Act, Ask, or Defer? Designing Human Oversight for Agentic AI

Wichinpong Park Sinchaisri
Haas School of Business, University of California, Berkeley, parksinchaisri@berkeley.edu

Haonan Deng
University of California, Berkeley, 121090096_dhn@berkeley.edu

Abstract. *Problem definition:* Firms using AI agents must decide which proposals to execute, which to send for human review, and which requests employees should handle independently. Review preserves the agent's work, but its value depends on whether employees inspect and what happens when they find an error. *Methodology/results:* We model a firm that chooses which cases enter a capacity-constrained review station and what information reviewers receive, anticipating that reviewers inspect more when they expect to find errors. Withholding information can sustain attention, but it may require admitting cases the firm would otherwise exclude. We show that full disclosure of the firm's assessment is optimal if and only if no feasible mixture of cases gains more from changing attention than the opportunity cost of the admissions it requires. With fixed work per admission and separate correction, expanding access to a review station can favor withholding information; the direction can reverse when reviewers also make repairs. A refund-desk experiment with 118 participants finds that inspection increases with the frequency of errors. In numerical comparisons built from e-commerce orders and a public benchmark of AI agents, review loses most of its value as correction becomes as costly as independent handling. In the reference comparison where inspection rises in proportion to catchable errors, optimizing admissions while revealing the full assessment captures over 98% of the attainable review gain. With only action and amount visible, assigning the same workloads without accounting for attention loses nearly half of review's gain. *Managerial implications:* The value of review depends on correction costs, who performs repairs, and how assignments affect attention. Firms setting monetary approval limits should also consider the agent's confidence. Before withholding information, firms should measure how reviewers inspect the proposed mix of cases.

Key words: Human–AI collaboration; human oversight; behavioral operations; admission control; information design
This version: November 12, 2026

1. Introduction

Organizations have long separated the authority to act from the duty to check. One employee, the maker, prepares a transaction; another, the checker, verifies it before it takes effect. Payments above a limit need a second signature, insurance claims above a threshold go to an examiner, and flagged transactions wait for an analyst before they settle. AI agents now take over the maker's role. With operational tools, they change orders, issue refunds, and modify accounts (Yao et al. 2025), and firms can be liable for what their agents tell customers (Moffatt v. Air Canada 2024). One response is to retain a human checker; oversight requirements for high-risk systems explicitly address the risk of over-relying on automated output (European Union 2024, Article 14). Yet a check protects the firm only if the reviewer chooses to look and can recognize an error when she does.

Consider a firm whose AI agent handles refund requests. For each request, the agent gathers the transaction record, applies the refund policy, and proposes a payment. The firm can then let the agent *act*, *ask* a reviewer to approve the payment, or *defer* the request to an employee who resolves it from scratch. Review keeps the agent's work and allows a check of the supporting record, but what happens next depends on the reviewer. If she inspects and blocks the payment, someone must still work out the right resolution. If she approves without inspecting, the payment goes through as proposed. We ask: *How should a firm choose which AI proposals to review and what reviewers see when these choices affect whether employees inspect?*

We address this question with an analytical model combining admission control and information design. Requests differ in confidence, the consequences of an error, and correction costs. Reviewers choose how often to inspect; when they do, they apply a fixed, imperfect screening rule. The firm's admissions and disclosures must respect capacity and the information employees already observe. We first analyze the choice of execution mode under full inspection, then study how the opportunity cost of admitting cases determines the value of disclosure when inspection is discretionary.

Even under full inspection, the riskiest proposals need not be the best use of review. A large refund may be better handled directly if its errors are hard to detect, or if correcting one takes nearly as much work as starting over. A smaller refund may deserve review if an employee can spot its errors reliably and fix them cheaply. Blocking a valid payment is costly as well, because it triggers investigation, delay, and restoration. What matters is the loss that inspection prevents net of these costs, compared with the better of letting the agent act and handling the request directly. Confidence alone omits the consequences of an error, whereas expected loss omits how much of it a check can economically prevent (Lemma 1 and Proposition 1).

When inspection is discretionary, realizing this saving also depends on the rest of the reviewer's queue. In visual search, rare targets are missed more often (Wolfe et al. 2005, 2013), partly because searchers stop looking sooner (Wolfe and Van Wert 2010). Food-safety inspectors cite more violations after encountering worse compliance at their previous site (Ibanez and Toffel 2020). In a model of AI oversight, rarer errors make inspection more costly to motivate (Bastani and Cachon 2025). Suppose, then, that scrutiny (the probability that the reviewer inspects) rises with how often inspection uncovers an error. Admitting proposals with fewer catchable errors than the existing pool dilutes the queue and lowers attention to requests the reviewer cannot distinguish from them. Proposals with more catchable errors than the pool have the opposite effect, but they also take capacity and create correction work.

This spillover operates only among requests that look alike to the reviewer before she inspects. The proposed payment and visible transaction fields already set some cases apart. When the agent's confidence is not already visible or inferable, the firm must also decide whether to show it. Showing confidence lets the reviewer focus on likely errors. Withholding it keeps cases with different error rates in a common pool, so the scrutiny each receives, and hence the value of withholding, depends on which cases the firm admits.

Suppose, for example, that two kinds of proposals look alike on the screen. The first contains frequent catchable errors that save little when corrected; the second contains rarer ones whose correction is valuable. With scrutiny increasing in the frequency of catchable errors, reviewing the two together draws more attention to the second kind than it gets alone. The first kind sustains that attention, but its own review need not be worthwhile. A third kind of proposal is visibly different and worth checking on its own merits. When review slots are scarce, pooling the first two pays only if the attention it adds is worth more than the reviews of the third kind it displaces.

Our main result formalizes this tradeoff. Under full disclosure, reviewers see the assessment the firm uses to route each case, including the agent's confidence and the assessed loss. We value each admission at a common price that reflects congestion and the work it displaces. Theorem 1 gives an admission-price certificate: full disclosure is optimal exactly when, at that price, no feasible grouping gains more from changing attention than the opportunity cost of the admissions it requires. For any continuous scrutiny response, it suffices to check mixtures of two request types. The certificate identifies when the simple policy of full disclosure and admissions ranked by their value under the resulting scrutiny cannot be improved by pooling cases. The closest analyses study score coarsening when a decision maker chooses whether to open costly review (Yang 2026) and the pay needed to induce inspection of AI output (Bastani and Cachon 2025); the certificate prices the cases needed to sustain attention under congestion.

The certificate implies that allowing more admissions to a fixed review station can make withholding optimal. When sustaining attention requires cases the firm would otherwise exclude, a tight admission limit makes those cases expensive, because each one displaces other work. Relaxing the limit lowers that cost, and once selective disclosure becomes strictly better, further relaxation cannot restore full disclosure. This direction relies on two features of the baseline: every admission requires the same service work, and other employees make corrections.

