully revealing.*

Proof. Appendix B.2 provides the geometric characterization and cutoff constructions underlying Theorem 1. □

Theorem 1 gives the main disclosure result. Startup frictions do not create a generic preference for opacity. They make the review gates the valuable posterior targets, so optimal disclosure pools strategically at those gates rather than preserving fine distinctions everywhere.

Appendix B.2 makes the remaining distinction precise. In the central-region case, both review gates remain the relevant posterior targets for the mass between them. If that central mass can be absorbed fully at the two gates while preserving Bayes plausibility, the design takes a two-sided gate-pooling form. If full absorption is infeasible, some interior structure must remain.

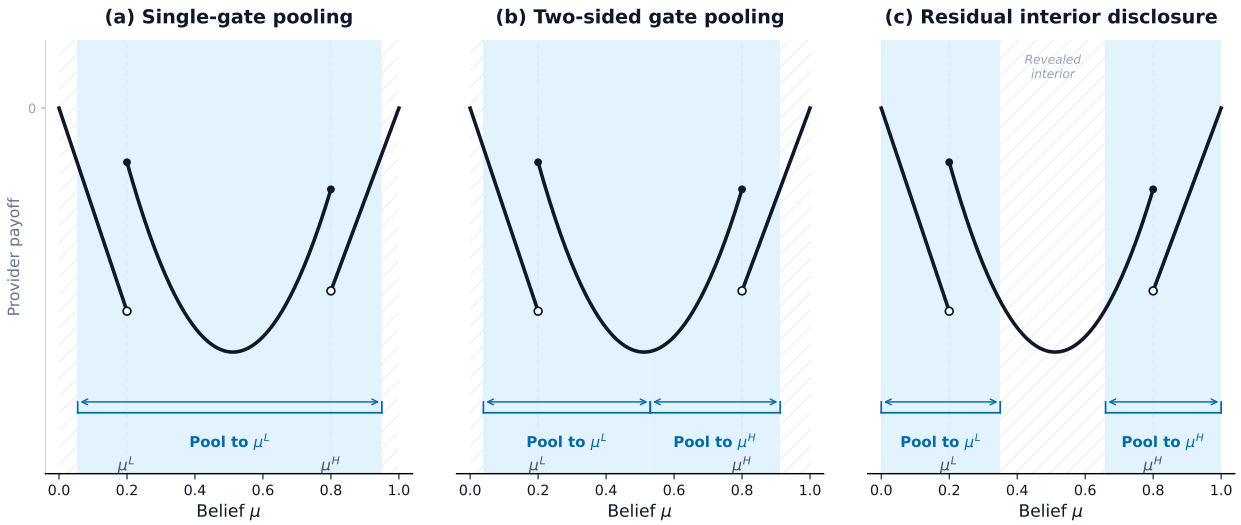


Figure 3: Three configurations of gate-targeting disclosure. Panel (a) shows single-gate pooling. Panel (b) shows two-sided gate pooling. Panel (c) shows residual interior disclosure, in which complete gate absorption is infeasible and a revealing interior interval remains. Appendix B.2 reports the finer geometric refinement within these three configurations.

Residual interior disclosure is the case in which the two review gates remain the relevant posterior targets for the central block, but the central block cannot be reassigned fully to those two gate posteriors while preserving pooled means. Economically, this happens when the central block is not sufficiently tail-rich relative to the two gate targets: the left and right tails of the block can support gate-targeting pooling near the boundaries, but not enough to absorb the entire middle. A positive-length interior interval therefore remains fully revealing. The revealing middle interval should thus be read as a feasibility residue, not as a separate motive for fine interior communication.

Figure 3 visualizes these three configurations. Corollary 1 then adds an implementation refinement: outside the residual case, the affine revealed pieces can be compressed without loss, so an optimal design can be implemented with finitely many recommendation categories.

Corollary 1 (Finite implementation outside the residual case)**.** *In the single-gate and two-sided gate-pooling cases, an optimal implementation uses finitely many recommendation categories, and at most four suffice.*

Proof. Appendix B.3 provides the implementation refinement underlying Corollary 1. □

Finite recommendation categories are therefore a consequence of gate-targeting, not a separate economic force.

V Liability Sharing, Review, and Welfare

V.1 Welfare Logic

The welfare analysis separates two questions: whether review is induced at all, and, conditional on inducement, whether gate-targeting sends the right cases to review.

Even holding score informativeness fixed, endogenous coarsening can distort welfare. The algorithm-only benchmark isolates the distortion from gate-targeting disclosure. Comparing the deployed system with the better standalone alternative adds a second failure mode: at low liability, review may not be induced, so the deployed system collapses to algorithm-only use.

For any implementable recommendation policy Γ, let total welfare be real surplus net of review costs:

$$W(\Gamma;\alpha) \;:=\; -\mathbb{E}_{G_\Gamma}[\ell(\nu)] - \mathbb{E}_{G_\Gamma}[\mathcal{C}(\tau;\mu)]\,, \tag{6}$$

where G_Γ is the induced marginal distribution of posterior beliefs, ν denotes the post-review posterior (with $\nu = \mu$ when the DM does not review), and expectations are taken under the joint distribution induced by Γ and the DM's equilibrium continuation strategy. When welfare depends only on the induced marginal posterior distribution, I write $W(G;\alpha)$ as shorthand.

Using the DM's continuation value $V(\mu;\alpha)$ from (4), welfare under policy Γ can be written as

$$W(\Gamma;\alpha) \;=\; \mathbb{E}_{G_\Gamma}\left[V(\mu;\alpha)\right] - (1-\alpha)\mathbb{E}_{G_\Gamma}\left[\ell(\nu)\right]. \tag{7}$$

The first term captures the DM's continuation value, which already nets out review costs and the DM's α-share of misclassification losses. The second term is the residual error cost that the DM does not internalize.

Equation (7) shows that liability matters in two ways. As α rises, the provider internalizes more of the welfare consequences of advice design, so the residual error externality shrinks and disappears at $\alpha = 1$. At the same time, review becomes privately more attractive to the DM, so disclosure has greater scope to move cases into review. Liability therefore governs both whether review is induced and how much of the resulting distortion the provider internalizes.

V.2 A low-liability benchmark: no review and full revelation

Let $\Gamma^\star(\alpha, K)$ denote a selected equilibrium recommendation policy, and let

$$G^\star(\alpha, K) := G_{\Gamma^\star(\alpha,K)}$$

denote the induced marginal posterior distribution. Define equilibrium welfare by

$$W(\alpha, K) := W(\Gamma^\star(\alpha, K); \alpha).$$

Define

$$\Delta^{\max}(\alpha) := \max_{\mu\in[0,1]} \Delta(\mu; \alpha, \lambda), \qquad \alpha_{\min}(K) := \inf\{\alpha \in [0,1] : \Delta^{\max}(\alpha) \geq K\},$$

with the convention that $\alpha_{\min}(K) = +\infty$ if the set is empty. If $\alpha_{\min}(K) = +\infty$, review is never induced for any $\alpha \in [0,1]$. Otherwise, $\alpha_{\min}(K)$ is the smallest liability level at which some posterior makes review privately worthwhile.

Define the algorithm-only benchmark by $W^{\text{algorithm}} := -\mathbb{E}_F[\ell(\omega)]$, that is, expected welfare when the score is fully revealed and review is never used.

Proposition 3 (Low-liability benchmark: no review)**.** *If $\alpha < \alpha_{\min}(K)$, then review is never induced in equilibrium. The provider's optimal recommendation policy is full revelation, so the induced posterior distribution satisfies $G^\star = F$, and equilibrium welfare equals the algorithm-only benchmark, $W(\alpha, K) = W^{\text{algorithm}}$.*

Proof. See Appendix C.1. □

Proposition 3 gives the low-liability endpoint. When liability is too low to make review privately worthwhile anywhere, equilibrium disclosure is fully revealing and the deployed system collapses to the algorithm-only benchmark. The relevant comparison in that region is therefore whether algorithm-only deployment already dominates human-only review.

V.3 Full decision-maker liability as a deployment-stage alignment benchmark

Conditional on fixed score informativeness, assigning full liability to the decision maker removes the deployment-stage disclosure distortion. When $\alpha = 1$, the residual welfare wedge in (7) disappears. If $\beta > 0$, the provider then maximizes a positive multiple of the DM's continuation value and therefore a positive multiple of total welfare.

Proposition 4 (Full-liability benchmark: deployment efficiency)**.** *Suppose the transfer rule satisfies* (2) *with* $\beta > 0$.[10] *For fixed score informativeness, equilibrium welfare is maximized at full decision-maker liability* $\alpha = 1$*; that is,* $W(1, K) \geq W(\alpha, K)$ *for all* $\alpha \in [0, 1]$.

Proof. See Appendix C.3. □

Proposition 4 is a deployment-stage benchmark, not a literal institutional prescription. Under the reduced-form transfer rule and for fixed score informativeness, full decision-maker liability aligns the provider's objective with continuation welfare. The remaining question is whether welfare rises monotonically between the low-liability endpoint and this benchmark. It need not.

V.4 Mis-sorting across the immediate-action cutoff

The welfare logic is easiest to see through mis-sorting. Review itself is not the problem: conditional on a given posterior, opening review is weakly socially beneficial. The hazard arises when gate-targeting pools cases from opposite sides of the immediate-action cutoff and changes which cases are reviewed.

For each posterior μ, let $B(\mu; \alpha)$ denote the social gain from allowing review to start from μ, and let $D_\Gamma(\mu)$ denote the distortion created by the set of scores pooled into that posterior under policy Γ.

For any implementable recommendation policy $\Gamma(\cdot \mid \omega)$, let G_Γ denote the induced marginal posterior distribution. Each posterior μ corresponds to a set of underlying scores that the recommendation maps into μ; let $\bar{\ell}_\Gamma(\mu)$ denote the average myopic loss among those scores. The *social continuation loss* at posterior μ captures what actually happens to welfare once the DM decides whether to review:

$$L(\mu; \alpha) := \mathbb{E}_{\nu \sim \tau^*(\mu;\alpha)}[\ell(\nu)] + \mathcal{C}(\tau^*(\mu; \alpha); \mu),$$

[10] If $\beta = 0$, the deployment-stage benchmark remains attainable at $\alpha = 1$, but the provider is indifferent across recommendation policies and the equilibrium selection need not be unique.

where $\tau^*(\mu;\alpha)$ is the DM's equilibrium review strategy; when no review occurs, $L(\mu;\alpha) = \ell(\mu)$. By iterated expectations, $\mathbb{E}_{G_\Gamma}[\bar{\ell}_\Gamma(\mu)] = \mathbb{E}_F[\ell(\omega)]$.

Lemma 3 (Two-margin decomposition)**.** *For any implementable recommendation policy Γ,*

$$W(\Gamma;\alpha) - W^{\text{algorithm}} \;=\; \mathbb{E}_{G_\Gamma}[B(\mu;\alpha)] - \mathbb{E}_{G_\Gamma}[D_\Gamma(\mu)]\,, \tag{8}$$

where

$$B(\mu;\alpha) := \ell(\mu) - L(\mu;\alpha) \qquad \textit{(social value of review at posterior } \mu\textit{)},$$
$$D_\Gamma(\mu) := \ell(\mu) - \bar{\ell}_\Gamma(\mu) \qquad \textit{(mis-sorting distortion at posterior } \mu\textit{)}.$$

Moreover:

(i) $B(\mu;\alpha) \geq 0$ *for every* μ *and every* $\alpha \in [0,1]$*;*

(ii) $D_\Gamma(\mu) \geq 0$ *for every* μ*, with equality if and only if* ℓ *is affine on the support of* $Q_\Gamma(\cdot \mid \mu)$*.*

Proof. See Appendix C.4. □

Part (i) is the key point: conditional on a posterior, human review is weakly socially beneficial. Welfare losses therefore come from how advice groups underlying cases before the review decision is made. Corollary 2 sharpens this observation: any welfare loss relative to the algorithm-only benchmark must come from advice categories that pool scores from opposite sides of the immediate-action cutoff.

The structure of the loss function ℓ pins down exactly which recommendation policies generate positive mis-sorting distortion.

Corollary 2 (Cross-cutoff mis-sorting is necessary for a welfare loss)**.** *Recall that* $\ell(\mu) = \min\{\mu D_{FN}, (1-\mu)D_{FP}\}$ *is affine on* $[0,\overline{\mu}]$ *and on* $[\overline{\mu},1]$*, where* $\overline{\mu} = D_{FP}/(D_{FP}+D_{FN})$*.*

(i) If every recommendation category induced by Γ *pools scores from only one side of* $\overline{\mu}$*, then* $D_\Gamma(\mu) = 0$ *for almost every realized posterior, and* $W(\Gamma;\alpha) \geq W^{\text{algorithm}}$ *for every* α*.*

(ii) If some recommendation categories with positive probability pool scores from both sides of $\overline{\mu}$*, then* $D_\Gamma(\mu) > 0$ *for those categories.*

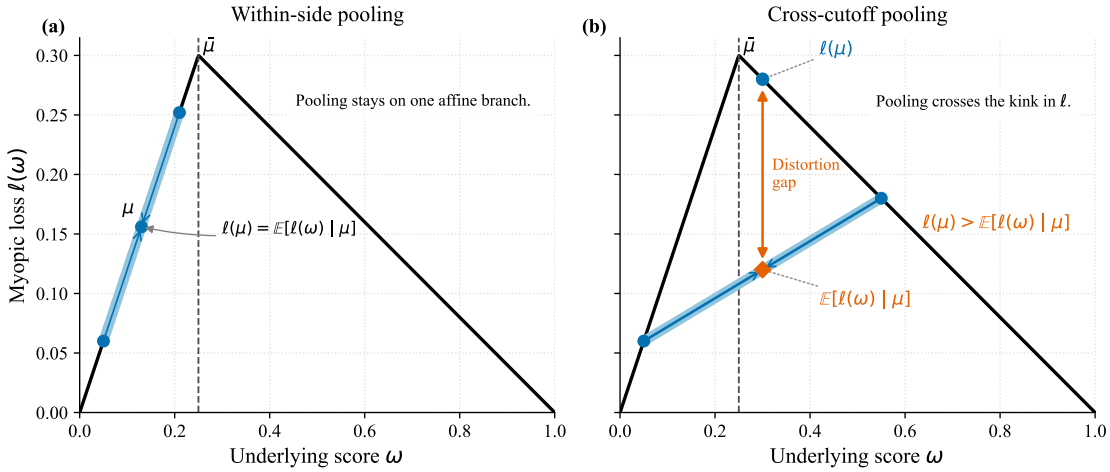


Figure 4: Why cross-cutoff pooling is harmful. Panel (a): if pooled scores lie on the same side of the immediate-action cutoff $\overline{\mu}$, the loss function is affine on that fiber, so the loss at the pooled posterior equals the average loss of the underlying scores. Panel (b): if pooled scores straddle $\overline{\mu}$, the fiber crosses the kink in ℓ, so Jensen's inequality is strict: the loss at the pooled posterior exceeds the average loss of the underlying scores. The vertical gap is the mis-sorting distortion. Any welfare loss relative to the algorithm-only benchmark must therefore come from mis-sorting induced by endogenous coarsening, not from review itself.

Consequently, any review trap requires a recommendation policy that pools scores from opposite sides of $\overline{\mu}$ into the same recommendation category.

Proof. See Appendix C.5. □

Corollary 2 shows that the welfare problem is a sorting problem, not a review problem. Pooling within one side of the immediate-action cutoff is harmless because ℓ is affine there. Distortion arises only when an advice category straddles the cutoff. This logic motivates the stronger comparison in the next subsection, which evaluates the combined system against the better standalone alternative.

V.5 Review traps relative to the algorithm-only benchmark

A *review trap* occurs when the mis-sorting distortion created by endogenous coarsening outweighs the social value of the review it induces, so the combined system performs worse than the algorithm-only benchmark, $W(\alpha, K) < W^{\text{algorithm}}$. This can already happen locally when review first becomes inducible.

Proposition 5 (Local review trap). *Fix K with $\alpha_{\min}(K) < 1$, and suppose there exists $\bar{\alpha} > \alpha_{\min}(K)$ and a selected equilibrium branch $\Gamma^\star(\alpha, K)$ on $(\alpha_{\min}(K), \bar{\alpha})$ such that: (i) the branch admits a weak-$*$ right limit at $\alpha_{\min}(K)$, and (ii) equilibrium welfare is right-continuous at $\alpha_{\min}(K)$ along that branch. If the post-*

activation limit design induces cross-cutoff pooling whose mis-sorting distortion exceeds the social value of the review it triggers, then there exists $\varepsilon > 0$ such that

$$W(\alpha, K) < W^{\text{algorithm}} \quad \textit{for all } \alpha \in (\alpha_{\min}(K), \alpha_{\min}(K) + \varepsilon).$$

Proof. See Appendix C.6. □

These are local selection conditions at the activation boundary, not new economic restrictions: they rule out discontinuous switching across equilibrium branches exactly when review first becomes inducible.

Combining Proposition 5 with the full-liability benchmark from Proposition 4 yields an immediate implication. If $\beta > 0$ and $\Delta^{\max}(1) > K$, then welfare is locally below the algorithm-only benchmark just after activation but strictly above it at full decision-maker liability. Welfare in α is therefore non-monotone, and the review-trap set is nonempty.

The uniform $\lambda = 0$ benchmark provides a closed-form illustration of the same mechanism. In that benchmark, a closed-form cutoff separates startup-cost values for which welfare rises at activation from those for which it falls, while welfare recovers at full decision-maker liability over the relevant parameter range. Figure 5 illustrates the resulting non-monotonicity. Appendix C.7 derives the benchmark calculation.