Assigning repairs to the checker changes both her incentive to inspect and the use of review capacity. A reviewer who must repair each error she finds anticipates that burden, and completed

repairs occupy time otherwise available for review. We construct an example in which added capacity then restores full disclosure. In the numerical comparisons, assigning repairs to reviewers cuts review value by about two-thirds, through both channels. If reviewers respond strongly enough to pay, bonuses paid for verified corrections restore part of that value and outperform selective disclosure at the same operating cost.

To examine the behavioral premise underlying these results, we conducted a supplementary refund-desk experiment with 118 participants. They decided how many AI proposals to audit in queues with different error frequencies but the same length and rewards. Inspection fell as errors became rarer, but in the low-error queues participants inspected more than would maximize their expected score at the observed error-detection and false-block rates. This pattern motivates one inspection response in our policy comparisons.

We quantify the operational consequences with policy comparisons that combine monetary variation in Brazilian e-commerce orders with agent success rates from a public benchmark of AI customer-service agents. Review loses most of its value as correction becomes as costly as independent handling; at the benchmark's success rates, much of its gain comes from preserving work that would otherwise be redone. Adding the agent's confidence to monetary approval limits improves these limits under the reference inputs. When inspection rises in proportion to catchable errors and reviewers see the action and amount, ignoring attention while assigning a fixed workload among cases that look alike loses value equal to the cost of about 19 independent resolutions per 1,000 arrivals. Once admissions are optimized, selective disclosure adds less than the cost of one such resolution per 1,000 arrivals over full disclosure. Optimized full disclosure is therefore a useful benchmark for joint design. Larger gains from withholding depend on inspection at error frequencies beyond those observed in the experiment.

For managers, the results make oversight a matter of work design. The firm sets which cases a reviewer receives, what her screen shows, and whether she repairs the errors she finds, and each choice changes how often she inspects. Similar tradeoffs can arise in payment approvals and claims examination when a check preserves prepared work, correction is costly, and attention responds to the assigned cases. For regulators, a rule that specifies how many cases a person must review does not determine how many errors that person catches.

2. Related Literature

Deciding whether an AI proposal executes, is reviewed, or is replaced by an employee's work is a problem of allocating authority. Aghion and Tirole (1997) distinguish formal authority from

effective control, and Dessein (2002) compares delegation with communication by an informed agent. At the request level, learning to defer routes cases using human and machine performance (Mozannar and Sontag 2020), and field evidence documents gains from AI assistance in customer support (Brynjolfsson et al. 2025). Our problem differs because review is a check before execution: it preserves the proposal, but the employee may choose not to inspect it.

Because the check precedes execution, review is related to gatekeeping and to the quality of human–AI collaboration. Gatekeepers decide when a case warrants specialist treatment, and workload and incentives affect these referrals (Shumsky and Pinker 2003, Hathaway et al. 2023). Dai and Singh (2025) compare AI as a gatekeeper with AI as a second opinion in patient pathways. In our review mode, an uninspected proposal executes, and a blocked proposal still requires resolution. Even when a person inspects, her judgment may be imperfect. Combined human–AI performance does not consistently exceed that of the better of the two (Vaccaro et al. 2024). Elicitation and disclosure can improve the use of private human information (Ibrahim et al. 2021, Balakrishnan et al. 2026); belief-updating errors and the communication of a deferral can change how advice is used (Bondi et al. 2022, Agarwal et al. 2023). We separate this quality of a check from the decision to perform it and account for the corrective work it creates.

The decision to inspect itself responds to machine assistance. Automation complacency and bias share an attentional component (Parasuraman and Manzey 2010), and more capable AI can discourage human information acquisition (Athey et al. 2020, Dell'Acqua 2022). Models of cognitive limitations explain how machine input changes attention and how explanations trade accuracy against cognitive burden (Boyacı et al. 2024, 2026), and algorithm quality and system load affect reliance and deliberation time (Snyder et al. 2026). Over repeated interaction, people learn about machine quality (de Véricourt and Gurkan 2026) and from AI (Noti et al. 2025), and interpretable tips can improve sequential decisions (Bastani et al. 2026). Feedback composition can affect rare-event detection (Epping et al. 2026). We hold the screening rule fixed to isolate how the expected frequency of interceptable errors changes the employee's incentive to use that rule.

Admission affects both the amount of review work and how it is done. Classical admission control prices the delay imposed on other customers (Naor 1969). Workload can change processing time and quality (Kc and Terwiesch 2009). Recent AI service models connect cognitive effort to congestion and staffing (Legros and de Véricourt 2026), verification to customers' willingness to wait (Yang et al. 2026), and automated screening to human rework bottlenecks (Lin and Zhang 2026). Our baseline holds service work per admission fixed. Even at the same workload within

every public-information group, replacing one admitted case with another can change scrutiny. The marginal value of an admission includes this effect on the cases already assigned.

Whether that composition effect benefits the firm depends on how inspection is rewarded. Multitask incentives and imperfect performance measures can direct effort away from organizational value (Holmström and Milgrom 1991, Baker 1992). Bastani and Cachon (2025) characterize the compensation needed to induce inspection and repair of AI output, showing why greater reliability can make oversight more expensive. Their principal chooses the contract and can restrict AI access; our baseline takes compensation as given and chooses admissions and disclosure at a congested review station. Related work jointly designs task assignment, review, and pay (Friedman 2026), or AI temperature and incentive pay (Dai and Taylor 2026). Admission and disclosure manage the gap between the employee's reward for catching errors and the firm's saving after correction.

Disclosure is the firm's second instrument, and information design studies how it influences attention within an assigned decision problem. A principal can restrict information to redirect costly attention (Lipnowski et al. 2020), and partial disclosure of a machine forecast or a reduction in its precision can preserve human information acquisition (Guan et al. 2025, Lu and Tomlin 2026). Dreyfuss and Hoong (2026) study calibrated coarsening when a human combines AI information with private evidence and may update imperfectly. The recipient's response also matters in recommendation problems with imperfect adherence (Grand-Clément and Pauphilet 2026). When the recipient may also consult personal AI, an advisor optimally counteracts its predictable recommendation (Liu and Sinchaisri 2026).

In service models, information already affects entry: customers use information about congestion to decide whether to join a service system (Lingenbrink and Iyer 2019, Anunrojwong et al. 2023). Closest to ours, Yang (2026) studies how a provider's coarsening of a calibrated score changes a human decision maker's choice to open costly review under shared liability. In that deployment model, the score distribution is given, but which cases receive review is endogenous. Our firm instead chooses admissions directly and designs disclosure for the admitted pool. Because attention to one request depends on the other admitted requests that the reviewer cannot distinguish from it, the disclosure decision must price the cases needed to create that pool, subject to capacity and public information. The analysis uses the comparison of experiments (Blackwell 1953) and committed information design (Kamenica and Gentzkow 2011). Rayo and Segal (2010) allow prospects to be excluded as well as pooled, and pairwise and dual methods also apply to assignment with peer effects (Kolotilin et al. 2025). Our certificate prices admission through congestion and displaced

work, identifying when a profitable mixture remains worth assembling and when broader access can favor selective disclosure.

3. Model: Service Requests and Human Oversight

We consider a firm whose AI agent proposes an action for each service request before the action binds the firm. The firm can authorize the proposal, submit it to review, or assign the request to an employee for independent resolution. A reviewed proposal executes unless the reviewer inspects it and a fixed screening rule, called the execution gate, blocks it. The firm commits to admissions and to the information reviewers see before they choose whether to inspect.

3.1. Model Primitives

Requests are indexed by i. An AI agent proposes an action of type g_i, and $Y_i \in \{0, 1\}$ indicates whether the proposal is appropriate. Before execution, the firm observes a routing state ξ_i containing the proposed action, its calibrated confidence, the assessed loss from an inappropriate execution, and the relevant case attributes. It does not observe Y_i. Let $\mathcal{F}_i = \sigma(\xi_i)$ denote its routing information. Confidence and autonomous risk are

$$x_i = \mathbb{P}(Y_i = 1 \mid \mathcal{F}_i), \qquad r_i = (1 - x_i)L_i, \tag{1}$$

where $L_i > 0$ is the conditional expected loss if the proposal is inappropriate and executes. Both quantities are assessed before execution, and the definition permits loss and correctness to be associated through the routing state. A raw model score need not equal x_i.

Let $Z_i = Z(\xi_i)$ collect the information available to the reviewer before costly inspection. In the refund workflow, Z_i includes the proposed payment and the transaction or policy fields visible on the review screen. The refund amount need not equal L_i, which can also include remediation and customer consequences, but any component of L_i that these fields reveal is public. Other internal assessments stay undisclosed only if the reviewer cannot infer them from the available records.

Before requests arrive, the firm commits to an authority rule and a disclosure policy. The authority rule assigns autonomous execution A, review K, or direct human handling D. If z_i is the probability of review, admissible rules satisfy $M = \mathbb{E}[z_i] \le \bar{M}$, where $\bar{M}$ limits entry into a fixed review station. For an admitted request, the disclosure policy generates a signal S_i from ξ_i; conditional on ξ_i, routing and disclosure randomization are independent of correctness and subsequent human evidence. The reviewer observes (Z_i, S_i) and knows that the request was admitted, and the policy and the distribution of routing states are common knowledge. Full disclosure reveals ξ_i but not the

unobserved correctness of the current proposal. Requests with the same public information form a public stratum, within which disclosure can group them.

The reviewer then chooses scrutiny $e \in [0, 1]$, the probability of inspecting the proposal; inspection requires the employee to acquire and evaluate the supporting evidence. If she inspects, she obtains the additional human signal and the AI evidence required by a fixed execution gate q, which either permits execution or blocks the proposal for corrective handling; if she does not, the proposed action executes. Thus S_i governs what she knows before the scrutiny decision, whereas q governs screening once she inspects. Because the gate is fixed, its conditional screening probabilities are unchanged when the firm varies admission or disclosure.

For a refund, the distinction is between seeing the proposed payment and checking the evidence that establishes eligibility. Opening correspondence, reconciling a return with an order, or checking an exception to the policy can require effort beyond reading the payment amount. The model applies when that verification decision precedes access to the decisive evidence. Information available at negligible cost before this choice belongs to Z_i and cannot be withheld through S_i.

3.2. Payoffs and Correction Costs

Each mode provides gross resolution value R, which cancels from authority comparisons and payoff differences. Autonomous execution and direct handling yield

$$U_{A,i} = R - r_i, \qquad U_{D,i} = R - d_i, \tag{2}$$

where $d_i > 0$ includes the labor, delay, and residual error cost of independent resolution, which replaces the AI proposal rather than inspecting it. The common-cost benchmark sets $d_i = d$.

Let $B_{q,i} = 1$ denote the event that gate q would block the proposal if activated. Define

$$\begin{aligned} \pi_{q,i} &= \mathbb{P}(Y_i = 0, B_{q,i} = 1 \mid \mathcal{F}_i), \\ u_{q,i} &= \mathbb{P}(Y_i = 1, B_{q,i} = 1 \mid \mathcal{F}_i). \end{aligned} \tag{3}$$

The first is the probability of an intercepted error and the second that of a false block, both evaluated as if the gate were activated and before the additional evidence is observed. They satisfy $0 \le \pi_{q,i} \le 1 - x_i$ and $0 \le u_{q,i} \le x_i$.

Each admission incurs a charge $c_q > 0$ for assignment, access to the review station, and noncongestion delay; corrective handling is charged separately. Let $k_i^E \ge 0$ be the expected cost of resolving a blocked inappropriate proposal and $k_i^V \ge 0$ the incremental cost of a false block, conditional on the corresponding event. These costs are assessed from the routing state, and we suppress their

dependence on q. Because these costs include subsequent handling, delay, and any loss that remains after correction, a blocked error need not be resolved perfectly or costlessly.

For a fully activated gate, expected correction cost, prevented loss, and residual risk are

$$\begin{aligned} \kappa_{q,i} &= k_i^E \pi_{q,i} + k_i^V u_{q,i}, \\ \ell_{q,i} &= L_i \pi_{q,i}, \qquad \rho_{q,i} = r_i - \ell_{q,i}. \end{aligned} \tag{4}$$

We assume that, conditional on ξ_i and an inappropriate proposal, its execution loss has mean L_i whether or not the gate would block it. This separates the magnitude of the original error from its detectability within a routing state. The net value of activating the gate is

$$v_{q,i} = \ell_{q,i} - \kappa_{q,i}. \tag{5}$$

This value can be negative even when the gate intercepts errors. At scrutiny e, review yields

$$U_{K,i}^q(e) = R - c_q - r_i + e\, v_{q,i}. \tag{6}$$

In particular, $U_{K,i}^q(1) = R - c_q - \rho_{q,i} - \kappa_{q,i}$.

For $r_i > 0$, define screening effectiveness and net review value under full inspection by

$$\eta_{q,i} = 1 - \frac{\rho_{q,i}}{r_i}, \tag{7}$$

$$\begin{aligned} \mathcal{V}_{q,i} &= \min\{r_i, d_i\} - \rho_{q,i} - \kappa_{q,i} \\ &= r_i \eta_{q,i} - \kappa_{q,i} - (r_i - d_i)^+. \end{aligned} \tag{8}$$

Here $(\cdot)^+$ denotes the positive part. Screening effectiveness is the fraction of original execution loss prevented. Net review value is the cost saved relative to the better nonreview alternative, net of correction costs but before the admission charge, so full inspection is worthwhile when $\mathcal{V}_{q,i} > c_q$. At scrutiny e, the corresponding incremental payoff is $\min\{r_i, d_i\} - r_i - c_q + e v_{q,i}$.