V.6 Two sources of negative deployment synergy

The algorithm-only benchmark isolates the distortion cleanly, but system evaluation requires a stronger benchmark: a deployed human–algorithm system should be compared with the better of algorithm-only deployment and human-only review. I use *deployment synergy* to denote the welfare gain of the combined human–algorithm system relative to the better standalone alternative.[11]

Define the human-only benchmark by

$$W^{\text{human}}(K) := V(\mu_0; 1),$$

that is, welfare when the decision maker receives no algorithmic input and decides whether to re-

[11] I use the term in the sense of the human–AI literature: the joint effect is evaluated relative to the relevant standalone alternatives and can be positive, negative, or neutral. See, for example, Vaccaro et al. (2024).

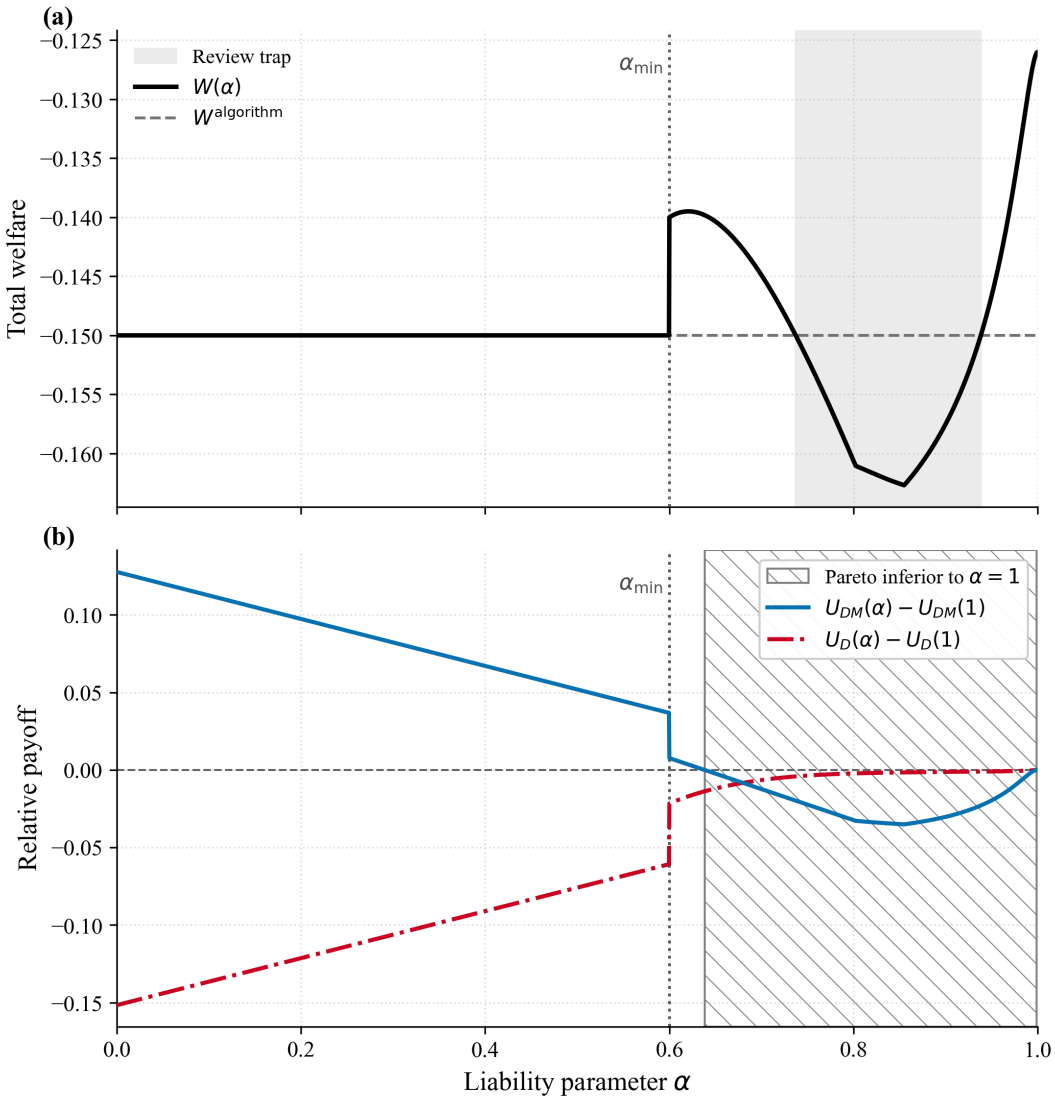


Figure 5: Benchmark illustration of a review trap (uniform prior, $\lambda = 0$, $K = 0.18$, $\beta = 0.03$). Panel (a) plots total welfare. Welfare coincides with the algorithm-only benchmark before review becomes inducible, can fall below it just after activation because gate-targeting causes cross-cutoff pooling, and later recovers as alignment strengthens. Panel (b) plots the DM's and provider's payoffs relative to their $\alpha = 1$ levels; the shaded region marks liability levels at which both are below their full-liability payoffs.

view starting from the prior alone. The argument uses $\alpha = 1$ because, in the absence of a provider, welfare equals the decision maker's continuation value with full internalization of misclassification losses. Define deployment synergy by

$$\text{Synergy}^{\text{deploy}}(\alpha, K) := W(\alpha, K) - \max\left\{W^{\text{algorithm}}, W^{\text{human}}(K)\right\}.$$

Negative deployment synergy means that the combined human–algorithm system performs worse than the better standalone alternative.

Negative deployment synergy can arise for two distinct reasons. The first is insufficient review. By Proposition 3, whenever $\alpha < \alpha_{\min}(K)$, the deployed system collapses to algorithm-only use. Therefore, if human-only review dominates algorithm-only deployment at $K = 0$, continuity of $W^{\text{human}}(K)$ implies the same ranking for sufficiently small startup costs. The second is distorted review. Once review becomes inducible, endogenous coarsening can route the wrong cases into review through cross-cutoff pooling, so the deployed system can still underperform the better standalone alternative when algorithm-only deployment is the relevant standalone benchmark.

Corollary 3 (Negative deployment synergy from insufficient review)**.** *Suppose*

$$W^{\text{human}}(0) > W^{\text{algorithm}}.$$

Then there exists $\bar{K} > 0$ such that, for every $K \in (0, \bar{K})$ and every $\alpha < \alpha_{\min}(K)$,

$$W(\alpha, K) = W^{\text{algorithm}} < W^{\text{human}}(K),$$

and hence

$$\text{Synergy}^{\text{deploy}}(\alpha, K) < 0.$$

Proof. See Appendix C.2. □

Corollary 3 identifies the low-liability mechanism behind the blue region in Figure 6: below activation, the deployed system collapses to algorithm-only use, so when startup frictions are small enough that human-only review is better, the combined system underperforms the better standalone alternative.

When algorithm-only deployment is the better standalone benchmark, Proposition 5 immediately yields the orange region in Figure 6. In that region,

$$\max\left\{W^{\text{algorithm}}, W^{\text{human}}(K)\right\} = W^{\text{algorithm}},$$

so a local review trap is automatically also a negative-deployment-synergy result: once review becomes inducible, endogenous coarsening distorts which cases are sent to review and lowers welfare below the relevant standalone alternative.

The uniform $\lambda = 0$ benchmark provides a closed-form illustration of both mechanisms. Figure 6 maps the sign of $\text{Synergy}^{\text{deploy}}(\alpha, K)$ in (α, K) space. The blue region corresponds to insufficient review relative to the human-only benchmark; the orange region corresponds to distorted review relative to the algorithm-only benchmark.

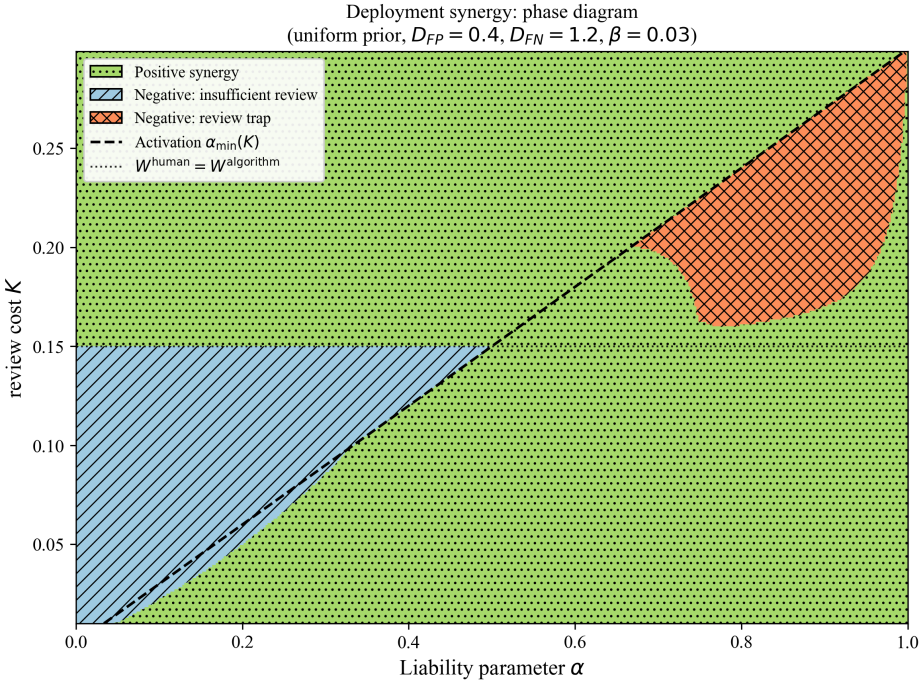


Figure 6: Deployment synergy relative to the better standalone alternative (uniform prior, $\lambda = 0$, $D_{FP} = 0.4$, $D_{FN} = 1.2$, $\beta = 0.03$). The figure compares the deployed human–algorithm system with the better of algorithm-only deployment and human-only review. The blue region corresponds to negative deployment synergy from insufficient review: liability is too low to induce review, so the deployed system collapses to algorithm-only use even though human-only review performs better. The orange region corresponds to distorted review when algorithm-only deployment is the relevant standalone benchmark: once review becomes inducible, endogenous coarsening distorts which cases are sent to review and lowers welfare below the better standalone alternative. The dashed curve is the activation boundary $\alpha_{\min}(K)$; the dotted line marks parameter values at which the human-only and algorithm-only benchmarks coincide.

V.7 Implications for review design

Three implications follow for workflow design. First, advice design is also escalation design: the relevant question is not only how informative advice is, but how it determines which cases receive review. Second, lowering the startup cost of review can improve performance even if the review technology itself is unchanged, because it weakens the incentive to coarsen advice around the review gates. Third, evaluation should look beyond calibration and final accuracy to whether recommendation categories pool cases from opposite sides of the immediate-action cutoff. In settings such as radiology worklists, fraud queues, or compliance escalation, such cross-cutoff pooling is the direct signature of endogenous coarsening.

VI Liability and Upstream Model Improvement

The baseline analysis treats score informativeness as fixed. This section adds an upstream margin by allowing the provider to invest in score quality before deployment. The question is whether the liability shift that improves deployment-stage advice design also weakens incentives to improve the score. Conditional on any fixed score technology, all earlier results continue to hold; the only new issue is how liability trades off better deployment-stage disclosure against weaker incentives to improve the score.

The provider chooses training effort $e \in [\underline{e}, \bar{e}]$ at cost $c(e)$, which determines a family of signal structures $\{H_e(\cdot \mid \theta)\}_{e\in[\underline{e},\bar{e}]}$. Higher effort makes the signal structure more informative in the Blackwell order. Each H_e induces a calibrated posterior score with distribution F_e. Conditional on effort e, the deployment game is exactly the baseline model with F replaced by F_e. Let $G^*(e,\alpha)$ denote the resulting equilibrium disclosure design, and define $M(e,\alpha) := \mathbb{E}_{G^*(e,\alpha)}[V(\mu;\alpha)]$ and $L(e,\alpha) := \mathbb{E}_{G^*(e,\alpha)}[\ell(\nu)]$, where ν denotes the post-review posterior ($\nu = \mu$ when review is not opened). Deployment-stage welfare and provider payoff are

$$W^{dep}(e,\alpha) = M(e,\alpha) - (1-\alpha)L(e,\alpha), \tag{9}$$

$$U^{dep}(e,\alpha) = \beta\big(M(e,\alpha) - V(\mu_0;\alpha)\big) - (1-\alpha)L(e,\alpha). \tag{10}$$

Let $e^*(\alpha)$ denote the provider's effort choice and $W^{tr}(\alpha) := W^{dep}(e^*(\alpha),\alpha) - c(e^*(\alpha))$ denote to-

tal welfare under the induced training level. The formal regularity conditions are collected in Appendix D.1.

Near full decision-maker liability, the welfare effect of α separates into a deployment margin and a training margin. The deployment margin is the same one isolated in Proposition 4: stronger decision-maker liability reduces the provider's disclosure distortion. The new margin is training: when the provider does not capture the full social return to a more informative score, stronger decision-maker liability can reduce her private return to investing in score informativeness. The key derivative gap is

$$W_e^{dep}(e,\alpha) - U_e^{dep}(e,\alpha) = (1-\beta)M_e(e,\alpha), \tag{11}$$

so the provider internalizes the full marginal social return to training if and only if $\beta = 1$.

Proposition 6 (Endogenous-training qualification near full liability)**.** *Let $e_1 = e^*(1)$. Suppose the local regularity conditions in Appendix D.1 hold and assume*

$$M_e(e_1,1) > 0, \qquad \beta M_{ee}(e_1,1) - c''(e_1) < 0, \qquad \beta M_{e\alpha}(e_1,1^-) + L_e(e_1,1^-) < 0.$$

Then

$$e^{*\prime}(1^-) = \frac{-(\beta M_{e\alpha}(e_1,1^-) + L_e(e_1,1^-))}{\beta M_{ee}(e_1,1) - c''(e_1)} < 0, \tag{12}$$

and

$$(W^{tr})'(1^-) = \underbrace{(1-\beta)M_e(e_1,1)e^{*\prime}(1^-)}_{\text{training effect } (<0)} + \underbrace{W_\alpha^{dep}(e_1,1^-)}_{\text{deployment effect } (\geq 0)}. \tag{13}$$

Hence, near $\alpha = 1$, stronger decision-maker liability has two opposing effects:

1. *a* deployment-stage alignment effect, *captured by $W_\alpha^{dep}(e_1,1^-) \geq 0$, which improves recommendation design as in Proposition 4; and*
2. *an* upstream training effect, *captured by $(1-\beta)M_e(e_1,1)e^{*\prime}(1^-) < 0$, which weakens the provider's incentive to invest in score informativeness.*

If $(1-\beta)M_e(e_1,1)\left|e^{*\prime}(1^-)\right| > W_\alpha^{dep}(e_1,1^-)$, then there exists $\bar{\alpha} < 1$, sufficiently close to 1, such that $W^{tr}(\bar{\alpha}) > W^{tr}(1)$.

Proof. See Appendix D.1. □

Proposition 6 qualifies the full-liability benchmark from Proposition 4. Conditional on a fixed score technology, stronger decision-maker liability still improves deployment-stage advice design. What can overturn that deployment-stage benchmark is the upstream training margin: when the provider captures only part of the return to a better score, shifting liability toward the decision maker can improve deployment-stage disclosure while weakening incentives to invest in score informativeness. When $\beta = 1$, that training wedge disappears, so the full-liability benchmark extends globally; Appendix D.1 records the formal corollary.

VII Conclusion

This paper studies advice design when algorithmic output helps determine whether a human decision maker opens costly review. With a startup cost, advice becomes part of the escalation rule: the provider gains from moving posterior beliefs to the thresholds at which review starts or stops rather than from preserving fine distinctions across the score distribution. Coarse advice therefore arises endogenously even without any exogenous communication constraint.

The welfare problem is not review itself but mis-sorting. Conditional on a posterior, review is weakly socially beneficial. Losses arise when endogenous coarsening pools cases from opposite sides of the immediate-action cutoff and changes which cases are reviewed. This can make a combined human–algorithm system underperform the relevant standalone benchmark even without behavioral frictions.

Holding score informativeness fixed, shifting liability toward the decision maker mitigates this deployment-stage distortion. When score informativeness is endogenous, however, the same liability shift can weaken incentives to improve the score. The design of human review systems therefore depends not only on prediction quality, but also on how advice shapes the allocation of review.

A Decision problem and review geometry

This appendix proves the results of Section III. Appendix A.1 proves Lemma 1, Appendix A.2 proves Lemma 2, and Appendix A.3 proves Proposition 1.

A.1 Proof of Lemma 1

Fix $\mu \in [0,1]$. The concavification identity

$$\sup_{\tau:\, \mathbb{E}_\tau[\nu]=\mu} \mathbb{E}_{\nu\sim\tau}[N(\nu)] = (\operatorname{cav} N)(\mu)$$

is standard in Bayesian persuasion (Kamenica and Gentzkow, 2011) and rational inattention with posterior-separable costs (Caplin and Dean, 2015).

For the upper bound, let h be any concave function with $h \geq N$. Then for any Bayes-plausible posterior law τ with $\mathbb{E}_\tau[\nu] = \mu$, Jensen's inequality implies

$$\mathbb{E}_\tau[N(\nu)] \leq \mathbb{E}_\tau[h(\nu)] \leq h(\mathbb{E}_\tau[\nu]) = h(\mu).$$

Taking the supremum over τ and then the infimum over such h yields

$$\sup_{\tau:\, \mathbb{E}_\tau[\nu]=\mu} \mathbb{E}_\tau[N(\nu)] \leq (\operatorname{cav} N)(\mu).$$

For the lower bound, by definition

$$(\operatorname{cav} N)(\mu) = \sup\left\{\sum_i w_i N(\mu_i) : w_i \geq 0,\ \sum_i w_i = 1,\ \sum_i w_i \mu_i = \mu\right\}.$$

Each feasible convex combination corresponds to a Bayes-plausible finite-support distribution τ with $\Pr_\tau(\nu = \mu_i) = w_i$, so every such value is attainable as $\mathbb{E}_\tau[N(\nu)]$. Hence

$$\sup_{\tau:\, \mathbb{E}_\tau[\nu]=\mu} \mathbb{E}_\tau[N(\nu)] = (\operatorname{cav} N)(\mu).$$

The DM therefore compares two options at posterior μ: abstain and obtain $N(\mu)$, or acquire information and obtain $(\operatorname{cav} N)(\mu) - K$. Rearranging yields

$$V(\mu;\alpha) + \lambda\phi(\mu) = \max\{N(\mu),\ (\operatorname{cav} N)(\mu) - K\}.$$

A.2 Proof of Lemma 2

The argument has three parts: affinity of the concave envelope on active components, uniqueness of the active interval, and tangency at the endpoints.

Step 1: Affinity on each active component. Let

$$\mathcal{I} := \{\mu \in [0,1] : (\operatorname{cav} N)(\mu) > N(\mu)\}$$

denote the set on which information acquisition is strictly valuable. Fix any connected component J of $\mathcal{I}$ and any $\mu \in J$.

By Lemma 1,

$$(\operatorname{cav} N)(\mu) = \sup_{\tau:\, \mathbb{E}_\tau[\nu]=\mu} \mathbb{E}_\tau[N(\nu)].$$

Because N is continuous and bounded on $[0,1]$, the supremum is attained. Since the posterior is one-dimensional and Bayes plausibility imposes a single moment restriction, an optimizer can be chosen with support size at most two.[12] Because $\mu \in \mathcal{I}$, the optimizer is nondegenerate, so it has support $\{x_1, x_2\}$ with

$$x_1 < \mu < x_2.$$

Let L denote the affine function through $(x_1, N(x_1))$ and $(x_2, N(x_2))$. Then

$$(\operatorname{cav} N)(\mu) = L(\mu).$$

Write $\mu = tx_1 + (1-t)x_2$ for the weight $t \in (0,1)$ induced by the two-point optimizer. Because $\{x_1, x_2\}$ is an optimal support at mean μ,

$$(\operatorname{cav} N)(\mu) = tN(x_1) + (1-t)N(x_2).$$

Concavity of $\operatorname{cav} N$ gives

$$(\operatorname{cav} N)(\mu) \geq t(\operatorname{cav} N)(x_1) + (1-t)(\operatorname{cav} N)(x_2),$$

and since $\operatorname{cav} N \geq N$ pointwise,

$$t(\operatorname{cav} N)(x_1) + (1-t)(\operatorname{cav} N)(x_2) \geq tN(x_1) + (1-t)N(x_2) = (\operatorname{cav} N)(\mu).$$

Hence both inequalities bind. Therefore

$$(\operatorname{cav} N)(x_1) = N(x_1), \qquad (\operatorname{cav} N)(x_2) = N(x_2).$$

So $\operatorname{cav} N$ and the affine line L coincide at x_1 and x_2.

It follows that $(\operatorname{cav} N)(z) = L(z)$ for every $z \in (x_1, x_2)$. Since $\operatorname{cav} N$ is concave and coincides with L at x_1 and x_2,

$$(\operatorname{cav} N)(z) \geq L(z) \qquad \text{for all } z \in (x_1, x_2).$$

[12] Equivalently, the feasible set is characterized by the two linear equalities $\tau([0,1]) = 1$ and $\mathbb{E}_\tau[\nu] = \mu$, so an extreme-point optimizer has support size at most two.

Suppose instead that there exists $z' \in (x_1, x_2)$ such that

$$(\operatorname{cav} N)(z') > L(z').$$

Without loss of generality, take $z' < \mu$ and write

$$\mu = tz' + (1-t)x_2 \qquad \text{for some } t \in (0,1).$$

Let $\tau^{z'}$ be an optimal Bayes-plausible experiment at mean z', so that

$$\mathbb{E}_{\tau^{z'}}[N(\nu)] = (\operatorname{cav} N)(z').$$

Consider the composite experiment at mean μ that, with probability t, runs $\tau^{z'}$ and, with probability $1-t$, places degenerate mass at x_2. This composite experiment is Bayes-plausible with mean μ and delivers expected value

$$t(\operatorname{cav} N)(z') + (1-t)N(x_2) > tL(z') + (1-t)L(x_2) = L(\mu) = (\operatorname{cav} N)(\mu),$$

contradicting the definition of $(\operatorname{cav} N)(\mu)$ as the maximal Bayes-plausible value at mean μ. Hence $(\operatorname{cav} N) = L$ on (x_1, x_2).

Because $\mu \in \mathcal{I}$, $L(\mu) > N(\mu)$, so by continuity there is a neighborhood of μ on which $L > N$. Repeating the same argument for every point of the connected component J and using overlap of neighborhoods implies that $\operatorname{cav} N$ is affine on all of J.

Step 2: The active set is a unique interval. Write

$$N(\mu) = \begin{cases} N_R(\mu) := -\alpha D_{FN}\mu + \lambda\phi(\mu), & \mu \in [0, \overline{\mu}], \\ N_A(\mu) := -(1-\mu)\alpha D_{FP} + \lambda\phi(\mu), & \mu \in [\overline{\mu}, 1]. \end{cases}$$

Each branch is strictly concave because ϕ is strictly concave. At the kink $\overline{\mu}$,

$$N_A'(\overline{\mu}^+) - N_R'(\overline{\mu}^-) = \alpha(D_{FP} + D_{FN}) > 0.$$

A concave function cannot have an upward jump in slope, so N is not concave. Therefore

$$\mathcal{I} \neq \varnothing.$$

Any connected component of $\mathcal{I}$ must contain $\overline{\mu}$. Suppose a component J were contained in $[0, \overline{\mu}]$. By Step 1, $\operatorname{cav} N$ is affine on J, and by continuity $\operatorname{cav} N = N$ at the boundary points of J. But N_R is strictly concave on $[0, \overline{\mu}]$, so the affine chord through two distinct points of N_R lies strictly below N_R in the interior, contradicting $\operatorname{cav} N \geq N$ on J. The same argument rules out a component contained in $[\overline{\mu}, 1]$.

Hence every connected component of $\mathcal{I}$ must straddle $\overline{\mu}$. Since two distinct connected components cannot both contain $\overline{\mu}$, the set $\mathcal{I}$ has exactly one connected component, namely $\mathcal{I} = (\mu_L^*, \mu_H^*)$ with $\mu_L^* < \overline{\mu} < \mu_H^*$. This establishes existence and uniqueness of the interval.

Step 3: Tangency conditions at the endpoints. Let $L(\mu) = a + s\mu$ denote the affine restriction of $\operatorname{cav} N$ to $[\mu_L^*, \mu_H^*]$. Because $\operatorname{cav} N = N$ at the endpoints,

$$s = \frac{N(\mu_H^*) - N(\mu_L^*)}{\mu_H^* - \mu_L^*}.$$

Since $\operatorname{cav} N$ is concave and equals N_R on $[0, \mu_L^*]$, its left-hand derivative at μ_L^* is $N_R'(\mu_L^*)$; its right-hand derivative at μ_L^* is s (the slope of the affine piece on $[\mu_L^*, \mu_H^*]$). Concavity requires

$$N_R'(\mu_L^*) \geq s.$$

Suppose $N_R'(\mu_L^*) > s$. Then for $\mu \in (\mu_L^*, \mu_L^* + \delta)$ with $\delta > 0$ small,

$$L(\mu) - N_R(\mu) \approx (s - N_R'(\mu_L^*))(\mu - \mu_L^*) < 0,$$

so $L < N_R = N$ on that interval. But $\operatorname{cav} N = L$ on (μ_L^*, μ_H^*), giving $\operatorname{cav} N < N$ there—contradicting $\operatorname{cav} N \geq N$ pointwise. Therefore

$$N_R'(\mu_L^*) = s.$$

By the same reasoning at μ_H^*: $\operatorname{cav} N$ has left-hand derivative s and right-hand derivative $N_A'(\mu_H^*)$ there. Concavity gives $s \geq N_A'(\mu_H^*)$, and if $s > N_A'(\mu_H^*)$ then $L < N_A = N$ on $(\mu_H^* - \delta, \mu_H^*)$ for small $\delta > 0$ (by the same first-order approximation), again contradicting $\operatorname{cav} N = L \geq N$. Therefore

$$N_A'(\mu_H^*) = s.$$

Substituting the branch derivatives,

$$N_R'(\mu_L^*) = -\alpha D_{FN} + \lambda\phi'(\mu_L^*), \qquad N_A'(\mu_H^*) = \alpha D_{FP} + \lambda\phi'(\mu_H^*),$$

yields

$$\frac{N(\mu_H^*) - N(\mu_L^*)}{\mu_H^* - \mu_L^*} = -\alpha D_{FN} + \lambda\phi'(\mu_L^*), \tag{3a}$$

$$\frac{N(\mu_H^*) - N(\mu_L^*)}{\mu_H^* - \mu_L^*} = \alpha D_{FP} + \lambda\phi'(\mu_H^*). \tag{3b}$$

Because the interval (μ_L^*, μ_H^*) is unique, the pair (μ_L^*, μ_H^*) satisfying these conditions is unique as well.

A.3 Proof of Proposition 1

The argument first characterizes the geometry of Δ on $[\mu_L^*, \mu_H^*]$ and then identifies the review region as the superlevel set $\{\mu : \Delta(\mu) \geq K\}$.

Lemma A.1 (Properties of Δ). *On $[\mu_L^*, \mu_H^*]$:*

1. *$\Delta(\mu_L^*) = \Delta(\mu_H^*) = 0$ and $\Delta(\mu) > 0$ for $\mu \in (\mu_L^*, \mu_H^*)$.*
2. *Δ is strictly increasing on $[\mu_L^*, \overline{\mu}]$ and strictly decreasing on $[\overline{\mu}, \mu_H^*]$, with maximum at $\overline{\mu}$.*
3. *For $K \in (0, \Delta(\overline{\mu}))$, $\Delta(\mu) = K$ has exactly two solutions $\mu^L \in (\mu_L^*, \overline{\mu})$ and $\mu^H \in (\overline{\mu}, \mu_H^*)$.*

Proof. **Part 1.** By construction, $\Delta(\mu) := (\operatorname{cav} N)(\mu) - N(\mu)$, which is zero at the tangency points and strictly positive on the strict concavification region. This follows directly from the characterization in the proof of Lemma 2.

Part 2. On $[\mu_L^*, \mu_H^*]$, $(\operatorname{cav} N)(\mu)$ is linear with slope s. Thus, for $\mu \neq \overline{\mu}$, $\Delta'(\mu) = s - N'(\mu)$. Differentiating again gives $\Delta''(\mu) = -N''(\mu) = -\lambda\phi''(\mu)$. Since ϕ is strictly concave, $\Delta''(\mu) > 0$, so $\Delta'(\mu)$ is strictly increasing on both $(\mu_L^*, \overline{\mu})$ and $(\overline{\mu}, \mu_H^*)$.

On the left branch: the tangency condition at μ_L^* (Eq. (3a)) implies $\Delta'(\mu_L^*) = 0$. Since Δ' is strictly increasing, $\Delta'(\mu) > 0$ for all $\mu \in (\mu_L^*, \overline{\mu})$, so Δ is strictly increasing there. On the right branch: symmetrically, the tangency condition at μ_H^* gives $\Delta'(\mu_H^*) = 0$, so $\Delta'(\mu) < 0$ for all $\mu \in (\overline{\mu}, \mu_H^*)$, and Δ is strictly decreasing there.

Thus, Δ is strictly increasing on $[\mu_L^*, \overline{\mu}]$ and strictly decreasing on $[\overline{\mu}, \mu_H^*]$. Hence it attains its unique maximum at $\overline{\mu}$.

Part 3. The function $\Delta(\mu)$ is continuous and strictly monotone on each side of $\overline{\mu}$. Since $\Delta(\mu_L^*) = 0$ and $\Delta(\overline{\mu}) > K$, the Intermediate Value Theorem implies there exists a unique $\mu^L \in (\mu_L^*, \overline{\mu})$ with $\Delta(\mu^L) = K$. A symmetric argument establishes the existence of a unique $\mu^H \in (\overline{\mu}, \mu_H^*)$. □

By Lemma 1, the DM reviews if and only if $\Delta(\mu) \geq K$. Lemma A.1 shows that Δ is continuous, strictly increasing on $[\mu_L^*, \overline{\mu}]$, and strictly decreasing on $[\overline{\mu}, \mu_H^*]$. Therefore the review region is empty when $K > \Delta(\overline{\mu})$ and otherwise is the closed interval $[\mu^L, \mu^H]$, where μ^L and μ^H are the unique solutions to $\Delta(\mu) = K$ on the two sides of $\overline{\mu}$. When $K = 0$, these two solutions are μ_L^* and μ_H^* because $\Delta(\mu_L^*) = \Delta(\mu_H^*) = 0$. When $K = \Delta(\overline{\mu})$, the two thresholds coincide at $\overline{\mu}$ because Δ attains its unique maximum there.

B Optimal-design geometry

This section proves the disclosure results of Section IV. Appendix B.1 states the convex-analytic verification tool used throughout. Appendix B.2 proves the general-case classification behind Theorem 1. Appendix B.3 justifies the finite-implementation counts in Corollary 1. Appendix B.4 studies the fixed-cost benchmark ($\lambda = 0$) used for the closed-form illustration.

B.1 Shared verification framework

All constructions in this section use the convex-majorant argument from Dworczak and Martini (2019). The following two dual facts are used repeatedly below. The case proofs then specify a

candidate design, a supporting convex price, and verify domination, support, and budget balance.

Verification tool (Dworczak and Martini (2019), Theorems 1–2). A distribution of posteriors $G \in \text{MPC}(F)$ is optimal if and only if[13] there exists a convex function $p : [0,1] \to \mathbb{R}$ such that: (i) *Domination:* $p(\mu) \geq \widetilde{u}(\mu)$ for all μ; (ii) *Support:* $\text{supp}(G) \subseteq \{\mu : p(\mu) = \widetilde{u}(\mu)\}$; (iii) *Budget balance:* $\int_0^1 p\, dF = \int_0^1 p\, dG$.

Dual minimization characterization (Dworczak and Martini (2019), Proposition 1). A price function p^* is an equilibrium price function if and only if it solves

$$\min_p \int_0^1 p(\mu)\, dF(\mu) \tag{14}$$

subject to p being convex on $[0,1]$ and $p(\mu) \geq \widetilde{u}(\mu)$ for all μ.

Verification template. In each case, optimality is verified in the same way. A candidate design G and a convex function p are specified such that: (i) $p \geq \widetilde{u}$ pointwise; (ii) the support of G lies on the contact set $\{p = \widetilde{u}\}$; and (iii) budget balance holds block by block. The Verification Tool then implies optimality. Uniqueness below refers to the geometric configuration up to payoff-equivalent refinements on affine revealed regions.

Proof-specific notation. Only a small set of proof-specific symbols appears repeatedly below. The review gates are μ^L, μ^H. The symbols $m'_1, \ldots, m'_4$ in the general case and $m_1, \ldots, m_4$ in the benchmark denote the critical intersections that delimit candidate pooled blocks; $\theta'_1, \theta'_{23}, \theta'_4$ and $\theta_1, \theta_{23}, \theta_4$ are the associated conditional means. The remaining symbols are local to the central-region analysis: $\overline{m}^L, \overline{m}^H$ are maximal one-sided reaches, m_L, m_H delimit the residual revealing interval when it exists, and $x_L(\cdot), x_H(\cdot), k^*, x_L^*, x_H^*$ are the cutoff objects for the residual-disclosure and single-touch cases. All other notation is defined on first use.

Let

$$u_R(\mu) := -\mu C_{FN}, \qquad u_A(\mu) := -(1-\mu) C_{FP}$$

denote the left and right affine wings of the provider payoff $\widetilde{u}$ outside the review region.

B.2 General geometric refinement ($\lambda > 0$) supporting Theorem 1

This subsection proves the general-case classification behind Theorem 1 when $\lambda > 0$. The main text groups optimal designs into three economic classes—single-gate pooling, two-sided gate pooling, and residual interior disclosure—and the finer geometric refinement is derived here. The one-gate configurations use the same affine-support construction and are verified first. The central-region configurations then separate according to whether the mass between the two review gates can be absorbed fully or requires an interior residue.

[13] Theorem 2 in Dworczak and Martini (2019) requires a regularity condition (a weaker condition is given by Dizdar and Kováč (2020)), which is satisfied by $\widetilde{u}$: bounded and upper-semicontinuous.

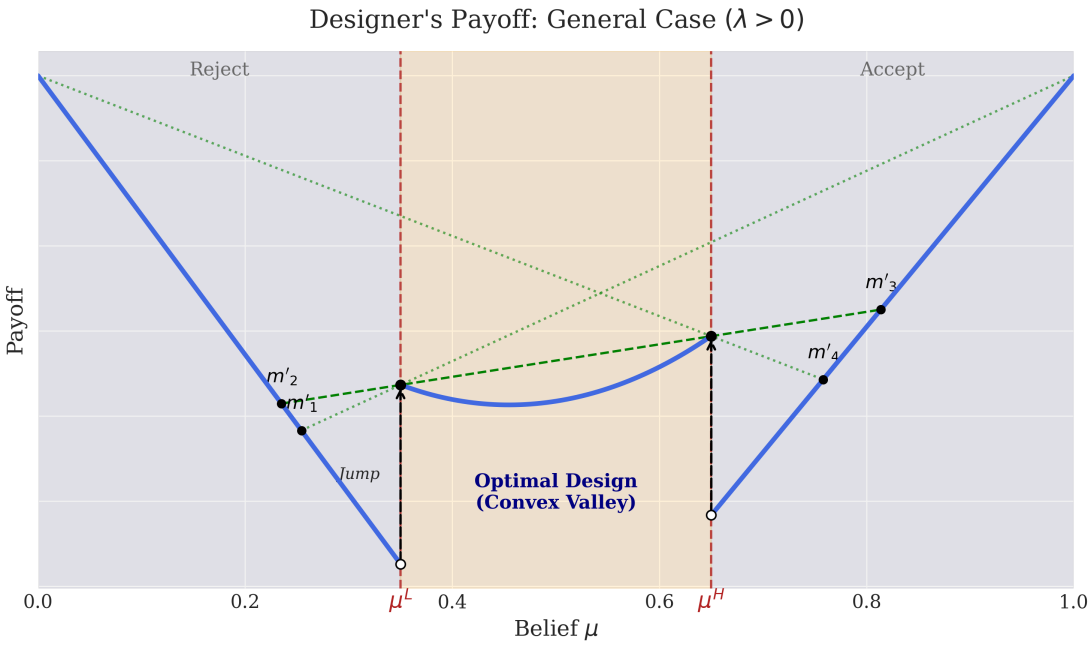


Figure B.1: Provider payoff valley in the general case ($\lambda > 0$)**.** Inside the review region $[\mu^L, \mu^H]$, the payoff $\widetilde{u}$ is strictly convex because of the information-cost term $-\beta\lambda I^*(\mu)$, with upward jumps at the review gates proportional to βK. This valley geometry underlies the central-region refinement in Appendix B.2.