3.3. Reviewer Scrutiny

We model the quality of a completed inspection separately from the employee's willingness to perform one. The gate captures quality: it specifies the decision made after inspection. In the stakes-blind specification, its probability of permitting an inappropriate proposal depends on confidence and action type but not on downstream loss. For $x < 1$,

$$\begin{aligned} \psi_q(x; g) &= \mathbb{P}(B_q = 0 \mid Y = 0, x, g), \\ \rho_q(x, L; g) &= \psi_q(x; g)(1 - x)L. \end{aligned} \tag{9}$$

Hence $\eta_q(x;g) = 1 - \psi_q(x;g)$ and $\pi_q(x;g) = (1-x)\eta_q(x;g)$; at $x = 1$, $\pi_q = 0$. A proportional gate has constant detection probability $\eta_q = \beta_q \in [0,1]$ conditional on an inappropriate proposal. Its interception probability is $\pi_q = (1-x)\beta_q$, and

$$\rho_q(r) = (1-\beta_q)r. \tag{10}$$

Even so, net value need not be proportional to risk when correction costs vary across requests.

For the signal-based implementation, let θ be a latent index of the proposal's appropriateness, with $Y = \mathbf{1}\{\theta \geq 0\}$, and

$$\begin{aligned} s_A &= \theta + \varepsilon_A, \qquad \varepsilon_A \sim N(0, \tau^2), \\ s_H &= \theta + \mu + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2), \end{aligned} \tag{11}$$

where the positive-variance errors are mutually independent and independent of θ, conditional on action type. Conditional on (s_A, g), other routing attributes carry no additional information about (θ, s_H) in this specification, so $x = \mathbb{P}(\theta \geq 0 \mid s_A, g)$. Gate q combines the AI and human signals and permits execution when its assessment is nonnegative. Calibration corrects the shift μ and adjusts the weight on human evidence, but a calibrated gate may still miss an inappropriate proposal or block an appropriate one, and both probabilities enter its economic evaluation.

Scrutiny, by contrast, is chosen before s_H is seen. With correct beliefs, the employee's perceived probability of an interceptable error is

$$p_q(Z,S) = \mathbb{E}[\pi_{q,i} \mid Z, S, K]. \tag{12}$$

Conditioning on K accounts for the firm's admission rule. Let $b_H > 0$ be the employee's value of an intercepted error under a fixed performance-evaluation rule. The employee incurs cognitive cost $\Psi_H(e)$, which is continuous, nondecreasing, and strictly convex on $[0,1]$, with $\Psi_H(0) = 0$. Thus

$$e = h(p_q), \qquad h(p) = \arg\max_{0 \leq e \leq 1} \{b_H p e - \Psi_H(e)\}. \tag{13}$$

Quadratic cost $\Psi_H(e) = \iota_H e^2/2$ gives $h(p) = \min\{\gamma p, 1\}$, where $\gamma = b_H/\iota_H$; the proportional benchmark uses $0 < \gamma \leq 1$, so scrutiny is proportional to the perceived interception probability. The evaluation rule creates an incentive gap: the employee values detected errors, whereas the firm values $v_{q,i}$, the avoided loss after correction costs. Error counts are thus an imperfect measure of the firm's gain from inspection (Baker 1992), and admission and disclosure determine how the employee's incentive to catch errors translates into that gain.

A standing benefit from checking can sustain scrutiny even when interceptable errors are rare. Adding $a_H e$, with $a_H \geq 0$, to the employee's objective and retaining quadratic cost gives

$$h_+(p) = \min\{1, (a_H + b_H p)/\iota_H\}. \tag{14}$$

We call h_+ the positive-default response: the employee values routine checking in addition to the reward for intercepted errors. The numerical comparison uses $a_H/\iota_H = 0.25$ and $b_H/\iota_H = 1$, so full scrutiny begins at $p = 0.75$. The positive-default form is motivated by the inspection levels observed in the experiment (Section 6.1). The general disclosure result allows either response, and the first-order admission conditions of Proposition 3 apply where the response is differentiable.

The baseline uses a common scrutiny response for reviewers who know the committed admission and disclosure policies. Their posterior interception probability governs inspection, while the fixed gate separates the decision to check from checking accuracy. Compensation and staffing are held fixed. Service times have a common distribution independent of inspection and outcomes, so congestion depends on admitted volume; corrections use a separate resource, with costs included in k_i^E and k_i^V. Appendix B.4 instead lets inspection and repairs take reviewer time, and Section 6.6 varies compensation. The employee's cognitive cost is not charged to the firm, and event-contingent payments enter the corresponding event costs.

4. Execution Authority Under Full Inspection

We first study authority choices when every reviewed proposal is inspected. This benchmark isolates the loss a check prevents and the corrective work it creates, and provides the request values used in the admission problem of Section 5. Relative to the classical allocation of inspection effort (Lindsay and Bishop 1964), the analysis adds the option to replace a proposal.

4.1. The Authority Frontier

Review competes with the better of autonomous execution and direct human handling. For a request with $r_i > 0$, define the cost-adjusted effectiveness frontier

$$\phi_q(i) = \frac{c_q + \kappa_{q,i} + (r_i - d_i)^+}{r_i}. \tag{15}$$

The frontier uses the correction cost that the specified gate generates at that routing state, and its final term accounts for the saving available from direct handling when autonomous risk exceeds d_i.

LEMMA 1 (**Authority with correction costs**). *Without review, autonomous execution is weakly optimal exactly when $r_i \le d_i$. Under full inspection, review is strictly optimal exactly when*

$$v_{q,i} > c_q + (r_i - d_i)^+, \tag{16}$$

or, for $r_i > 0$, when $\eta_{q,i} > \phi_q(i)$. The incremental value of offering review is

$$\mathcal{G}_{q,i} = [v_{q,i} - c_q - (r_i - d_i)^+]^+ = [\mathcal{V}_{q,i} - c_q]^+. \tag{17}$$

At $r_i = 0$, autonomous execution is strictly optimal.

Each intercepted error saves L_i but incurs k_i^E, and each false block adds k_i^V. Review must recover these costs as well as its admission charge and any advantage of direct handling, so a gate that catches many errors can still have little value on requests that are expensive to correct.

The two outside options also explain why the highest-risk requests need not receive review. At low risk, there is little loss to recover against the admission charge. At high risk, direct handling limits the firm's exposure, so review must justify keeping the proposal rather than replacing it. In between, inspection can preserve useful work while preventing enough loss to cover correction.

With common d and constant correction cost $\bar{\kappa}_q$, let $\bar{c}_q = c_q + \bar{\kappa}_q < d$. The frontier is minimized at $r = d$, where autonomy and direct handling tie. Under constant effectiveness $\beta \in (0, 1)$, review is attractive for an interval around this boundary if and only if $\beta d > \bar{c}_q$ (Appendix A).

4.2. Confidence and Net Review Value

Autonomous risk alone does not settle the comparison: requests with the same risk can differ in confidence, and a more confident proposal errs less often but loses more when it does. Whether review favors it depends on which errors the gate detects and what correcting them costs, and the signal structure of Section 3.3 separates the two.