Critical intersections and conditional means (general case). The cutoff objects used in Proposition B.1 are the general-case analogues of the benchmark intersections, with the plateau level $-\beta K$ replaced by the boundary continuation values $\widetilde{u}(\mu^L)$ and $\widetilde{u}(\mu^H)$:

- Let $\psi_{bridge}(\cdot)$ denote the line through $(\mu^L, \widetilde{u}(\mu^L))$ and $(\mu^H, \widetilde{u}(\mu^H))$. Define m'_2 and m'_3 by $\psi_{bridge}(m'_2) = u_R(m'_2)$ and $\psi_{bridge}(m'_3) = u_A(m'_3)$.
- Let $\ell^L_R(\cdot)$ denote the line through $(1, 0)$ and $(\mu^L, \widetilde{u}(\mu^L))$. Define m'_1 as the (unique) μ solving $u_R(\mu) = \ell^L_R(\mu)$.
- Let $\ell^H_A(\cdot)$ denote the line through $(0, 0)$ and $(\mu^H, \widetilde{u}(\mu^H))$. Define m'_4 as the (unique) μ solving $u_A(\mu) = \ell^H_A(\mu)$.

Define the associated conditional expectations:

$$\theta'_1 := \mathbb{E}[\mu \mid \mu \geq m'_1], \qquad \theta'_{23} := \mathbb{E}[\mu \mid \mu \in [m'_2, m'_3]], \qquad \theta'_4 := \mathbb{E}[\mu \mid \mu \leq m'_4].$$

Maximal one-sided reaches. In the central-region refinement, the first question is whether the bridge interval $[m'_2, m'_3]$ can be absorbed fully by the two gate posteriors. Define the maximal reaches $\overline{m}^L, \overline{m}^H$ by

$$\mathbb{E}[\mu \mid \mu \in [m'_2, \overline{m}^L]] = \mu^L, \qquad \mathbb{E}[\mu \mid \mu \in [\overline{m}^H, m'_3]] = \mu^H. \tag{15}$$

When $\overline{m}^L \geq \overline{m}^H$, the bridge mass can be absorbed directly by the two gates. When $\overline{m}^L < \overline{m}^H$, bridge pooling is infeasible, and the residual-feasibility objects defined next determine whether

the optimal design is residual interior disclosure or single-touch pooling.

Auxiliary objects when bridge pooling is infeasible. When $\overline{m}^L < \overline{m}^H$, bridge pooling is infeasible. The following objects distinguish residual interior disclosure from single-touch pooling. Fix any pivot $k \in [\mu^L, \mu^H]$. Define

- $\psi_L(\cdot; k) : [0,1] \to \mathbb{R}$ as the affine function defining the line passing through $(\mu^L, \widetilde{u}(\mu^L))$ and $(k, \widetilde{u}(k))$;
- $\psi_H(\cdot; k) : [0,1] \to \mathbb{R}$ as the affine function defining the line passing through $(k, \widetilde{u}(k))$ and $(\mu^H, \widetilde{u}(\mu^H))$.

Lemma B.1 (Existence/uniqueness of cutoffs)**.** *For each $k \in [\mu^L, \mu^H]$, the affine functions $\psi_L(\cdot; k)$ and $u_R(\cdot)$ intersect at a unique point $\hat{x}_L(k) \in \mathbb{R}$, and $\psi_H(\cdot; k)$ and $u_A(\cdot)$ intersect at a unique point $\hat{x}_H(k) \in \mathbb{R}$. Moreover, the mappings $k \mapsto \hat{x}_L(k)$ and $k \mapsto \hat{x}_H(k)$ are continuous.*

Proof. Both $\psi_L(\cdot; k)$ and $u_R(\cdot)$ are affine in μ. Write the slope of u_R as $-C_{FN}$ and the slope of $\psi_L(\cdot; k)$ as $s_L(k) := \frac{\widetilde{u}(k) - \widetilde{u}(\mu^L)}{k - \mu^L}$. Because $\widetilde{u}$ is strictly convex on (μ^L, μ^H) and lies strictly above the affine continuation u_R at μ^L (upward boundary jump), the secant slope $s_L(k)$ is strictly greater than $-C_{FN}$, so the two lines cannot be parallel. Hence, they intersect at a unique point obtained by solving a linear equation. Continuity in k follows because $s_L(k)$ depends continuously on $\widetilde{u}(k)$, and the intersection point is a rational expression in the coefficients of the two affine functions. The argument for $\psi_H(\cdot; k)$ and $u_A(\cdot)$ is identical. □

Define effective cutoffs $x_L(k) := \max\{\hat{x}_L(k), 0\}$ and $x_H(k) := \min\{\hat{x}_H(k), 1\}$.

Define the gap functions

$$\begin{aligned} \Delta_L(k) &:= \mathbb{E}[\mu \mid \mu \in [x_L(k), k]] - \mu^L, \\ \Delta_H(k) &:= \mathbb{E}[\mu \mid \mu \in [k, x_H(k)]] - \mu^H. \end{aligned} \tag{16}$$

These functions measure whether the left/right pooling blocks generated by the pivot k can hit the target posteriors μ^L and μ^H.

Because conditional means vary continuously with interval endpoints, Lemma B.1 and the continuity of truncation ($\max\{\cdot, 0\}$ and $\min\{\cdot, 1\}$) imply that Δ_L and Δ_H are continuous on $[\mu^L, \mu^H]$.

Boundary signs and existence. As $k \to \mu^H$ from below, $\psi_L(\cdot; k) \to \psi_{bridge}(\cdot)$, hence $x_L(\mu^H) = m_2'$ and

$$\Delta_L(\mu^H) = \mathbb{E}[\mu \mid \mu \in [m_2', \mu^H]] - \mu^L > 0$$

when $\overline{m}^L < \overline{m}^H$. As $k \to \mu^L$ from above, the left pooling interval collapses around μ^L, so $\Delta_L(\mu^L) < 0$. Symmetrically, $x_H(\mu^L) = m_3'$ implies $\Delta_H(\mu^L) < 0$, while $\Delta_H(\mu^H) > 0$ as the right pooling interval collapses around μ^H.

Therefore, by the Intermediate Value Theorem, each target-mean constraint admits (possibly

non-unique) solutions. Define the extremal solutions

$$m_L := \inf\{k \in (\mu^L, \mu^H] : \Delta_L(k) \geq 0\}, \qquad m_H := \sup\{k \in [\mu^L, \mu^H) : \Delta_H(k) \leq 0\}. \tag{17}$$

The remaining two central-region cases are separated by whether the left and right gate-targeting blocks can be chosen while still leaving a positive gap between them. If they can, the optimal design is residual interior disclosure, so a nondegenerate interior interval remains fully revealing. If they cannot, the two pooled blocks meet and the design collapses to a single-touch configuration.

Proposition B.1 (Geometric classification (general case $\lambda > 0$))**.** *Let $\theta'_1, \theta'_{23}, \theta'_4$ be the conditional means defined above, and let $\overline{m}^L, \overline{m}^H$ be the maximal reaches and m_L, m_H the extremal feasibility thresholds defined above. Then the optimal recommendation policy, equivalently the optimal implementable posterior distribution, takes one of the following mutually exclusive and exhaustive forms:*

1. ***Tail pooling toward the left gate*** *($\theta'_1 \leq \mu^L$): pool $\mu \in [x_1, 1]$ to posterior μ^L and reveal outside, where $x_1 \in (m'_1, 1)$ solves $\mathbb{E}[\mu \mid \mu \geq x_1] = \mu^L$.*
2. ***Cross-region pooling toward the left gate*** *($\theta'_1 > \mu^L > \theta'_{23}$): pool $\mu \in [x_2, x_3]$ to posterior μ^L and reveal outside, where $x_2 \in (m'_1, \mu^L)$ and $x_3 \in (\mu^H, 1)$ satisfy*
   (i) the points $(x_2, u_R(x_2))$, $(\mu^L, \widetilde{u}(\mu^L))$, and $(x_3, u_A(x_3))$ are collinear;
   (ii) Bayes plausibility holds: $\mathbb{E}[\mu \mid \mu \in [x_2, x_3]] = \mu^L$.
3. ***Central region*** *($\mu^L \leq \theta'_{23} \leq \mu^H$):*
   *(a) **Bridge pooling** ($\overline{m}^L \geq \overline{m}^H$): pool $[m'_2, m'_3]$ and split the pooled mass between posteriors μ^L and μ^H with weight $\pi = (\mu^H - \theta'_{23})/(\mu^H - \mu^L)$ on μ^L; reveal outside.*
   *(b) **Residual interior disclosure** ($\overline{m}^L < \overline{m}^H$ and $m_L < m_H$): pool a left block to μ^L and a right block to μ^H, leaving a nondegenerate interior interval (m_L, m_H) fully revealing.*
   *(c) **Single-touch pooling** ($\overline{m}^L < \overline{m}^H$ and $m_L \geq m_H$): there exist cutoffs $x_L^* \leq k^* \leq x_H^*$ such that the design pools $\mu \in [x_L^*, k^*]$ to μ^L and pools $\mu \in [k^*, x_H^*]$ to μ^H. In both cases the two pooled blocks meet at a common point k^*; when $m_L > m_H$ that common point is the strict pivot, and when $m_L = m_H$ it is the shared boundary point.*
4. ***Cross-region pooling toward the right gate*** *($\theta'_{23} > \mu^H > \theta'_4$): pool $\mu \in [x_5, x_4]$ to posterior μ^H and reveal outside, where $x_4 \in (\mu^H, 1)$ and $x_5 \in (0, \mu^L)$ satisfy*
   (i) the points $(x_5, u_R(x_5))$, $(\mu^H, \widetilde{u}(\mu^H))$, and $(x_4, u_A(x_4))$ are collinear;
   (ii) Bayes plausibility holds: $\mathbb{E}[\mu \mid \mu \in [x_5, x_4]] = \mu^H$.
5. ***Tail pooling toward the right gate*** *($\theta'_4 \geq \mu^H$): pool $\mu \in [0, x_6]$ to posterior μ^H and reveal outside, where $x_6 \in (0, m'_4)$ solves $\mathbb{E}[\mu \mid \mu \leq x_6] = \mu^H$.*

Remark B.1 (Implementation degeneracy within single-touch pooling)**.** *In Case 3(c), the supporting-price geometry is the same regardless of whether the pooling cutoffs x_L^*, x_H^* lie in the interior of $[0, 1]$ or at the boundaries. However, the implementation cardinality varies:*

- *If $0 < x_L^* < k^* < x_H^* < 1$: **quaternary** (two gate posteriors plus two residual posteriors).*
- *If exactly one of $x_L^* = 0$ or $x_H^* = 1$: **ternary** (one revealing tail vanishes).*

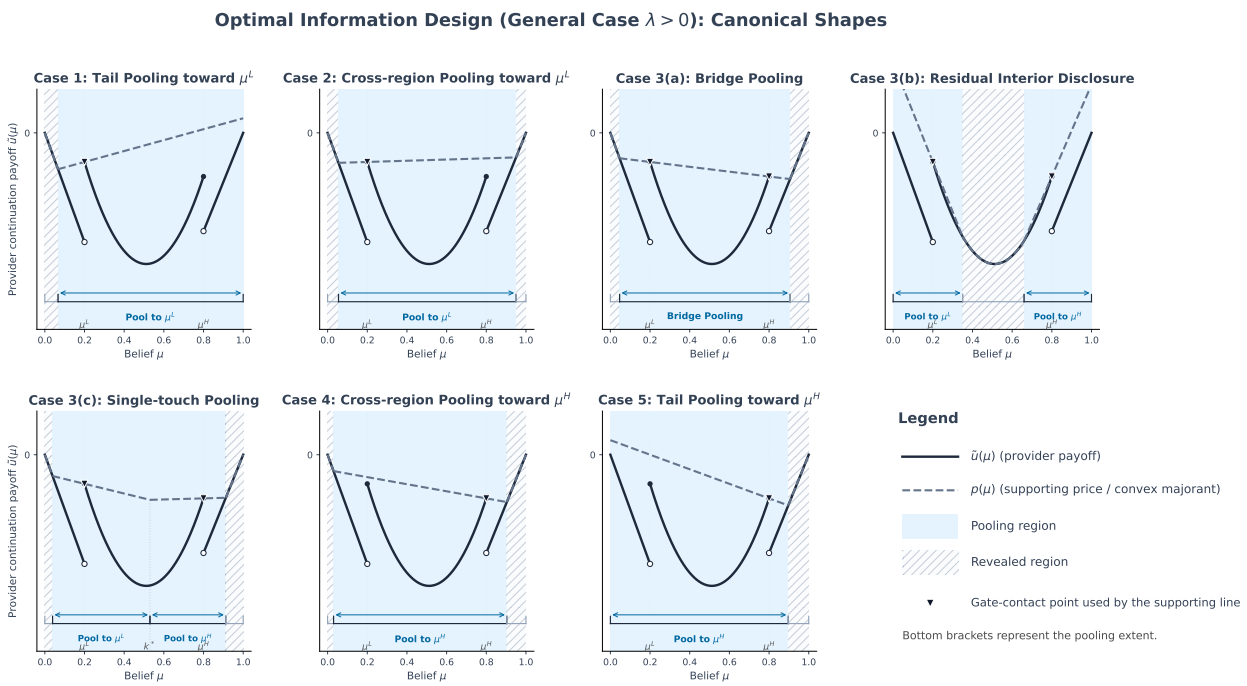


Figure B.2: Exhaustive geometric refinement of the three economic classes ($\lambda > 0$). Shaded regions indicate pooling; unshaded regions indicate full revelation. Cases 1, 2, 4, and 5 refine the single-gate pooling class. Cases 3(a) and 3(c) refine the two-sided gate-pooling class. Case 3(b) is the residual interior-disclosure class. This figure supplements Proposition B.1 and Theorem 1.

- *If $x_L^* = 0$ and $x_H^* = 1$: **binary** (both revealing tails vanish).*

This is an implementation-complexity refinement within the same geometric case, not a new optimality classification.

Cases 1, 2, 4, and 5 use the same affine-support construction, so the candidate design and supporting line are stated for each case while repeated verification details are suppressed. On the pooled block, the supporting price is affine and passes through the relevant gate-contact point; outside that block, the price coincides with $\widetilde{u}$. Because the pooled block is sent to its conditional mean, budget balance follows block by block.

Case 1. Tail pooling toward the left gate.

(i) *Candidate design G.* Reveal μ on $[0, x_1)$ and pool $\mu \in [x_1, 1]$ to posterior μ^L, where $x_1 \in (m_1', 1)$ uniquely solves $\mathbb{E}[\mu \mid \mu \geq x_1] = \mu^L$.

(ii) *Construct p.* Let ψ_1 be the affine function passing through $(x_1, u_R(x_1))$ and $(\mu^L, \widetilde{u}(\mu^L))$, and define

$$p(\mu) = \begin{cases} \widetilde{u}(\mu), & \mu < x_1, \\ \psi_1(\mu), & \mu \geq x_1. \end{cases}$$

Case 2. Cross-region pooling toward the left gate.

(i) *Candidate design G.* Reveal μ outside $[x_2, x_3]$ and pool $\mu \in [x_2, x_3]$ to posterior μ^L, where $x_2 \in (m_1', \mu^L)$ and $x_3 \in (\mu^H, 1)$ satisfy the collinearity and Bayes-plausibility conditions in

Proposition B.1(2).

(ii) *Construct p.* Let ψ_2 be the affine function through $(x_2, u_R(x_2))$ and $(x_3, u_A(x_3))$ (equivalently, ψ_2 also passes through $(\mu^L, \widetilde{u}(\mu^L))$), and define

$$p(\mu) = \begin{cases} \widetilde{u}(\mu), & \mu < x_2, \\ \psi_2(\mu), & x_2 \le \mu \le x_3, \\ \widetilde{u}(\mu), & \mu > x_3. \end{cases}$$

Case 4. Cross-region pooling toward the right gate.

(i) *Candidate design G.* Reveal μ outside $[x_5, x_4]$ and pool $\mu \in [x_5, x_4]$ to posterior μ^H, where $x_5 \in (0, \mu^L)$ and $x_4 \in (\mu^H, 1)$ satisfy the collinearity and Bayes-plausibility conditions in Proposition B.1(4).

(ii) *Construct p.* Let ψ_4 be the affine function through $(x_5, u_R(x_5))$ and $(x_4, u_A(x_4))$ (equivalently, ψ_4 also passes through $(\mu^H, \widetilde{u}(\mu^H))$), and define

$$p(\mu) = \begin{cases} \widetilde{u}(\mu), & \mu < x_5, \\ \psi_4(\mu), & x_5 \le \mu \le x_4, \\ \widetilde{u}(\mu), & \mu > x_4. \end{cases}$$

Case 5. Tail pooling toward the right gate.