ASSUMPTION 1 (**Anchored execution gate**). *Within action type g, confidence $x(s_A, g)$ is strictly increasing in s_A, and the posteriors of θ given (s_A, g) are ordered by the monotone likelihood ratio property. The gate permits execution if and only if $\widehat{\theta}_q(s_A, s_H; g) \ge 0$, where $\widehat{\theta}_q$ is nondecreasing in each signal. Conditional on (s_A, g), the remaining routing attributes carry no additional information about (θ, s_H) and do not enter the gate.*

PROPOSITION 1 (**Confidence and net review value**). *Under Assumption 1, screening effectiveness $\eta_q(x; g)$ is nonincreasing in confidence. For same-type requests with common risk $r > 0$, direct-handling cost d, and confidences $x_2 > x_1$, net review value is no higher at x_2 if and only if*

$$\kappa_{q,2} - \kappa_{q,1} \ge r\{\eta_q(x_2; g) - \eta_q(x_1; g)\}. \tag{18}$$

Hence nondecreasing correction cost along the risk contour preserves the confidence ordering.

Higher AI evidence favors approval through both the gate and the distribution of errors that reach it. Under these ordering assumptions, the fraction of errors intercepted therefore does not increase with confidence. Greater confidence can also reduce corrective work, however; when this saving exceeds the reduction in prevented loss, net review value rises. A firm should rank equal-risk requests by the saving inspection produces after correction, not by their error frequencies alone.

The same distinction determines whether confidence supports an upper cutoff. Fix g and common d, and suppose expected correction cost depends on confidence but not separately on L. Write

$$\bar{c}_q(x) = c_q + \kappa_q(x), \qquad H_q(x) = d\eta_q(x) - \kappa_q(x). \tag{19}$$

Here $H_q(x)$ is net review value at the autonomy–direct-handling boundary.

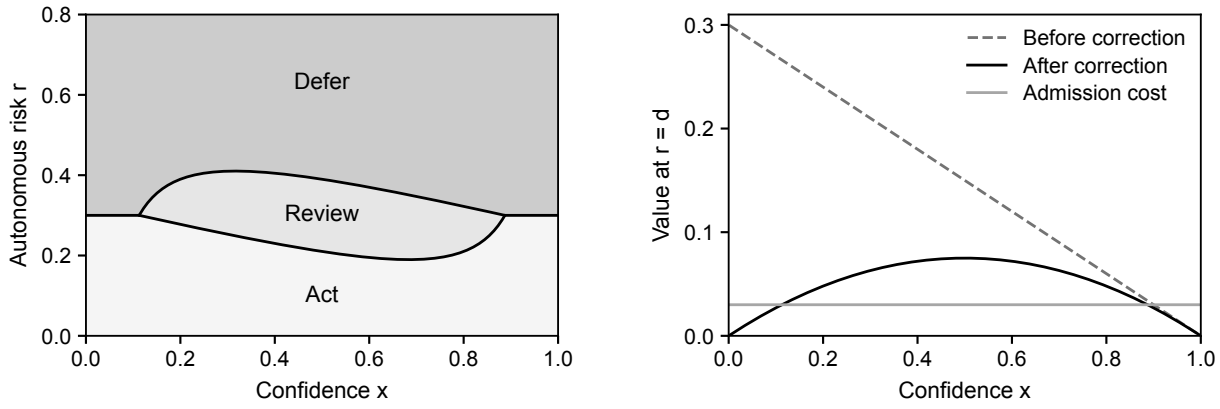
COROLLARY 1 **(Conditions for a confidence ceiling)**. *Suppose $\eta_q(x) \in (0,1)$ and $\kappa_q(x) \geq 0$ are continuous on a confidence interval $\mathcal{X}_g \subset (0,1)$. Allowing risk to vary over $(0,\infty)$, review is strictly optimal at some risk exactly when $H_q(x) > c_q$, in which case its region is*

$$\frac{\bar{c}_q(x)}{\eta_q(x)} < r < \frac{d - \bar{c}_q(x)}{1 - \eta_q(x)}. \tag{20}$$

On the request population, this interval is intersected with attainable risks conditional on confidence. If H_q is strictly decreasing and crosses c_q from above, there is a unique confidence ceiling over unrestricted risks. If, in addition, η_q is nonincreasing and κ_q is nondecreasing, the intervals in (20) are nested and contract to $r = d$ at the ceiling.

The ceiling is thus determined by net review value at $r = d$, where autonomy and direct handling tie. Correction costs can instead confine review to an interior range of confidences; Appendix A gives an example and the ordering condition for stakes-dependent gates.

Figure 1 uses $d = k^E = 0.30$, $c_q = 0.03$, $k^V = 0$, and $\eta(x) = 1 - x$. Screening effectiveness falls with confidence, yet net review value at $r = d$, $H(x) = 0.30x(1-x)$, initially rises because fewer errors require correction. At high confidence, the remaining errors are too difficult to intercept for review to cover its admission charge. Review is therefore confined to a band of confidences, so an upper cutoff alone would admit low-confidence proposals that are too costly to correct.



(a) Execution authority **(b)** Confidence and net value

Figure 1 Execution Authority and Net Review Value under Full Inspection

Note. Confidence-band example. Panel (b) evaluates value at the act-or-defer boundary $r = d$.

4.3. Calibration and the Value of Screening

Calibration changes both intercepted errors and false blocks. Compare a calibrated gate q with an uncorrected gate q' on the same request under full inspection, holding the conditional correction costs fixed across gates. Let $\Delta\pi_i = \pi_{q,i} - \pi_{q',i}$ and $\Delta u_i = u_{q,i} - u_{q',i}$. Their payoff difference is

$$\begin{aligned} &U^{q}_{K,i}(1) - U^{q'}_{K,i}(1) \\ &\qquad = (L_i - k^E_i)\Delta\pi_i - k^V_i \Delta u_i - (c_q - c_{q'}). \end{aligned} \tag{21}$$

When event costs also differ across gates, the comparison is $v_{q,i} - v_{q',i} - (c_q - c_{q'})$. In the Gaussian implementation, the approval thresholds of the calibrated and uncorrected gates can cross, so calibration makes the gate stricter for some AI signals and more lenient for others (Appendix A). Where it raises the approval threshold, it increases both interception and false blocking, so calibration pays only if the additional loss prevented exceeds the added correction and false-block costs and any change in the admission charge (Section 6.3).

5. Admission and Disclosure with Discretionary Inspection

We now let the firm choose which requests enter a capacity-constrained review station and what information reviewers receive. Under full inspection, an admission is worth its net value less the delay and displaced work it imposes. With discretionary inspection, it also changes attention to other assigned cases, and admissions and disclosure must be chosen jointly, subject to what reviewers observe and how they are rewarded.

Fix a gate q and drop its subscript. Finitely many routing types have shares $w_i > 0$ summing to one. The firm admits mass $y_i \in [0, w_i]$ of type i, with total mass $M = \sum_i y_i \le \bar{M} \in (0, 1)$. Define

$$t_i = \min\{r_i, d_i\} - r_i - c_q, \qquad v_i = L_i \pi_i - \kappa_i. \tag{22}$$

Thus a reviewed request contributes $t_i + e_i v_i$ relative to its best nonreview alternative. Congestion costs $C(M)$ are continuously differentiable, strictly convex, and nondecreasing, with $C(0) = 0$. Service-time distributions are identical across admissions, and the admission limit constrains access to this fixed station; corrections use the separate resource of Section 3, with cost included in κ_i.

5.1. The Admission Price

Under full inspection, type i has admission value $a_i^I = t_i + v_i = \mathcal{V}_i - c_q$. The firm solves

$$\max_{0 \le y \le w,\ \mathbf{1}^\top y \le \bar{M}} \left\{ \sum_i a_i^I y_i - C(\mathbf{1}^\top y) \right\}. \tag{23}$$

PROPOSITION 2 (**Admission under full inspection**). *An optimal allocation admits types in decreasing order of a_i^I. Its total mass M_I is unique, and there is an admission price*

$$\varpi_I = C'(M_I) + \nu_I, \qquad \nu_I \ge 0, \quad \nu_I(\bar{M} - M_I) = 0, \tag{24}$$

such that $a_i^I > \varpi_I$ implies full admission and $a_i^I < \varpi_I$ implies exclusion. Types at the price can be partially admitted. Review is unused exactly when

$$\max_i a_i^I \le C'(0). \tag{25}$$

The admission price charges for marginal congestion and the opportunity cost of the admission limit. If $C(M) = M\bar{C}(M)$, marginal congestion is $\bar{C}(M) + M\bar{C}'(M)$, because a request imposes delay on other reviews as well as experiencing delay itself. This is the standard admission externality (Naor 1969), applied to the saving after correction costs and against the better nonreview alternative.

5.2. The Marginal Value of an Admission

Proposition 2 ranks admissions by individual net value. When inspection is discretionary, however, an admission also changes how carefully the reviewer checks the other cases assigned. To isolate this effect, consider types with the same public information and a common disclosure signal. With probability $\omega \in [0, 1]$, an independent cue reveals the routing type before scrutiny, but not whether

the proposal is appropriate; otherwise, the reviewer learns only that the request was admitted. Thus $\omega = 0$ pools all admitted types, whereas $\omega = 1$ reveals them. With correct beliefs, define

$$p(y) = \frac{\sum_i y_i \pi_i}{M}, \qquad \bar{v}(y) = \frac{\sum_i y_i v_i}{M}, \qquad M > 0. \tag{26}$$

The response h from Section 3.3 gives expected scrutiny

$$e_i(y) = \omega h(\pi_i) + (1-\omega) h(p(y)). \tag{27}$$

The firm's gain, with $F(0) = 0$, is

$$F(y) = \sum_i y_i \{t_i + e_i(y) v_i\} - C(M). \tag{28}$$

PROPOSITION 3 (**The marginal value of an admission**). *A global optimum exists. Suppose h is differentiable at $p(y)$. At any nonzero local optimum, let*

$$\begin{aligned} \widetilde{a}_i(y) = t_i + e_i(y) v_i \\ + (1-\omega) h'(p(y)) \bar{v}(y) \{\pi_i - p(y)\}. \end{aligned} \tag{29}$$

There is a price $\varpi = C'(M) + \nu$, with $\nu \geq 0$ and $\nu(\bar{M} - M) = 0$, such that $\widetilde{a}_i(y) \geq \varpi$ for a fully admitted type, $\widetilde{a}_i(y) \leq \varpi$ for an excluded type, and equality holds for partial admission. These conditions are sufficient when $\omega = 1$ or h is constant on the relevant probability range.

The first two terms in $\widetilde{a}_i(y)$ are the request's own contribution, and the third prices its effect on scrutiny of the existing pool. An above-average interception probability raises scrutiny wherever the employee's response is increasing and the type remains pooled. If average net inspection value is positive, this extra scrutiny benefits the other admissions. This benefit can justify a request whose own contribution is negative, although correction costs reduce it and can reverse its sign. Revealing the routing type removes this spillover by separating inspection decisions, which is why admissions and disclosure must be chosen together.

5.3. When Full Disclosure Is Optimal

The firm now chooses both which cases enter review and the signals that distinguish them. A disclosure policy consists of signal masses $m_s \geq 0$ and posterior type probabilities ϑ_s. Each posterior is supported within one public stratum $\{i : Z_i = z\}$. The induced admissions satisfy

$$y_i = \sum_s m_s \vartheta_{s,i} \leq w_i, \qquad M = \sum_s m_s \leq \bar{M}. \tag{30}$$

Signals concern the routing type and provide no additional information about correctness conditional on that type, and the independent cue is retained. A posterior then yields gain per admission

$$\begin{aligned}\Gamma(\vartheta) = \sum_i \vartheta_i [t_i + \omega h(\pi_i) v_i] \\ + (1-\omega) h(\pi^\top \vartheta) v^\top \vartheta.\end{aligned} \tag{31}$$

The firm maximizes $\sum_s m_s \Gamma(\vartheta_s) - C(M)$. All signals use the same station and incur no separate setup cost. Here $h : [0,1] \to [0,1]$ need only be continuous; it can be nonlinear or saturated.

Under full disclosure, the admission score is $a_i^F = t_i + h(\pi_i) v_i$. Let M_F be its optimal volume and ϖ_F^+ its highest compatible admission price. A compatible price equals $C'(M_F) + \nu$, satisfies complementary slackness, and supports admission in decreasing score order. The highest price resolves ties at a binding admission limit; when the limit is slack, it equals $C'(M_F)$. Write

$$\delta_i(\varpi) = (\varpi - a_i^F)^+. \tag{32}$$

This is the shortfall of type i's full-disclosure score below the admission price.

For a same-stratum pair, let $p_{ij}(\alpha) = (1-\alpha)\pi_i + \alpha \pi_j$ and $v_{ij}(\alpha) = (1-\alpha) v_i + \alpha v_j$. Its pooling premium is

$$\begin{aligned}\Delta_{ij}(\alpha) = (1-\omega) \big[h(p_{ij}(\alpha)) v_{ij}(\alpha) \\ - (1-\alpha) h(\pi_i) v_i - \alpha h(\pi_j) v_j \big].\end{aligned} \tag{33}$$

At a fixed posterior mean interception probability, Γ is linear in the posterior, so pairwise signals suffice, as in established information-design methods (Rayo and Segal 2010, Kolotilin et al. 2025). For n routing types, some optimal joint policy uses at most $n+1$ such signals within public strata.

THEOREM 1 **(Full disclosure and the admission price)**. *Full disclosure is globally optimal among joint admission and disclosure policies if and only if every same-stratum pair and every* $\alpha \in [0,1]$ *satisfy*

$$\Delta_{ij}(\alpha) \le (1-\alpha)\delta_i(\varpi_F^+) + \alpha \delta_j(\varpi_F^+). \tag{34}$$

Holding the station, type primitives, public information, cue probability, and scrutiny response fixed, the admission limits for which full disclosure is optimal form a lower interval, possibly empty or the entire domain.

In the proof, the pair inequalities bound each signal's gain per admission by ϖ_F^+ plus its types' posterior-weighted surpluses $(a_i^F - \varpi_F^+)^+$. Full disclosure attains the aggregate bound, and duality

gives the converse (Appendix B.1). Both conclusions allow negative v_i. Pooling must cover the admission shortfalls of the cases it requires. If both scores reach the full-disclosure price, any positive pooling premium makes full disclosure suboptimal. If one score falls short, the change in scrutiny must also compensate for admitting otherwise excluded work. A favorable information comparison within a fixed pool is therefore insufficient for the joint decision.