(i) *Candidate design G.* Pool $\mu \in [0, x_6]$ to posterior μ^H and reveal μ on $(x_6, 1]$, where $x_6 \in (0, m_4')$ uniquely solves $\mathbb{E}[\mu \mid \mu \le x_6] = \mu^H$.

(ii) *Construct p.* Let ψ_5 be the affine function passing through $(x_6, u_A(x_6))$ and $(\mu^H, \widetilde{u}(\mu^H))$, and define

$$p(\mu) = \begin{cases} \psi_5(\mu), & \mu \le x_6, \\ \widetilde{u}(\mu), & \mu > x_6. \end{cases}$$

Central-region cases. The remaining configurations place the critical mass between the two review gates. The distinction among Cases 3(a), 3(b), and 3(c) is whether that central mass can be absorbed entirely by the two gates or whether some interior structure must survive in the optimal design.

Case 3(a). Bridge pooling. In the central-region case $\mu^L \le \theta_{23}' \le \mu^H$, let ψ_{bridge} be the bridge line through $(\mu^L, \widetilde{u}(\mu^L))$ and $(\mu^H, \widetilde{u}(\mu^H))$, with m_2', m_3' defined above. The objects $\overline{m}^L$ and $\overline{m}^H$, defined in (15), measure how far one-sided gate absorption can extend while still hitting the target gate posteriors. When these maximal reaches overlap, the central bridge mass can be split directly between the two gates. When they do not overlap, the central region must instead be handled by residual interior disclosure or a single-touch design.

Define the total mass on the bridge interval $M := F(m_3') - F(m_2') > 0$ and the target mixing weight $\pi := \frac{\mu^H - \theta_{23}'}{\mu^H - \mu^L} \in [0,1]$, so that Bayes plausibility for a binary split on $\{\mu^L, \mu^H\}$ requires assigning mass πM to posterior μ^L and mass $(1-\pi)M$ to posterior μ^H.

(i) *Candidate design G.* Pool beliefs $\mu \in [m_2', m_3']$ and split the pooled mass between posteriors μ^L and μ^H with weights π and $1-\pi$, while revealing beliefs outside $[m_2', m_3']$.

(ii) *Construct p.* Define p_{bridge} by

$$p_{bridge}(\mu) := \begin{cases} u_R(\mu), & \mu < m_2', \\ \psi_{bridge}(\mu), & \mu \in [m_2', m_3'], \\ u_A(\mu), & \mu > m_3'. \end{cases}$$

(iii) *Feasibility and implementation check.* Since $\psi_{bridge}(\mu) > \widetilde{u}(\mu)$ for all $\mu \in (\mu^L, \mu^H)$ (strict convexity), the Verification Tool, condition (ii), forces any optimal G to place no mass in (μ^L, μ^H). Hence, conditional on $\mu \in [m_2', m_3']$, the design must be a split supported on $\{\mu^L, \mu^H\}$, and budget balance pins down the split weight to π. Such a split is Bayes-plausible if and only if $\overline{m}^L \geq \overline{m}^H$.

If $\overline{m}^L \geq \overline{m}^H$, this binary split is feasible. An explicit Bayes-plausible implementation can be described by an assignment rule $s(\mu) \in [0,1]$ giving the probability of sending posterior μ^L at belief μ: take $s(\mu) = 1$ for $\mu \in [m_2', \overline{m}^H)$, take $s(\mu) = 0$ for $\mu \in (\overline{m}^L, m_3']$, and choose s on the overlap $[\overline{m}^H, \overline{m}^L]$ so that simultaneously

$$\int_{m_2'}^{m_3'} s(\mu)\, dF(\mu) = \pi M, \qquad \int_{m_2'}^{m_3'} \mu\, s(\mu)\, dF(\mu) = \pi M\, \mu^L.$$

Because F has a continuous strictly positive density and the overlap interval is nonempty, such a choice is always possible by varying the (possibly partial) assignment on the overlap.

The induced MPC G that pools $\mu \in [m_2', m_3']$ to a binary signal on $\{\mu^L, \mu^H\}$ can be represented by the CDF

$$G(\mu) = \begin{cases} F(\mu), & \mu \in [0, m_2'), \\ F(m_2'), & \mu \in [m_2', \mu^L), \\ F(m_2') + \pi\big(F(m_3') - F(m_2')\big), & \mu \in [\mu^L, \mu^H), \\ F(m_3'), & \mu \in [\mu^H, m_3'], \\ F(\mu), & \mu \in (m_3', 1], \end{cases}$$

where $\pi = (\mu^H - \theta_{23}')/(\mu^H - \mu^L)$.

Bridge infeasibility ($\overline{m}^L < \overline{m}^H$). When $\overline{m}^L < \overline{m}^H$, the overlap interval is empty and the bridge split becomes infeasible. Define the left-capacity $M_L := F(\overline{m}^L) - F(m_2')$ and right-capacity $M_H := F(m_3') - F(\overline{m}^H)$. Because strict convexity forces any optimal G to place no mass in (μ^L, μ^H) (the

Verification Tool, condition (ii)), Bayes plausibility requires mass πM at μ^L. But $M_L < \pi M$ when the overlap is empty: the interval $[m_2', \overline{m}^L]$ has conditional mean exactly μ^L, so any assignment exceeding M_L must include beliefs above $\overline{m}^L$, raising the conditional mean strictly above μ^L. The symmetric argument applies to M_H.

But when $\overline{m}^L < \overline{m}^H$, $M_L + M_H < M$ (the gap carries positive mass). Therefore, the two necessary inequalities cannot hold simultaneously: if $\pi M \leq M_L$, then

$$(1-\pi)M = M - \pi M \geq M - M_L > M_H,$$

contradicting $(1-\pi)M \leq M_H$. This proves bridge infeasibility.

Case 3(b). Residual interior disclosure.

(i) *Candidate design G.* Suppose $\overline{m}^L < \overline{m}^H$ and $m_L < m_H$. Pool $\mu \in [x_L(m_L), m_L]$ to posterior μ^L and pool $\mu \in [m_H, x_H(m_H)]$ to posterior μ^H, revealing beliefs outside these two pooled blocks.

(ii) *Construct p.* Define

$$p(\mu) = \begin{cases} u_R(\mu), & \mu < x_L(m_L), \\ \psi_L(\mu; m_L), & \mu \in [x_L(m_L), m_L], \\ \widetilde{u}(\mu), & \mu \in (m_L, m_H), \\ \psi_H(\mu; m_H), & \mu \in [m_H, x_H(m_H)], \\ u_A(\mu), & \mu > x_H(m_H). \end{cases}$$

Verification of Case 3(b). By definition of m_L and m_H, the left pooled block $[x_L(m_L), m_L]$ has conditional mean μ^L, and the right pooled block $[m_H, x_H(m_H)]$ has conditional mean μ^H. Hence Bayes plausibility holds block by block.

It remains to verify domination and convexity of p. On the interval $[\mu^L, m_L]$, the function $\psi_L(\cdot; m_L)$ is the chord joining $(\mu^L, \widetilde{u}(\mu^L))$ and $(m_L, \widetilde{u}(m_L))$. Because $\widetilde{u}$ is strictly convex on (μ^L, μ^H),

$$\psi_L(\mu; m_L) \geq \widetilde{u}(\mu) \qquad \text{for all } \mu \in [\mu^L, m_L].$$

On $[x_L(m_L), \mu^L]$, both $\psi_L(\cdot; m_L)$ and u_R are affine; they coincide at $x_L(m_L)$, and $\psi_L(\mu^L; m_L) = \widetilde{u}(\mu^L) \geq u_R(\mu^L)$, so

$$\psi_L(\mu; m_L) \geq u_R(\mu) = \widetilde{u}(\mu) \qquad \text{for all } \mu \in [x_L(m_L), \mu^L].$$

The symmetric argument gives

$$\psi_H(\mu; m_H) \geq \widetilde{u}(\mu) \qquad \text{for all } \mu \in [m_H, x_H(m_H)].$$

For convexity, note that the slope of u_R is weakly below the chord slope of $\psi_L(\cdot; m_L)$ (because u_R and ψ_L meet at $x_L(m_L)$ and ψ_L lies above u_R to the right). By convexity of $\widetilde{u}$, the chord slope

of ψ_L is weakly below the one-sided slope of $\widetilde{u}$ at m_L; similarly the chord slope of ψ_H is weakly above the one-sided slope of $\widetilde{u}$ at m_H, and weakly below the slope of u_A. Since $\widetilde{u}$ is convex on (m_L, m_H), the slopes of the five pieces u_R, ψ_L, $\widetilde{u}|_{(m_L,m_H)}$, ψ_H, u_A are weakly increasing from left to right. Therefore p is convex on $[0,1]$.

Hence $p(\mu) \geq \widetilde{u}(\mu)$ for all μ, with equality on the revealed regions and at the target posteriors μ^L and μ^H. Budget balance holds block by block: on each pooled block p is affine and the design sends the block to its conditional mean, so Jensen binds; on revealed regions budget balance is immediate. Therefore the candidate in Case 3(b) satisfies domination, support, and budget balance.

Case 3(c). Single-touch pooling.

(i) *Candidate design G.* Suppose $\overline{m}^L < \overline{m}^H$ and $m_L > m_H$. The single-touch design pools $\mu \in [x_L^*, k^*]$ to posterior μ^L and pools $\mu \in [k^*, x_H^*]$ to posterior μ^H, while revealing beliefs outside $[x_L^*, x_H^*]$.

(ii) *Construct p^*.* Let p^* be the (pointwise) smallest convex function that dominates $\widetilde{u}$ and coincides with the affine wings u_R on the left and u_A on the right. In Case 3(c) the optimal majorant satisfies $p^*(\mu) > \widetilde{u}(\mu)$ for all $\mu \in (\mu^L, \mu^H)$. Lemma B.2 shows that p^* has a unique kink $k^* \in (\mu^L, \mu^H)$ and that the pooling cutoffs x_L^*, x_H^* are determined by Bayes plausibility.

Lemma B.2 (Single-touch pivot and cutoffs)**.** *Assume $\overline{m}^L < \overline{m}^H$ and $m_L > m_H$. Let p^* be the optimal convex majorant in the Verification Tool. Then:*

(i) (Existence of a pivot.) *$p^*(\mu) > \widetilde{u}(\mu)$ for all $\mu \in (\mu^L, \mu^H)$ and p^* has a kink at some $k^* \in (\mu^L, \mu^H)$. Moreover, p^* can have at most one such kink on (μ^L, μ^H).*

(ii) (Cutoffs.) *Given k^*, there exist unique cutoffs $x_L^* \in [0, \mu^L)$ and $x_H^* \in (\mu^H, 1]$ such that*

$$\mathbb{E}[\mu \mid \mu \in [x_L^*, k^*]] = \mu^L, \qquad \mathbb{E}[\mu \mid \mu \in [k^*, x_H^*]] = \mu^H.$$

Consequently, pooling $\mu \in [x_L^*, k^*]$ to posterior μ^L and pooling $\mu \in [k^*, x_H^*]$ to posterior μ^H (revealing outside) yields an MPC design that satisfies the conditions of the Verification Tool.

Proof. (i) In Case 3(c), residual interior disclosure is ruled out by $m_L > m_H$, so an optimal verified design cannot contain a nondegenerate revealed interval inside (μ^L, μ^H). Hence

$$p^*(\mu) > \widetilde{u}(\mu) \qquad \text{for all } \mu \in (\mu^L, \mu^H).$$

On any connected component of $\{p^* > \widetilde{u}\}$, the pointwise-minimal convex majorant must be affine: otherwise one could replace p^* on a strict subinterval by the chord between two interior points and obtain a lower convex majorant, contradicting minimality. Thus p^* is piecewise affine on (μ^L, μ^H).

If p^* were affine on all of (μ^L, μ^H), then the support condition would force the entire central bridge block $[m_2', m_3']$ to be assigned only to the gate posteriors $\{\mu^L, \mu^H\}$. That is exactly the bridge-

pooling geometry of Case 3(a). But bridge pooling is infeasible when $\overline{m}^L < \overline{m}^H$. Therefore p^* cannot be affine on the whole interval, so it must have at least one kink $k^* \in (\mu^L, \mu^H)$.

To prove that there can be at most one kink, suppose instead that p^* has two distinct kinks $k_1 < k_2$ in (μ^L, μ^H). Because $p^*(\mu) > \widetilde{u}(\mu)$ for all $\mu \in (\mu^L, \mu^H)$ and $\widetilde{u}$ is continuous, choose $\delta > 0$ small enough that

$$[k_1 - \delta,\ k_2 + \delta] \subset (\mu^L, \mu^H)$$

and

$$\eta := \min_{\mu \in [k_1 - \delta,\, k_2 + \delta]} \big(p^*(\mu) - \widetilde{u}(\mu)\big) > 0.$$

Let ℓ_-, ℓ_0, and ℓ_+ denote the affine pieces of p^* on $[k_1 - \delta, k_1]$, $[k_1, k_2]$, and $[k_2, k_2 + \delta]$, respectively. Since k_1 and k_2 are kinks of a convex function, their slopes satisfy

$$\text{slope}(\ell_-) < \text{slope}(\ell_0) < \text{slope}(\ell_+).$$

Fix any $\varepsilon \in (0, \eta)$. Define a new function $\tilde{p}$ that agrees with p^* outside $[k_1 - \delta, k_2 + \delta]$ and is the unique continuous piecewise-affine function on $[k_1 - \delta, k_2 + \delta]$ that agrees with p^* at $k_1 - \delta$ and $k_2 + \delta$ and satisfies

$$\tilde{p}(k_1) = p^*(k_1) - \varepsilon, \qquad \tilde{p}(k_2) = p^*(k_2) - \varepsilon.$$

For ε small enough, the three resulting affine pieces still have weakly increasing slopes, so $\tilde{p}$ is convex. Because the graph is moved downward by at most $\varepsilon < \eta$ inside a region where the slack above $\widetilde{u}$ is at least η,

$$\tilde{p}(\mu) \geq \widetilde{u}(\mu) \qquad \text{for all } \mu \in [k_1 - \delta, k_2 + \delta].$$

Thus $\tilde{p}$ is a feasible convex majorant of $\widetilde{u}$ with $\tilde{p} < p^*$ on a set of positive F-measure. Since $f > 0$,

$$\int_0^1 \tilde{p}(\mu)\, dF(\mu) < \int_0^1 p^*(\mu)\, dF(\mu),$$

contradicting dual minimality. Hence p^* has exactly one kink.

(ii) Any admissible single-touch design with pivot k must send the left pooled block to μ^L and the right pooled block to μ^H. By definition of the extremal feasibility thresholds, the left target is attainable iff $k \leq m_L$, while the right target is attainable iff $k \geq m_H$. Because Case 3(c) assumes $m_H < m_L$, feasibility requires

$$k \in [m_H, m_L].$$

The optimal pivot k^* must therefore lie in this interval.

Fix $k^* \in [m_H, m_L]$. Under $f > 0$, the truncated-mean map

$$x \mapsto \mathbb{E}[\mu \mid \mu \in [x, k^*]]$$

is continuous and strictly increasing on $[0, k^*)$, with limit k^* as $x \uparrow k^*$; similarly,

$$x \mapsto \mathbb{E}[\mu \mid \mu \in [k^*, x]]$$

is continuous and strictly increasing on $(k^*, 1]$, with limit k^* as $x \downarrow k^*$. Since $k^* \leq m_L$ makes the left target attainable and $k^* \geq m_H$ makes the right target attainable, there exist unique cutoffs $x_L^* \in [0, \mu^L)$ and $x_H^* \in (\mu^H, 1]$ such that

$$\mathbb{E}[\mu \mid \mu \in [x_L^*, k^*]] = \mu^L, \qquad \mathbb{E}[\mu \mid \mu \in [k^*, x_H^*]] = \mu^H.$$

With these cutoffs, Bayes plausibility holds block by block. Because p^* is affine on each pooled block, Jensen binds on each block, so budget balance follows from the Verification Tool. □

Verification of Case 3(c). Let ψ_L^* and ψ_H^* denote the two affine pieces of p^* on $[x_L^*, k^*]$ and $[k^*, x_H^*]$, respectively. By construction, p^* equals $\widetilde{u}$ on the revealed tails, is affine on each pooled block, and strictly dominates $\widetilde{u}$ throughout the interior valley (μ^L, μ^H).

Therefore the support condition implies that beliefs in the two pooled blocks cannot remain in the interior valley: they must be sent to the gate posteriors μ^L and μ^H. The cutoff equations

$$\mathbb{E}[\mu \mid \mu \in [x_L^*, k^*]] = \mu^L, \qquad \mathbb{E}[\mu \mid \mu \in [k^*, x_H^*]] = \mu^H$$

then give Bayes plausibility block by block.

Budget balance also holds block by block. Because p^* is affine on $[x_L^*, k^*]$,

$$\int_{x_L^*}^{k^*} p^*(\mu)\, dF(\mu) = F([x_L^*, k^*])\, p^*(\mu^L),$$

and similarly, because p^* is affine on $[k^*, x_H^*]$,

$$\int_{k^*}^{x_H^*} p^*(\mu)\, dF(\mu) = F([k^*, x_H^*])\, p^*(\mu^H).$$

On the revealed tails, budget balance is immediate because the design leaves beliefs unchanged. Hence the candidate single-touch design satisfies the three conditions of the Verification Tool.