Expanding review access weakly lowers the full-disclosure admission price and these shortfalls, so once selective disclosure is strictly preferable, further expansion cannot restore full disclosure. Lower marginal congestion at every feasible volume has the same effect, as when the three-reviewer queue in Section 6.2 adds a fourth identical reviewer (Appendix B.1). Both comparisons fix work per admission and assign corrections to a separate resource. When the inspecting employee also repairs blocked proposals, capacity use depends on scrutiny and case composition, and Appendix B.4 gives an example in which expanding capacity instead restores full disclosure.

For proportional scrutiny $h(p) = \gamma p$, with $0 < \gamma \leq 1$, define

$$J_{ij} = -\gamma(1-\omega)(\pi_i - \pi_j)(v_i - v_j). \tag{35}$$

Only pairs with $J_{ij} > 0$ can improve on full disclosure. For such a pair, let $\bar{a}_{ij} = \max\{a_i^F, a_j^F\}$ and $\underline{a}_{ij} = \min\{a_i^F, a_j^F\}$ denote the higher and lower full-disclosure scores.

COROLLARY 2 (**Critical admission prices**). *Under proportional scrutiny, the critical price of a same-stratum pair with $J_{ij} > 0$ is*

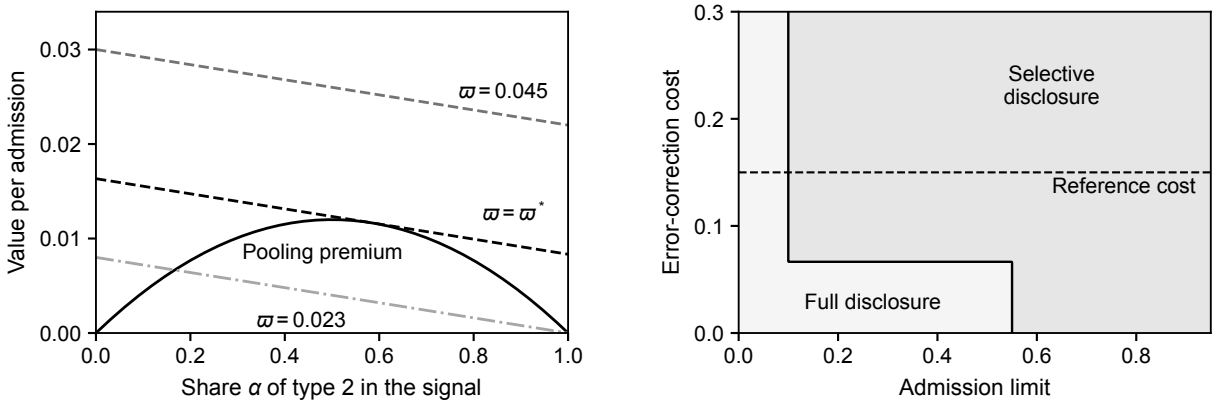
$$\varpi_{ij}^* = \begin{cases} \underline{a}_{ij} + J_{ij}, \\ \quad \text{if } J_{ij} \leq \bar{a}_{ij} - \underline{a}_{ij}, \\ \dfrac{\bar{a}_{ij} + \underline{a}_{ij}}{2} + \dfrac{J_{ij}}{4} + \dfrac{(\bar{a}_{ij} - \underline{a}_{ij})^2}{4J_{ij}}, \\ \quad \text{if } J_{ij} > \bar{a}_{ij} - \underline{a}_{ij}. \end{cases} \tag{36}$$

Full disclosure is optimal exactly when $\varpi_F^+ \geq \varpi_{ij}^*$ for every such pair. If no such pair exists, it is optimal at every admission limit.

Figure 2 illustrates why a positive pooling premium is insufficient. Two types in the same public stratum have interception probabilities 0.20 and 0.60 and net inspection values 0.20 and 0.08, and both have autonomous risk below the direct-handling cost. With $h(p) = p$ and a base admission cost of 0.025, their full-disclosure admission scores are 0.015 and 0.023. A third type, in a different

public stratum, has score 0.045 and accounts for 10% of arrivals, so it is the best use of a scarce review slot. Mixing the first two types improves their combined inspection value, but their premium cannot cover the opportunity cost of displacing this third type.

Once the admission limit exceeds that 10% share, the full-disclosure price falls below the pair's critical price, approximately 0.0313, and selective disclosure becomes strictly preferable. The employee's response and the two pooled types are unchanged; the firm can now obtain the mixture without displacing its most valuable work. Appendix B.2 gives the underlying loss and correction primitives. The example also shows why the theorem uses the highest compatible price: exactly at a 10% admission limit, that price can support full disclosure although a lower compatible price would fail the pair test. In panel (b), the access level at which selective disclosure pays moves with the error-correction charge, which changes both the ranking of admissions and the pooling benefit.



(a) Pooling and admission shortfalls **(b)** Optimal disclosure regimes

Figure 2 The Admission-Price Certificate in the Three-Type Example

Note. Panel (a) fixes $k^E = 0.15$; panel (b) varies correction cost and the admission limit, holding the other primitives fixed.

5.4. Public Information and Reviewer Incentives

Which pairings are feasible depends on what the reviewer already observes. In the refund workflow, the proposed payment and visible transaction fields define the public strata. If exact confidence and action type are public under the known stakes-blind gate, then $\pi_i = (1 - x_i)\eta_q(x_i; g_i)$ is known before scrutiny. Every additional disclosure then induces the same response, and full disclosure attains the joint optimum. This holds for nonlinear responses and heterogeneous correction costs when corrections use a separate resource (Appendix B.3). Public loss alone, by contrast, does not

ensure optimality of full disclosure: correction and false-block costs can reverse the ordering of interception probabilities and net inspection values within a loss stratum.

Under proportional scrutiny, agreement between the rankings of π_i and v_i within each public stratum makes every J_{ij} nonpositive, so full disclosure is optimal at every admission limit. At a fixed pool, the effect of an additional randomized refinement instead depends on the conditional covariance of posterior interception probabilities and net inspection values, and with nonlinear scrutiny, changes in mean inspection also matter (Appendix B.3).

An error-count rule uses verifiable catches; aligning rewards with net saving also requires valuing prevented losses and charging for false blocks. With quadratic effort and neither a standing checking benefit nor upper saturation, alignment makes refinements weakly beneficial for a fixed admitted pool, even with negative net values (Corollary EC.1). Announced case-specific bonuses can reveal stakes, whereas contingent pay can influence inspection through expected rewards without revealing them. Section 6.6 compares uniform bonuses with disclosure at equal cost, treating compensation as a separate design choice (Bastani and Cachon 2025).