Boundary case $m_L = m_H$. Let $\bar{m} := m_L = m_H$. In this case the residual interval from Case 3(b) collapses to a common boundary point, so the left and right gate-targeting blocks meet at $\bar{m}$. Define the candidate design by pooling $\mu \in [x_L(\bar{m}), \bar{m}]$ to posterior μ^L and pooling $\mu \in [\bar{m}, x_H(\bar{m})]$ to

posterior μ^H, revealing outside these two blocks. Define

$$p(\mu) = \begin{cases} u_R(\mu), & \mu < x_L(\bar{m}), \\ \psi_L(\mu; \bar{m}), & \mu \in [x_L(\bar{m}), \bar{m}], \\ \psi_H(\mu; \bar{m}), & \mu \in [\bar{m}, x_H(\bar{m})], \\ u_A(\mu), & \mu > x_H(\bar{m}). \end{cases}$$

By definition of $\bar{m}$, the left block has conditional mean μ^L and the right block has conditional mean μ^H, so Bayes plausibility holds block by block. The domination and convexity arguments are the same as in the strict case, except that the two affine pieces now meet at the common boundary point $\bar{m}$ rather than being separated by a revealed interval. Hence the candidate satisfies domination, support, convexity, and budget balance. This verifies the common-boundary-point case stated in Proposition B.1.

Conclusion of the proof. In each case above, the constructed price function is a convex majorant of $\widetilde{u}$, the induced design is supported on the contact set, and budget balance holds block by block. The Verification Tool therefore implies optimality of the candidate design. Under $f > 0$, the verified majorant also pins down the geometric configuration uniquely, up to payoff-equivalent refinements on affine revealed regions. Appendix B.3 addresses those implementation refinements.

B.3 Minimum-complexity implementation

This subsection justifies the complexity counts reported in Corollary 1. Appendix B.2 has already characterized the economically distinct posterior structures. The remaining issue is implementation: among payoff-equivalent optimal representations, we select one with the smallest number of recommendation categories. Outside the residual case, this selection is straightforward because any revealed affine piece can be compressed to its conditional mean without changing either Bayes plausibility or payoff, whereas strategic gate posteriors are indispensable.

Fix an optimal design G in either the single-gate or the two-sided gate-pooling class. Any remaining implementation freedom is confined to revealed regions on which $\widetilde{u}$ is affine. Let B be such a component, with induced posterior law τ_B. Since $\widetilde{u}$ is affine on B,

$$\mathbb{E}_{\mu\sim\tau_B}[\widetilde{u}(\mu)] = \widetilde{u}\big(\mathbb{E}_{\mu\sim\tau_B}[\mu]\big)\,.$$

Hence B can be replaced by a single recommendation category at its conditional mean without changing feasibility or payoff. Repeating this compression on every nondegenerate affine revealed component yields the minimum-complexity optimal implementation within the same posterior geometry.

The remaining categories are therefore exactly the indispensable gate posteriors plus one category for each surviving affine revealed component. In the single-gate case, this gives a binary or ternary implementation depending on whether one or two revealed tails survive. In the two-sided

case, both gate posteriors are indispensable, and each surviving revealed tail contributes one additional category; hence at most four categories suffice. In the residual interior-disclosure case, no analogous finite compression is available, because the remaining interior interval lies on a region where $\widetilde{u}$ is strictly convex rather than affine, so any nontrivial pooling there would strictly reduce expected payoff.

B.4 Benchmark geometric classification ($\lambda = 0$)

This subsection studies the fixed-cost benchmark ($\lambda = 0$) used in the closed-form illustration in Section V.5. Relative to the general $\lambda > 0$ case, the only new geometric feature is that the provider's payoff is flat on the review interval, so an interior pooled posterior can be optimal. That plateau case disappears once $\lambda > 0$.

Critical intersections and conditional means. The benchmark geometry is indexed by four critical belief thresholds derived by equating the provider's branch payoffs with the plateau level $-\beta K$ or the relevant line segments:

$$m_1 = \frac{\beta K}{\beta K + (1-\mu^L)\, C_{FN}},\ m_2 = \frac{\beta K}{C_{FN}},\ m_3 = 1 - \frac{\beta K}{C_{FP}},\ m_4 = \frac{C_{FP}\,\mu^H}{C_{FP}\,\mu^H + \beta K} \tag{18}$$

where $\mu^L = \frac{K}{\alpha D_{FN}}$ and $\mu^H = 1 - \frac{K}{\alpha D_{FP}}$. Define the associated conditional means:

$$\theta_1 := \mathbb{E}[\mu \mid \mu \geq m_1], \qquad \theta_{23} := \mathbb{E}[\mu \mid \mu \in [m_2, m_3]], \qquad \theta_4 := \mathbb{E}[\mu \mid \mu \leq m_4]. \tag{19}$$

Proposition B.2 below records the five benchmark configurations indexed by these objects, and Figure B.3 provides the corresponding visual guide.

Proposition B.2 (Geometric classification (benchmark $\lambda = 0$))**.** *Let $\theta_1, \theta_{23}, \theta_4$ be the conditional means defined above. The optimal recommendation policy takes one of the following five mutually exclusive and exhaustive forms:*

1. ***Tail pooling toward the left gate*** *($\theta_1 \leq \mu^L$): pool $\mu \in [x_1, 1]$ to posterior μ^L and reveal μ on $[0, x_1)$, where $x_1 \in (m_1, 1)$ solves $\mathbb{E}[\mu \mid \mu \geq x_1] = \mu^L$.*
2. ***Cross-region pooling toward the left gate*** *($\theta_1 > \mu^L > \theta_{23}$): pool $\mu \in [x_2, x_3]$ to μ^L and reveal outside, where $x_2 \in (m_1, \mu^L)$ and x_3 satisfy collinearity and Bayes plausibility.*
3. ***Central plateau pooling*** *($\mu^H \geq \theta_{23} \geq \mu^L$): pool $\mu \in [m_2, m_3]$ to posterior θ_{23} and reveal outside; the pooled posterior lies on the flat region.*
4. ***Cross-region pooling toward the right gate*** *($\theta_{23} > \mu^H > \theta_4$): pool $\mu \in [x_5, x_4]$ to μ^H and reveal outside, where $x_4 \in (\mu^H, 1)$ and x_5 satisfy collinearity and Bayes plausibility.*
5. ***Tail pooling toward the right gate*** *($\theta_4 \geq \mu^H$): pool $\mu \in [0, x_6]$ to posterior μ^H and reveal μ on $(x_6, 1]$, where $x_6 \in (0, m_4)$ solves $\mathbb{E}[\mu \mid \mu \leq x_6] = \mu^H$.*

Figure B.3 records the five benchmark configurations that support Proposition B.2.

Proof of Proposition B.2. The benchmark proof parallels Appendix B.2. Cases 1, 2, 4, and 5 are direct

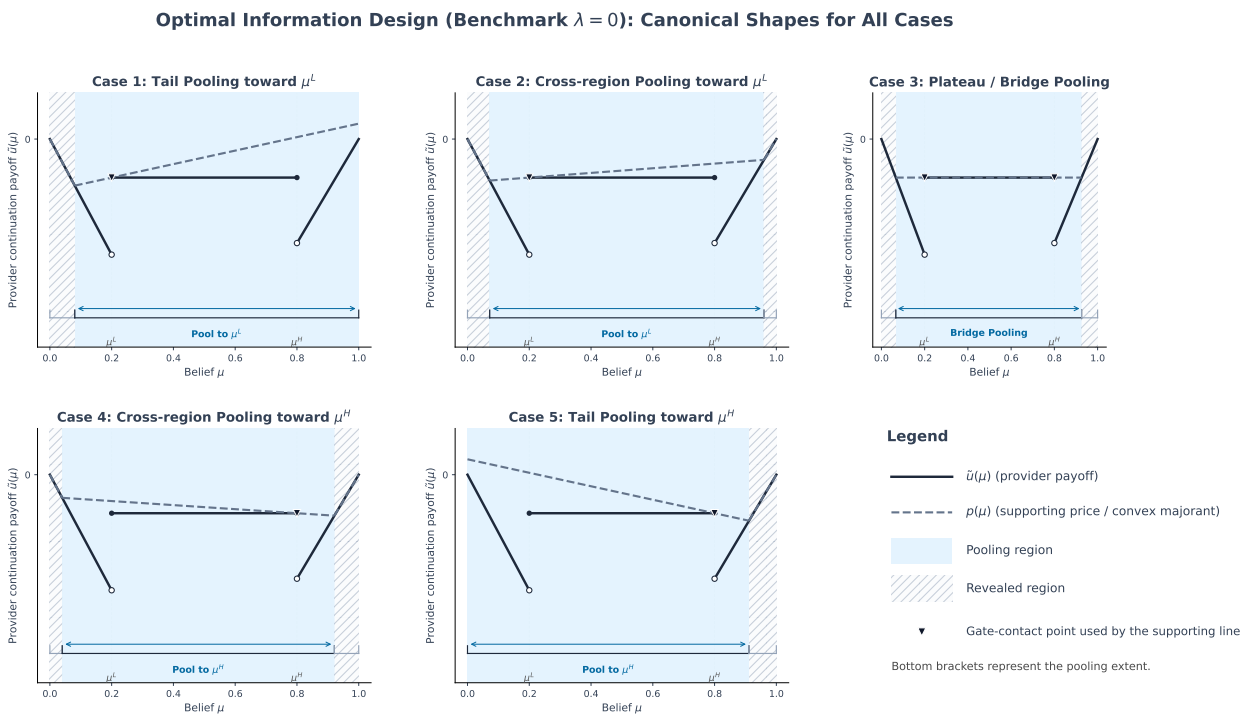


Figure B.3: Five benchmark configurations in the fixed-cost case $(\lambda = 0)$**.** Shaded regions indicate pooled intervals and unshaded regions indicate revealed intervals. Cases 1 and 5: single-gate tail pooling. Cases 2 and 4: single-gate cross-region pooling. Case 3: plateau pooling, in which the provider payoff is flat on the review interval and an interior pooled posterior is optimal. The figure illustrates Proposition B.2 and the closed-form benchmark in Appendix C.7.

fixed-cost analogues of the corresponding one-gate constructions: the stated cutoffs make the pooled block Bayes-plausible, the supporting price is affine on the pooled block and coincides with $\widetilde{u}$ on the revealed region, and Jensen binds block by block. The proof therefore summarizes those cases and focuses on the only genuinely new benchmark configuration.

In Case 3, when $\mu^L \le \theta_{23} \le \mu^H$, the provider payoff is constant on $[\mu^L, \mu^H]$, equal to $-\beta K$. Define

$$p(\mu) = \begin{cases} u_R(\mu), & \mu < m_2, \\ -\beta K, & \mu \in [m_2, m_3], \\ u_A(\mu), & \mu > m_3. \end{cases}$$

By construction, p is convex. It coincides with $\widetilde{u}$ on the revealed tails and equals the plateau value $-\beta K$ on $[m_2, m_3]$, so $p(\mu) \ge \widetilde{u}(\mu)$ for all $\mu \in [0,1]$. Pooling $\mu \in [m_2, m_3]$ to the posterior

$$\theta_{23} = \mathbb{E}[\mu \mid \mu \in [m_2, m_3]]$$

is Bayes-plausible and supported on the contact set $\{p = \widetilde{u}\}$. Budget balance on the pooled block is immediate because p is constant there:

$$\int_{m_2}^{m_3} p(\mu)\, dF(\mu) = F([m_2, m_3])\, p(\theta_{23}).$$

Hence plateau pooling is optimal in Case 3.

Mutual exclusivity and exhaustiveness follow because the relative positions of $\theta_1, \theta_{23}, \theta_4$ with respect to the gate interval $[\mu^L, \mu^H]$ partition the benchmark geometry into the five listed cases. Under $f > 0$, the supporting price is unique in each case, so the classification is exhaustive. □

C Welfare proofs and benchmark support

This section supports Section V. Appendix C.1 and Appendix C.3 establish the two endpoint benchmarks. Appendix C.4 and Appendix C.5 prove the accounting identity and the distortion characterization. Appendix C.6 proves the local review-trap result, and Appendix C.2 proves the insufficient-review negative-synergy result. The final two subsections record the fixed-cost illustration.

C.1 Proof of Proposition 3

Proof. If $\alpha < \alpha_{\min}(K)$, then by definition $\Delta^{\max}(\alpha) < K$. Hence

$$\Delta(\mu; \alpha, \lambda) < K \qquad \text{for all } \mu \in [0,1].$$

By Proposition 1, the DM never reviews in equilibrium.

Therefore, after any posterior μ, the DM acts immediately and

$$V(\mu; \alpha) = -\alpha \ell(\mu).$$

Substituting this into the reduced-form transfer rule (2), the provider's objective becomes

$$-(1-\alpha)\mathbb{E}_G[\ell(\mu)] + \beta\Big(\mathbb{E}_G[-\alpha\ell(\mu)] - V(\mu_0;\alpha)\Big) = -\gamma(\alpha)\mathbb{E}_G[\ell(\mu)] - \beta V(\mu_0;\alpha),$$

where $\gamma(\alpha) = (1-\alpha) + \beta\alpha$.

Since $\ell(\mu)$ is concave, the function $-\gamma(\alpha)\ell(\mu)$ is convex. Any feasible posterior distribution $G \in \mathrm{MPC}(F)$ is a mean-preserving contraction of F. By Jensen's inequality, the expectation of a convex function is maximized by the most dispersed feasible distribution, namely F itself. Hence the provider's optimal design is full revelation:

$$G^\star = F.$$

Finally, because review never occurs, the post-review posterior equals the posterior induced by disclosure, so under full revelation $\nu = \omega$ and

$$W(\alpha, K) = -\mathbb{E}_F[\ell(\omega)] = W^{\text{algorithm}}.$$

This establishes the claim. □

C.2 Proof of Corollary 3

Proof. By Proposition 3, if $\alpha < \alpha_{\min}(K)$, then

$$W(\alpha, K) = W^{\text{algorithm}}.$$

Next, under full decision-maker liability and no algorithmic input,

$$W^{\text{human}}(K) = V(\mu_0; 1).$$

By Lemma 1,

$$V(\mu_0;1) + \lambda\phi(\mu_0) = \max\left\{N(\mu_0;1),\ (\operatorname{cav} N)(\mu_0;1) - K\right\},$$

where the objects inside the maximum are independent of K except for the linear term $-K$. Hence $W^{\text{human}}(K)$ is continuous in K.

Since $W^{\text{human}}(0) > W^{\text{algorithm}}$, continuity implies that there exists $\bar{K} > 0$ such that

$$W^{\text{human}}(K) > W^{\text{algorithm}} \qquad \text{for all } K \in (0, \bar{K}).$$

Therefore, for every such K and every $\alpha < \alpha_{\min}(K)$,

$$W(\alpha, K) = W^{\text{algorithm}} < W^{\text{human}}(K),$$

which implies

$$\text{Synergy}^{\text{deploy}}(\alpha, K) < 0.$$

□

C.3 Proof of Proposition 4

Proof. Fix any posterior belief $\mu \in [0,1]$. Define the *full-liability continuation welfare* (the planner's value when the DM bears all losses) at belief μ as

$$V^{FB}(\mu) := \max_{\tau:\, \mathbb{E}_{\nu\sim\tau}[\nu]=\mu} \Big(\mathbb{E}_{\nu\sim\tau}[-\ell(\nu)] - \mathcal{C}(\tau;\mu)\Big).$$

By construction, this is exactly the DM's continuation value under full liability: $V^{FB}(\mu) = V(\mu;1)$.

Step 1 (pointwise upper bound). Fix any liability α and any implementable recommendation policy G. Let $\tau^*(\mu;\alpha)$ be the DM's equilibrium continuation strategy after observing μ. For each realized μ, the induced continuation welfare equals

$$w(\mu;\alpha) := \mathbb{E}_{\nu\sim\tau^*(\mu;\alpha)}[-\ell(\nu)] - \mathcal{C}(\tau^*(\mu;\alpha);\mu).$$

Since $\tau^*(\mu;\alpha)$ is feasible in the definition of $V^{FB}(\mu)$, it follows that $w(\mu;\alpha) \le V^{FB}(\mu) = V(\mu;1)$ for every μ. Taking expectations over $\mu \sim G$ gives

$$W(G;\alpha) = \mathbb{E}_G[w(\mu;\alpha)] \le \mathbb{E}_G[V(\mu;1)].$$

Therefore,

$$W(G^\star(\alpha,K);\alpha) \le \sup_{G\in \mathrm{MPC}(F)} \mathbb{E}_G[V(\mu;1)]. \tag{20}$$

Step 2 (full liability attains the upper bound). When $\alpha = 1$, the provider bears no misclassification loss, hence her payoff is

$$U_D(G;1) = T(G) = \beta\big(\mathbb{E}_G[V(\mu;1)] - V(\mu_0;1)\big).$$

Because $\beta > 0$, maximizing $U_D(G;1)$ is equivalent to maximizing $\mathbb{E}_G[V(\mu;1)]$ over $G \in \mathrm{MPC}(F)$. Moreover, under $\alpha = 1$ there is no residual misclassification externality, so $W(G;1) = \mathbb{E}_G[V(\mu;1)]$. Hence, any equilibrium choice $G^\star(1,K)$ satisfies

$$W(G^\star(1,K);1) = \mathbb{E}_{G^\star(1,K)}[V(\mu;1)] = \sup_{G\in \mathrm{MPC}(F)} \mathbb{E}_G[V(\mu;1)].$$

Step 3 (conclude). Combining this equality with the upper bound (20) yields $W(G^\star(1,K);1) \ge W(G^\star(\alpha,K);\alpha)$ for all $\alpha \in [0,1]$. □

C.4 Two-margin decomposition

Proof of Lemma 3. **Identity.** An implementable recommendation policy $\Gamma(\cdot \mid \omega)$ together with the score distribution F induces a joint law $H_\Gamma(d\omega, d\mu) = F(d\omega)\,\Gamma(d\mu \mid \omega)$, which disintegrates as

$H_\Gamma = Q_\Gamma(d\omega \mid \mu)\, G_\Gamma(d\mu)$, where G_Γ is the marginal posterior distribution and $Q_\Gamma(\cdot \mid \mu)$ is the conditional distribution of underlying scores pooled into posterior μ. Define $\bar{\ell}_\Gamma(\mu) := \int \ell(\omega)\, Q_\Gamma(d\omega \mid \mu)$.

By definition (6), $W(\Gamma;\alpha) = -\mathbb{E}_{G_\Gamma}[L(\mu;\alpha)]$, where $L(\mu;\alpha) = \mathbb{E}_{\nu\sim\tau^*}[\ell(\nu)] + \mathcal{C}(\tau^*;\mu)$ is the social continuation loss at posterior μ. The algorithm-only benchmark is $W^{\text{algorithm}} = -\mathbb{E}_F[\ell(\omega)]$. By iterated expectations under the joint law H_Γ, $\mathbb{E}_{G_\Gamma}[\bar{\ell}_\Gamma(\mu)] = \mathbb{E}_F[\ell(\omega)]$. Therefore:

$$\begin{aligned} W(\Gamma;\alpha) - W^{\text{algorithm}} &= -\mathbb{E}_{G_\Gamma}[L(\mu;\alpha)] + \mathbb{E}_{G_\Gamma}[\bar{\ell}_\Gamma(\mu)] \\ &= \mathbb{E}_{G_\Gamma}[\bar{\ell}_\Gamma(\mu) - L(\mu;\alpha)] \\ &= \mathbb{E}_{G_\Gamma}\big[\big(\ell(\mu) - L(\mu;\alpha)\big) - \big(\ell(\mu) - \bar{\ell}_\Gamma(\mu)\big)\big] \\ &= \mathbb{E}_{G_\Gamma}[B(\mu;\alpha)] - \mathbb{E}_{G_\Gamma}[D_\Gamma(\mu)]. \end{aligned}$$

Part (i): $B(\mu;\alpha) \geq 0$**.** If μ lies outside the review region, no review occurs and $L(\mu;\alpha) = \ell(\mu)$, so $B = 0$. If μ lies inside the review region, the DM reviews. Since the DM could have chosen no review (incurring payoff $-\alpha\ell(\mu)$), private optimality requires $-\alpha\mathbb{E}[\ell(\nu)] - \mathcal{C} \geq -\alpha\ell(\mu)$, i.e., $\alpha(\ell(\mu) - \mathbb{E}[\ell(\nu)]) \geq \mathcal{C}$. Therefore

$$B(\mu;\alpha) = \ell(\mu) - \mathbb{E}[\ell(\nu)] - \mathcal{C} \geq (1-\alpha)\big(\ell(\mu) - \mathbb{E}[\ell(\nu)]\big) \geq 0,$$

where the last inequality uses $\ell(\mu) \geq \mathbb{E}[\ell(\nu)]$ (since ℓ is concave and the post-review posteriors are a mean-preserving spread of μ restricted to stopping points).

Part (ii): $D_\Gamma(\mu) \geq 0$**.** Since ℓ is concave and $\int \omega\, Q_\Gamma(d\omega \mid \mu) = \mu$ by Bayes' rule, Jensen's inequality gives $\bar{\ell}_\Gamma(\mu) \leq \ell(\mu)$, so $D_\Gamma(\mu) \geq 0$. Equality holds if and only if ℓ is affine on the support of $Q_\Gamma(\cdot \mid \mu)$. □

C.5 Cross-cutoff distortion characterization

This subsection proves the distortion test used in the text: pooling is harmless within each side of the myopic cutoff and harmful exactly when a recommendation category mixes scores from opposite sides of that cutoff.

Proof of Corollary 2. **Part (i).** On each side of $\overline{\mu}$, the myopic loss function ℓ is affine. Therefore, if $Q_\Gamma(\cdot \mid \mu)$ places support only on one side of $\overline{\mu}$, then

$$\bar{\ell}_\Gamma(\mu) = \ell\left(\int \omega\, Q_\Gamma(d\omega \mid \mu)\right) = \ell(\mu),$$

so $D_\Gamma(\mu) = 0$. The welfare inequality then follows from (8) and $B(\mu;\alpha) \geq 0$.

Part (ii). The function ℓ is affine on $[0,\overline{\mu}]$ and on $[\overline{\mu},1]$ but has a kink at $\overline{\mu}$. Hence it is not affine on any interval that crosses $\overline{\mu}$. If $Q_\Gamma(\cdot \mid \mu)$ puts positive mass on both sides of $\overline{\mu}$, Jensen's inequality is strict, so $\bar{\ell}_\Gamma(\mu) < \ell(\mu)$ and therefore $D_\Gamma(\mu) > 0$. □

Lemma C.1 (Exact distortion formula). *Let $\overline{\mu} = D_{FP}/(D_{FP} + D_{FN})$. For any implementable recommendation policy Γ,*

$$D_\Gamma(\mu) = (D_{FP} + D_{FN}) \times \begin{cases} \mathbb{E}_{Q_\Gamma}[(\omega - \overline{\mu})_+ \mid \mu], & \mu \le \overline{\mu}, \\ \mathbb{E}_{Q_\Gamma}[(\overline{\mu} - \omega)_+ \mid \mu], & \mu \ge \overline{\mu}. \end{cases}$$

In particular, $D_\Gamma(\mu) > 0$ if and only if $Q_\Gamma(\cdot \mid \mu)$ puts positive mass on both sides of $\overline{\mu}$.

Proof. For $\mu \le \overline{\mu}$, $\ell(\mu) = \mu D_{FN}$, and $\bar{\ell}_\Gamma(\mu) = \int [\omega D_{FN} \mathbf{1}_{\omega \le \overline{\mu}} + (1-\omega) D_{FP} \mathbf{1}_{\omega > \overline{\mu}}]\, Q_\Gamma(d\omega \mid \mu)$. Then $D_\Gamma(\mu) = \mu D_{FN} - \bar{\ell}_\Gamma(\mu)$. Using $\mu = \int \omega\, Q_\Gamma(d\omega \mid \mu)$,

$$\begin{aligned} D_\Gamma(\mu) &= \int [\omega D_{FN} - \ell(\omega)]\, Q_\Gamma(d\omega \mid \mu) \\ &= \int_{\omega > \overline{\mu}} [\omega D_{FN} - (1-\omega) D_{FP}]\, Q_\Gamma(d\omega \mid \mu) \\ &= (D_{FP} + D_{FN}) \int (\omega - \overline{\mu})_+\, Q_\Gamma(d\omega \mid \mu). \end{aligned}$$

The case $\mu \ge \overline{\mu}$ is symmetric. □

C.6 Local review-trap proof and branch regularity

Branch-regularity conditions at activation. Fix K and suppose there exists $\bar{\alpha} > \alpha_{\min}(K)$ and a selected equilibrium branch $\Gamma^\star(\alpha, K)$ on $(\alpha_{\min}(K), \bar{\alpha})$ such that:

(a) the branch admits a weak-* right limit (G^+, Q^+) as $\alpha \downarrow \alpha_{\min}(K)$;

(b) equilibrium welfare is right-continuous at $\alpha_{\min}(K)$ along that branch.

These are local selection conditions at the activation boundary, not new economic restrictions. They rule out discontinuous switching across equilibrium branches exactly when review first becomes inducible.

Under these conditions, let $B^+(\mu) := \ell(\mu) - L^+(\mu)$ and $D^+(\mu) := \ell(\mu) - \bar{\ell}^+(\mu)$, where $L^+(\mu)$ and $\bar{\ell}^+(\mu) := \int \ell(\omega)\, Q^+(d\omega \mid \mu)$ are evaluated under the post-activation limit design and G^+ denotes the induced posterior distribution. The local review trap occurs when the distortion term dominates the review value at that limit:

$$\mathbb{E}_{G^+}\left[B^+(\mu)\right] < \mathbb{E}_{G^+}\left[D^+(\mu)\right], \tag{21}$$

which is equivalent to the three-margin representation

$$\underbrace{\mathbb{E}_{G^+}[\ell(\mu) - \ell(\nu)]}_{\text{correction value}} < \underbrace{\mathbb{E}_{G^+}[\mathcal{C}(\tau;\mu)]}_{\text{review cost}} + \underbrace{\Big(\mathbb{E}_{G^+}[\ell(\mu)] - \mathbb{E}_F[\ell(\omega)]\Big)}_{\text{coarsening distortion}}. \tag{22}$$

Proof of Proposition 5. **Step 1 (negative gap at the post-activation limit).** By Lemma 3 applied to

the limit design (G^+, Q^+),

$$W(G^+; \alpha_{\min}(K)) - W^{\text{algorithm}} = \mathbb{E}_{G^+}[B^+(\mu)] - \mathbb{E}_{G^+}[D^+(\mu)].$$

The activation-loss inequality (21) states that the right-hand side is strictly negative, so

$$W(G^+; \alpha_{\min}(K)) < W^{\text{algorithm}}.$$

Step 2 (extending to an interval via right continuity). By assumption (b), $W(\alpha, K) \to W(G^+; \alpha_{\min}(K))$ as $\alpha \downarrow \alpha_{\min}(K)$. Since the right limit is strictly below $W^{\text{algorithm}}$, there exists $\varepsilon > 0$ such that

$$W(\alpha, K) < W^{\text{algorithm}} \qquad \text{for all } \alpha \in (\alpha_{\min}(K), \alpha_{\min}(K) + \varepsilon).$$

This is a nonempty review-trap interval immediately above activation. □

Non-monotonicity implication. If $\beta > 0$ and $\Delta^{\max}(1) > K$, Proposition 4 implies $W(1, K) > W^{\text{algorithm}}$: because $\beta > 0$ the equilibrium design at $\alpha = 1$ maximizes $\mathbb{E}_G[V(\mu; 1)]$, and since $V(\mu; 1) \geq -\ell(\mu)$ with strict inequality on the positive-F-measure review set, $W(1, K) > \mathbb{E}_F[-\ell(\mu)] = W^{\text{algorithm}}$. Combined with the local-trap result, welfare is locally below the algorithm-only benchmark near activation but strictly above it at full decision-maker liability, so welfare is non-monotone in α and the review-trap set is nonempty.

C.7 Uniform benchmark: review traps and non-monotonicity

This subsection derives the closed-form illustration behind Figure 5. It adds no new mechanism; it makes the activation comparison from Proposition 5 explicit in the uniform fixed-cost benchmark.

In the uniform benchmark with $F(\mu) = \mu$ and $\lambda = 0$, two startup-cost thresholds organize the welfare comparison at activation. Suppose there exists $\bar{\alpha} > \alpha_{\min}(K)$ such that, for every $\alpha \in (\alpha_{\min}(K), \bar{\alpha})$, the equilibrium design near activation takes the single-gate tail-pooling form in which a lower tail of scores is pooled to the upper review gate μ^H.[14] Let

$$W(\alpha^+_{\min}, K) := \lim_{\alpha \downarrow \alpha_{\min}} W(\alpha, K).$$

Then $W(\alpha^+_{\min}, K) > W^{\text{algorithm}}$ for low enough startup costs, whereas $W(\alpha^+_{\min}, K) < W^{\text{algorithm}}$ for an intermediate range of startup costs; Appendix C.8 records the corresponding cutoff formulas. Moreover, $W(1, K) > W^{\text{algorithm}}$ for all K in the relevant range. Hence, in that intermediate startup-cost region, welfare is non-monotone in α and a review trap arises for some intermediate liability level.

[14] This corresponds to Case 5 of Proposition B.2. Appendix C.8 reports the corresponding parameter restriction.

Derivation. Let $\mathcal{L}_{\text{soc}}(\alpha, K) := -W(\alpha, K)$ denote equilibrium social loss. For $K < \overline{K}$, the activation threshold satisfies $\alpha_{\min} = \frac{K(D_{FP}+D_{FN})}{D_{FP}D_{FN}} < 1$. By Proposition 3, for every $\alpha < \alpha_{\min}$, welfare equals the algorithm-only benchmark, so $\mathcal{L}_{\text{soc}}(\alpha, K) = \mathcal{L}_{\text{soc}}^{(0)}$ where $\mathcal{L}_{\text{soc}}^{(0)} = \frac{D_{FP}D_{FN}}{2(D_{FP}+D_{FN})}$.

By the Case 5 parameter restriction reported in Appendix C.8, there exists a right-neighborhood of $\alpha_{\min}$ on which the equilibrium remains in Case 5. The closed-form expressions collected in Appendix C.8 then imply that the right-limit social loss at activation, $\mathcal{L}_{\text{soc}}^{(5)}(\alpha_{\min}^{+})$, satisfies

$$\mathcal{L}_{\text{soc}}^{(5)}(\alpha_{\min}^{+}) < \mathcal{L}_{\text{soc}}^{(0)} \quad \text{if and only if} \quad K < K^*,$$

and $\mathcal{L}_{\text{soc}}^{(5)}(\alpha_{\min}^{+}) > \mathcal{L}_{\text{soc}}^{(0)}$ if and only if $K^* < K < \overline{K}$. Since welfare is the negative of social loss, these comparisons establish the sign of welfare relative to $W^{\text{algorithm}}$ at activation.

For full liability, Appendix C.8 shows that as $\alpha \to 1$ the equilibrium enters Case 3 and social loss converges to $\mathcal{L}_{\text{soc}}^{(\infty)} = K\left(1 - \frac{\alpha_{\min}}{2}\right) < \mathcal{L}_{\text{soc}}^{(0)}$ for all $K \in (0, \overline{K})$, which gives $W(1, K) > W^{\text{algorithm}}$.

When $K^* < K < \overline{K}$, welfare is strictly below $W^{\text{algorithm}}$ at activation but strictly above it at $\alpha = 1$. Because the closed-form welfare expression is continuous in α on a right-neighborhood of $\alpha_{\min}$ within Case 5, there exists $\varepsilon > 0$ such that $W(\alpha, K) < W^{\text{algorithm}}$ for all $\alpha \in (\alpha_{\min}, \alpha_{\min} + \varepsilon)$. Hence equilibrium welfare is non-monotone in α in this parameter range.

C.8 Closed-form benchmark objects for Figure 5

This subsection collects the closed-form objects used in Appendix C.7 and Figure 5. The equilibrium review incidence and welfare objects are organized case by case so that the benchmark comparisons can be verified without redoing the derivation. Conditional on review, the DM learns θ perfectly and $V(\mu; \alpha) = -\min\{\alpha\ell(\mu), K\}$. Throughout, K is fixed and suppressed in some expressions when no confusion can arise.

Objects and welfare levels. Let $\mathcal{S}(\alpha, K) := [\mu^L(\alpha, K), \mu^H(\alpha, K)]$ denote the equilibrium review region, let $\mathcal{P}(\alpha, K)$ denote the equilibrium review incidence, and let $L_{\text{out}}(\alpha, K) := \mathbb{E}_{G^*(\alpha,K)}[\ell(\mu)\mathbf{1}\{\mu \notin \mathcal{S}(\alpha, K)\}]$ be the expected misclassification loss on posteriors outside the equilibrium review region. Since review reveals θ perfectly when $\lambda = 0$, the welfare levels are

$$\begin{aligned}
W(\alpha, K) &= -K\,\mathcal{P}(\alpha, K) - L_{\text{out}}(\alpha, K), \\
U_{DM}(\alpha, K) &= -(1-\beta)\big(K\,\mathcal{P}(\alpha, K) + \alpha\, L_{\text{out}}(\alpha, K)\big) + \beta V(\mu_0; \alpha), \\
U_D(\alpha, K) &= \beta\big(-K\,\mathcal{P}(\alpha, K) - \alpha\, L_{\text{out}}(\alpha, K) - V(\mu_0; \alpha)\big) - (1-\alpha)L_{\text{out}}(\alpha, K),
\end{aligned}$$

with $W = U_{DM} + U_D$ (transfers are welfare-neutral). In the figures of Section V.5, the prior mean μ_0 lies outside the equilibrium review region for the plotted parameter range, so $V(\mu_0; \alpha) = -\alpha\ell(\mu_0)$ and the expressions simplify accordingly.

Closed forms for $\mathcal{P}(\alpha)$ and $W(\alpha)$ under the uniform prior. Let $F(\mu) = \mu$, $\overline{\mu} = \frac{D_{FP}}{D_{FP}+D_{FN}}$. The loss integral over any interval $[a, b] \subseteq [0, 1]$ is

$$\mathcal{L}(a,b) = \int_a^b \ell(\mu)\, d\mu = \begin{cases} \frac{D_{FN}}{2}(b^2 - a^2), & b \le \overline{\mu}, \\ D_{FP}\Big((b-a) - \frac{1}{2}(b^2 - a^2)\Big), & a \ge \overline{\mu}, \\ \frac{D_{FN}}{2}(\overline{\mu}^2 - a^2) + D_{FP}\Big((b-\overline{\mu}) - \frac{1}{2}(b^2 - \overline{\mu}^2)\Big), & a < \overline{\mu} < b. \end{cases}$$

Case 5 (left-tail pooling toward the right gate). The pooled region is $[0, x_6]$ with $x_6 = 2\mu^H(\alpha) = \mathcal{P}(\alpha)$. Under Case 5, $\mu^H \le 1/2$ (so $x_6 \le 1$). Review occurs on $[0, x_6]$, yielding $\mathcal{L}^{(5)}_{\text{soc}}(\alpha) = K \cdot \mathcal{P}(\alpha) + \mathcal{L}(x_6, 1)$.

Case 4 (cross-region pooling toward the right gate). The pooling interval is $[x_5, x_4]$ with $x_4 = \mu^H + \frac{1}{2}\mathcal{P}(\alpha)$ and $x_5 = \mu^H - \frac{1}{2}\mathcal{P}(\alpha)$, where $\mathcal{P}(\alpha) = \frac{2}{D_{FP}+D_{FN}}\big((1-\mu^H)D_{FP} + \mu^H D_{FN} - \frac{2\beta K}{\gamma(\alpha)}\big)$. Let $\underline{x} := \min\{x_5, \mu^L\}$; then $\mathcal{L}^{(4)}_{\text{soc}}(\alpha) = K\,(x_4 - \underline{x}) + \mathcal{L}(0, \underline{x}) + \mathcal{L}(x_4, 1)$.

Case 3 (Central Pooling). The pooled region is $[m_2, m_3]$ with $m_2 = \frac{\beta K}{\gamma D_{FN}}$, $m_3 = 1 - \frac{\beta K}{\gamma D_{FP}}$, $\gamma(\alpha) = 1 - \alpha(1-\beta)$. Review occurs on $[m_2, m_3]$: $\mathcal{L}^{(3)}_{\text{soc}}(\alpha) = K \cdot \mathcal{P}(\alpha) + \mathcal{L}(0, m_2) + \mathcal{L}(m_3, 1)$, where $\mathcal{P}(\alpha) = m_3 - m_2$ is strictly decreasing in α.

Activation comparison in the uniform benchmark. Define $\alpha_{\min} := \frac{K(D_{FP}+D_{FN})}{D_{FP}D_{FN}}$, the inactive social loss $\mathcal{L}^{(0)}_{\text{soc}} := \frac{D_{FP}D_{FN}}{2(D_{FP}+D_{FN})}$, and the **critical activation cost** $K^* := \frac{D_{FP}(3D_{FN}-D_{FP})}{4(D_{FP}+D_{FN})}$. In the parameterization of Section V.5 ($D_{FP} = 0.4$, $D_{FN} = 1.2$, $\beta = 0.03$), the design at activation falls in Case 5 whenever $\beta \le \frac{D_{FN}-D_{FP}}{D_{FN}+D_{FP}}\big[\frac{D_{FP}D_{FN}}{K(D_{FP}+D_{FN})} - 1\big]$, which is satisfied by the chosen parameters. Then:

- **Efficiency Jump**: If $K < K^*$, then $\mathcal{L}^{(5)}_{\text{soc}}(\alpha^+_{\min}) < \mathcal{L}^{(0)}_{\text{soc}}$ (welfare improves at activation).
- **Review Trap at Activation**: If $K^* < K < \overline{K} := \frac{D_{FP}D_{FN}}{D_{FP}+D_{FN}}$, then $\mathcal{L}^{(5)}_{\text{soc}}(\alpha^+_{\min}) > \mathcal{L}^{(0)}_{\text{soc}}$.

Recovery under full decision-maker liability. As $\alpha \to 1$, the system enters Case 3 with $m_2^{(\infty)} = K/D_{FN}$, $m_3^{(\infty)} = 1 - K/D_{FP}$, $\mathcal{P}^{(\infty)} = 1 - \alpha_{\min}$, yielding $\mathcal{L}^{(\infty)}_{\text{soc}} = K(1 - \alpha_{\min}/2) < \mathcal{L}^{(0)}_{\text{soc}}$ for all $K \in (0, \overline{K})$, establishing that full decision-maker liability attains the deployment-stage benchmark.

D Local formal support for the endogenous-training extension

This appendix supports Section VI. Conditional on any fixed effort level, the deployment game is exactly the baseline model with F replaced by F_e, so Sections II–V apply verbatim. The only new issue is the date-1 training margin.

D.1 Setup and regularity conditions

Relative to the baseline model, the only new ingredient is a date-1 investment choice that affects score informativeness. Conditional on effort, the deployment game is unchanged. At date 0, the regulator chooses $\alpha \in [0, 1]$. At date 1, the provider chooses training effort $e \in [\underline{e}, \bar{e}]$ at cost $c(e)$. Effort determines a signal structure $H_e(\cdot \mid \theta)$, which in turn induces a calibrated posterior score $\omega_e(s) := \Pr(\theta = \text{Good} \mid s)$ with distribution F_e satisfying the baseline regularity: F_e admits a continuous density f_e with full support on $[0, 1]$ and $\mathbb{E}_{F_e}[\omega] = \mu_0$. The review technology, recom-

mendation problem, and liability rule are exactly the same as in the baseline model; Sections II–V apply verbatim with F replaced by F_e.

For each fixed (e, α), let $G^*(e, \alpha)$ denote the selected equilibrium recommendation design in the date-2 game, and define

$$M(e, \alpha) := \mathbb{E}_{G^*(e,\alpha)}[V(\mu; \alpha)], \qquad L(e, \alpha) := \mathbb{E}_{G^*(e,\alpha)}[\ell(\nu)].$$

As in the main text,

$$W^{dep}(e, \alpha) = M(e, \alpha) - (1 - \alpha)L(e, \alpha), \tag{23}$$

$$U^{dep}(e, \alpha) = \beta\big(M(e, \alpha) - V(\mu_0; \alpha)\big) - (1 - \alpha)L(e, \alpha). \tag{24}$$

The provider's date-1 problem is

$$e^*(\alpha) \in \arg\max_{e \in [\underline{e}, \bar{e}]} \Big\{ U^{dep}(e, \alpha) - c(e) \Big\}, \tag{25}$$

and equilibrium total welfare is $W^{tr}(\alpha) := W^{dep}(e^*(\alpha), \alpha) - c(e^*(\alpha))$.

Local regularity conditions near $\alpha = 1$. Assume that, in a left-neighborhood of $\alpha = 1$ and for e near $e_1 := e^*(1)$:

1. the equilibrium selection $G^*(e, \alpha)$ is single-valued;
2. the induced objects $M(e, \alpha)$ and $L(e, \alpha)$ are twice continuously differentiable in e and left-differentiable in α;
3. the date-1 problem (25) admits a unique interior solution $e^*(\alpha)$.

Economic sign conditions for Proposition 6. In addition, for Proposition 6, assume

$$M_e(e_1, 1) > 0, \qquad \beta M_{ee}(e_1, 1) - c''(e_1) < 0, \qquad \beta M_{e\alpha}(e_1, 1^-) + L_e(e_1, 1^-) < 0. \tag{26}$$

These conditions require, respectively, that a more informative score is locally valuable, that the provider's date-1 problem is locally concave in effort, and that stronger decision-maker liability lowers the provider's marginal return to training near $\alpha = 1$.

Lemma D.1 (Derivative gap)**.** *For every (e, α),*

$$W_e^{dep}(e, \alpha) - U_e^{dep}(e, \alpha) = (1 - \beta)M_e(e, \alpha).$$

Proof. Differentiate (23) and (24) with respect to e:

$$W_e^{dep}(e, \alpha) = M_e(e, \alpha) - (1 - \alpha)L_e(e, \alpha),$$

$$U_e^{dep}(e, \alpha) = \beta M_e(e, \alpha) - (1 - \alpha)L_e(e, \alpha),$$

because $V(\mu_0;\alpha)$ does not depend on e. Subtracting the two identities yields

$$W_e^{dep}(e,\alpha) - U_e^{dep}(e,\alpha) = (1-\beta)M_e(e,\alpha).$$

□

Proof of Proposition 6. Let $e_1 := e^*(1)$. Because the date-1 training problem is interior in a left-neighborhood of $\alpha = 1$, the first-order condition is

$$U_e^{dep}(e^*(\alpha),\alpha) = c'(e^*(\alpha)). \tag{27}$$

Using (24), this becomes

$$\beta M_e(e^*(\alpha),\alpha) - (1-\alpha)L_e(e^*(\alpha),\alpha) = c'(e^*(\alpha)).$$

Step 1: Derivative of the training choice. Differentiate (27) with respect to α from the left at $\alpha = 1$. By the chain rule,

$$\Big[\beta M_{ee}(e_1,1) - c''(e_1)\Big]e^{*\prime}(1^-) + \beta M_{e\alpha}(e_1,1^-) + L_e(e_1,1^-) = 0.$$

Rearranging gives

$$e^{*\prime}(1^-) = \frac{-(\beta M_{e\alpha}(e_1,1^-) + L_e(e_1,1^-))}{\beta M_{ee}(e_1,1) - c''(e_1)}.$$

This is (12). Under (26), the denominator is strictly negative and the numerator is strictly negative, hence

$$e^{*\prime}(1^-) < 0.$$

Step 2: Chain-rule decomposition of total welfare. By definition,

$$W^{tr}(\alpha) = W^{dep}(e^*(\alpha),\alpha) - c(e^*(\alpha)).$$

Differentiate from the left at $\alpha = 1$:

$$(W^{tr})'(1^-) = W_e^{dep}(e_1,1)e^{*\prime}(1^-) + W_\alpha^{dep}(e_1,1^-) - c'(e_1)e^{*\prime}(1^-).$$

Since $e_1 = e^*(1)$ solves the date-1 problem, the first-order condition at $\alpha = 1$ gives

$$c'(e_1) = U_e^{dep}(e_1,1).$$

Substituting,

$$(W^{tr})'(1^-) = \Big(W_e^{dep}(e_1,1) - U_e^{dep}(e_1,1)\Big)e^{*\prime}(1^-) + W_\alpha^{dep}(e_1,1^-).$$

By Lemma D.1,

$$(W^{tr})'(1^-) = (1-\beta)M_e(e_1,1)e^{*\prime}(1^-) + W_\alpha^{dep}(e_1,1^-),$$

which is (13).

Step 3: Sign of each margin. Because $\beta < 1$, $M_e(e_1,1) > 0$, and $e^{*\prime}(1^-) < 0$, the first term

$$(1-\beta)M_e(e_1,1)e^{*\prime}(1^-)$$

is strictly negative.

It remains to show that the second term is weakly positive. Fix any effort level e. Conditional on e, the date-2 game is exactly the baseline model with score distribution F_e. Therefore, Proposition 4 applies pointwise in e:

$$W^{dep}(e,\alpha) \le W^{dep}(e,1) \qquad \text{for all } \alpha \in [0,1].$$

Hence the left derivative at $\alpha = 1$ satisfies

$$W_\alpha^{dep}(e_1,1^-) \ge 0.$$

Step 4: Dominance condition and local improvement away from full liability. If

$$(1-\beta)M_e(e_1,1)\left|e^{*\prime}(1^-)\right| > W_\alpha^{dep}(e_1,1^-),$$

then (13) implies

$$(W^{tr})'(1^-) < 0.$$

By left differentiability of W^{tr} at $\alpha = 1$, there exists $\bar{\alpha} < 1$, arbitrarily close to 1, such that

$$W^{tr}(\bar{\alpha}) > W^{tr}(1).$$

This proves the proposition. □

Corollary D.1 (Global full-liability benchmark when $\beta = 1$)**.** *Suppose $\beta = 1$. Then full decision-maker liability remains globally welfare-optimal even when training is endogenous:*

$$W^{tr}(\alpha) \le W^{tr}(1) \qquad \textit{for all } \alpha \in [0,1].$$

Proof. Fix any effort level e. Conditional on e, the date-2 game is exactly the baseline model with score distribution F_e. Therefore, Proposition 4 applies pointwise:

$$W^{dep}(e,\alpha) \le W^{dep}(e,1) \qquad \text{for all } \alpha \in [0,1]. \tag{28}$$

Now suppose $\beta = 1$. By (24),

$$U^{dep}(e,1) = M(e,1) - V(\mu_0;1),$$

because the misclassification term disappears when $\alpha = 1$. Since $V(\mu_0;1)$ is constant in e, the provider's date-1 problem at $\alpha = 1$ is equivalent to

$$\max_e \{M(e,1) - c(e)\}.$$

Thus, $e^*(1)$ solves

$$M(e^*(1),1) - c(e^*(1)) \geq M(e,1) - c(e) \qquad \text{for all } e.$$

For any $\alpha \in [0,1]$,

$$W^{tr}(\alpha) = W^{dep}(e^*(\alpha),\alpha) - c(e^*(\alpha)) \leq W^{dep}(e^*(\alpha),1) - c(e^*(\alpha))$$

by (28). But when $\alpha = 1$,

$$W^{dep}(e,1) = M(e,1),$$

so

$$W^{tr}(\alpha) \leq M(e^*(\alpha),1) - c(e^*(\alpha)) \leq M(e^*(1),1) - c(e^*(1)) = W^{tr}(1).$$

Therefore,

$$W^{tr}(\alpha) \leq W^{tr}(1) \qquad \text{for all } \alpha \in [0,1],$$

which proves the corollary. □

References

AMERICAN COLLEGE OF RADIOLOGY (2025): "ACR Breast Imaging Reporting and Data System (BI-RADS®)," https://www.acr.org/Clinical-Resources/Clinical-Tools-and-Reference/Reporting-and-Data-Systems/BI-RADS, official ACR page describing standardized breast imaging terminology, assessment structure, and management recommendations. Accessed March 19, 2026.

ARIELI, I., Y. BABICHENKO, R. SMORODINSKY, AND T. YAMASHITA (2023): "Optimal persuasion via bi-pooling," *Theoretical Economics*, 18, 15–36.

BARTLING, B. AND U. FISCHBACHER (2012): "Shifting the Blame: On Delegation and Responsibility," *The Review of Economic Studies*, 79, 67–87.

BERGEMANN, D. AND S. MORRIS (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57, 44–95.

BUÇINCA, Z., M. B. MALAYA, AND K. Z. GAJOS (2021): "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," *Proceedings of the ACM on Human-Computer Interaction*, 5, 1–21.

BYRNE, D. P. AND N. DE ROOS (2022): "Start-up Search Costs," *American Economic Journal: Microeconomics*, 14, 81–112.

CAPLIN, A. AND M. DEAN (2015): "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *American Economic Review*, 105, 2183–2203.

DIETVORST, B. J., J. P. SIMMONS, AND C. MASSEY (2015): "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," *Journal of Experimental Psychology: General*, 144, 114–126.

DIZDAR, D. AND E. KOVÁČ (2020): "A simple proof of strong duality in the linear persuasion problem," *Games and Economic Behavior*, 122, 407–412.

DWORCZAK, P. AND G. MARTINI (2019): "The simple economics of optimal persuasion," *Journal of Political Economy*, 127, 1993–2048.

EU AI ACT (2024): "Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (AI Act)," https://eur-lex.europa.eu/eli/reg/2024/1689/oj?locale=en, issued by the European Parliament and Council of the European Union. Official Journal text; see Articles 14, 16, 26, and 113. Accessed February 27, 2026.

EU PRODUCT LIABILITY DIRECTIVE (2024): "Directive (EU) 2024/2853 on Liability for Defective Products (Revised Product Liability Directive)," https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202402853, issued by the European Parliament and Council of the European Union. Applies to products placed on the market after 9 December 2026; classifies software and AI systems as "products" subject to strict liability; includes evidence-disclosure obligations and rebuttable presumptions of defect and causation. Accessed March 1, 2026.

GENTZKOW, M. AND E. KAMENICA (2016): "A Rothschild-Stiglitz approach to Bayesian persuasion," *American Economic Review*, 106, 597–601.

HALAC, M. AND P. YARED (2020): "Commitment versus Flexibility with Costly Verification," *Journal of Political Economy*, 128, 4523–4573.

HOLMSTRÖM, B. (1982): "Moral Hazard in Teams," *The Bell Journal of Economics*, 13, 324–340.

KAMENICA, E. AND M. GENTZKOW (2011): "Bayesian persuasion," *American Economic Review*, 101, 2590–2615.

KOLOTILIN, A. (2018): "Optimal information disclosure: A linear programming approach," *Theoretical Economics*, 13, 607–635.

MACKOWIAK, B., F. MATEJKA, AND M. WIEDERHOLT (2023): "Rational Inattention: A Review," *Journal of Economic Literature*, 61, 226–273.

MATEJKA, F. AND A. MCKAY (2015): "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, 105, 272–298.

MATYSKOVA, L. AND A. MONTES (2023): "Bayesian persuasion with costly information acquisition," *Journal of Economic Theory*, 211, 105678.

MELLO, M. M. AND N. GUHA (2024): "Understanding Liability Risk from Using Health Care Artificial Intelligence Tools," *New England Journal of Medicine*, 390, 271–278.

MOOKHERJEE, D. AND I. PNG (1989): "Optimal auditing, insurance, and redistribution," *The Quarterly Journal of Economics*, 104, 399–415.

NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION (2025a): "Standing General Order on Crash Reporting," https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting, crash-reporting requirements for ADS and Level 2 ADAS; public data notes through December 15, 2025. Accessed February 27, 2026.

——— (2025b): "Third Amended Standing General Order 2021-01: Incident Reporting for Automated Driving Systems (ADS) and Level 2 Advanced Driver Assistance Systems (ADAS)," https://www.nhtsa.gov/document/sgo-crash-reporting-adas-ads, effective June 16, 2025. Accessed February 27, 2026.

PRICE, W. N. I. AND I. G. COHEN (2024): "Locating Liability for Medical AI," *DePaul Law Review*, 73, 339.

SIMS, C. A. (2003): "Implications of rational inattention," *Journal of Monetary Economics*, 50, 665–690.

TEXAS OFFICE OF THE ATTORNEY GENERAL (2024): "Assurance of Voluntary Compliance: Pieces Technologies, Inc." https://www.texasattorneygeneral.gov/sites/default/files/images/press/Pieces%20Technologies%20AVC.pdf, settlement concerning alleged misrepresentations about generative-AI clinical documentation accuracy. Accessed March 1, 2026.

TOWNSEND, R. M. (1979): "Optimal contracts and competitive markets with costly state verification," *Journal of Economic Theory*, 21, 265–293.

U.S. FOOD AND DRUG ADMINISTRATION (2024): "510(k) Summary: VinDr-Mammo (K233108)," https://www.accessdata.fda.gov/cdrh_docs/pdf23/K233108.pdf, fDA clearance summary stating that VinDr-Mammo provides exam-level output for prioritization/worklist triage and that interpreting physicians remain responsible for reviewing each exam. Accessed March 19, 2026.

VACCARO, M., A. ALMAATOUQ, AND T. MALONE (2024): "When combinations of humans and AI are useful: A systematic review and meta-analysis," *Nature Human Behaviour*, 8, 2293–2303.