The disclosure choice concerns only an internal routing assessment; the proposal and verification evidence remain available to the reviewer. In practice, the certificate evaluates a proposed review interface against optimized full disclosure, with admissions allowed to adjust under either policy. Applying this test requires more than aggregate accuracy: the firm needs interception probabilities, correction requirements, the scrutiny that an assignment induces, and the value of displaced work.

6. Policy Comparisons in a Refund Workflow

We evaluate the admission and disclosure decisions in a refund workflow and examine whether inspection responds to queue composition, as the theory assumes. Both exercises use public order data from Olist, a Brazilian e-commerce company that connects small merchants to online marketplaces (Olist 2018). Item prices and freight charges describe the payment at issue in a refund, but these records do not reveal whether a reviewer would inspect a proposal.

A supplementary experiment with 118 participants examines that response by varying queue error frequency at a fixed workload and reward schedule (Section 6.1). The policy comparisons that follow combine Olist's transaction heterogeneity with agent success rates from a public benchmark of AI customer-service agents and scenarios for correction costs and scrutiny.

6.1. Inspection in a Refund Task

The experiment asks whether the same reviewer inspects more when her queue contains more errors, holding queue length, rewards, and audit costs fixed. Its refund cases are built from sampled order

characteristics, with the refund requests, return status, and validity assigned by design. Participants supervised a simulated AI agent that proposed approving every refund. A refund was valid only if the return had been received, the request was within 30 days of delivery, and the amount did not exceed the customer's payment. Before seeing any individual amount or record, each participant chose how many of ten proposals to audit, and the system selected that many proposals at random. An audit revealed the supporting record, after which the participant released or blocked the proposal; unaudited proposals executed. Correct outcomes earned points and audits became increasingly costly, so a perfect checker would audit eight, four, and two proposals in the enriched (40% invalid), neutral (20%), and depleted (10%) queues, respectively. The task ended at approval or blocking and imposed no corrective work.

The 118 participants were working professionals from various industries enrolled in a part-time MBA program at a U.S. business school, and points counted toward the course grade. Each participant faced the three queues in counterbalanced order. The screen provided the routing information needed to calculate each queue's invalid rate, and for 61 randomly assigned participants it also stated the rates; the comparisons below pool both displays. Composition was thus conveyed before inspection, and learning from review experience was not separately varied. Appendix D gives the scoring, assignment, and inference procedures.

Participants inspected more in queues with more errors. The mean share audited was 0.578, 0.428, and 0.321 in the enriched, neutral, and depleted queues; the enriched-minus-depleted contrast was 0.257, with a 95% interval of $[0.195, 0.318]$.

Conditional on inspection, participants blocked 84.7% of invalid and 8.9% of valid proposals. Using these pooled detection and false-block rates, the score-maximizing continuous audit count is close to the observed mean in the enriched queue (5.7 versus 5.8), but far below it in the depleted queue (0.1 versus 3.2). Inspection exceeded this benchmark most when errors were rare; Figure EC.4 compares all three queues.

The higher share audited in error-rich queues supports the premise that scrutiny rises with error frequency, which gives admissions an attention effect in Proposition 3 and the policy comparisons. At the pooled detection rate, the queues span interception probabilities of approximately 0.085–0.339, over which the mean share audited is close to $0.25 + p$. This pattern motivates the positive-default response $h_+(p) = \min\{1, 0.25 + p\}$ used in the policy comparisons. Inspection at zero errors and at higher interception frequencies remains unmeasured, so Section 6.5 varies the response beyond the observed range.

6.2. Numerical Design

The policy comparisons extend this setting to a firm that chooses admissions and bears correction and congestion costs. We return to Olist's full population of 99,441 orders from 2016–2018, which links item prices and freight charges to customers, delivery records, and reviews. Aggregating item prices and freight by order and retaining positive matched item values gives 98,666 orders from 95,420 customers. A customer-disjoint split assigns 49,320 orders to construction and 49,346 to evaluation. Construction customers determine normalization and practical policy thresholds, which are then fixed for evaluation. An assessed-loss index combines item value and freight, both available before execution, with an action category: account update, refund, delivery promise, replacement, or escalation. Its weights and category multipliers specify the consequences of an inappropriate execution, with median loss normalized to twice the cost of independent handling. The reference environment assigns categories and quality independently of order attributes (relaxed in Section 6.7).

Historical retail tasks in τ-bench (Yao et al. 2025), a public benchmark for AI customer-service agents, provide reliability references. In its retail domain, a tool-using agent converses with a simulated customer under a written service policy, and a task succeeds when the resulting database state matches the task's goal. We classify 432 historical GPT-4o runs on 108 tasks by the final prescribed write action. Category success rates range from 0.40 to 0.63; two categories with fewer than twenty runs use assumed rates. We use these task-level success rates to set the reference scenario; uniform targets of 0.90 and 0.95 represent more reliable agents. The Gaussian gate in Section 3.3 generates conditional confidence and the probabilities of interception and false blocking.

All payoffs use normalized resolution value $R = 1$. Independent handling costs $d = 0.30$, including labor, delay, and residual error. The reference scenario charges $c_q = 0.025$ per admission, $k^E = 0.15$ per intercepted error, and $k^V = 0.03$ per false block, so correcting an error costs half as much as resolving the request independently. Inspection has net value

$$v_{q,i} = (L_i - k^E)\pi_{q,i} - k^V u_{q,i}. \tag{37}$$

These charges are scenario inputs; Section 6.3 varies the correction charge. Unless stated otherwise, monetary gains and losses below are per 1,000 arrivals. Because independent handling costs $d = 0.30$, a gain of 3 per 1,000 arrivals equals the cost of independently resolving 10 requests.

The review station is an Erlang-C queue with three unit-rate reviewers, arrival rate five, and delay cost 0.08; an admission limit of 0.54 gives utilization 0.90. Correction and independent handling initially use separate resources; Section 6.6 changes the correction assignment.

The comparisons use full inspection, proportional scrutiny $h(p) = p$, and the positive-default response h_+. Unlike the experiment, where audit counts were chosen before any case was seen, the numerical screens reveal amount before inspection and condition scrutiny on that information. Table EC.1 summarizes the sources and their roles.

6.3. The Value of Review

We begin by asking whether review improves on the best policy without it, and how much of that improvement practical admission rules capture. Without review, the firm chooses between autonomous execution and independent handling for each request. Following Lemma 1, the best such policy yields 801 per 1,000 arrivals. Full automation yields 575. Independent handling of every request yields 700, while a construction-fitted confidence-only guardrail that ignores stakes yields 756. All review policies use this best nonreview rule for requests they do not admit. Table 1 therefore measures the gain from checking a proposal relative to the better nonreview option.

Admission rules. All policies in Table 1 reveal the routing type before inspection. The net-value policy optimizes admissions using the inspection probability associated with each disclosed type and accounts for congestion. Practical rules admit requests whose amount or confidence lies in a fitted interval, or in both intervals, and a separate rule combines risk and confidence cutoffs. These thresholds are selected on construction customers and frozen for evaluation. Risk priority instead orders requests by autonomous expected loss and chooses its own evaluation-population volume.

Table 1 Admission Policies and Review Value

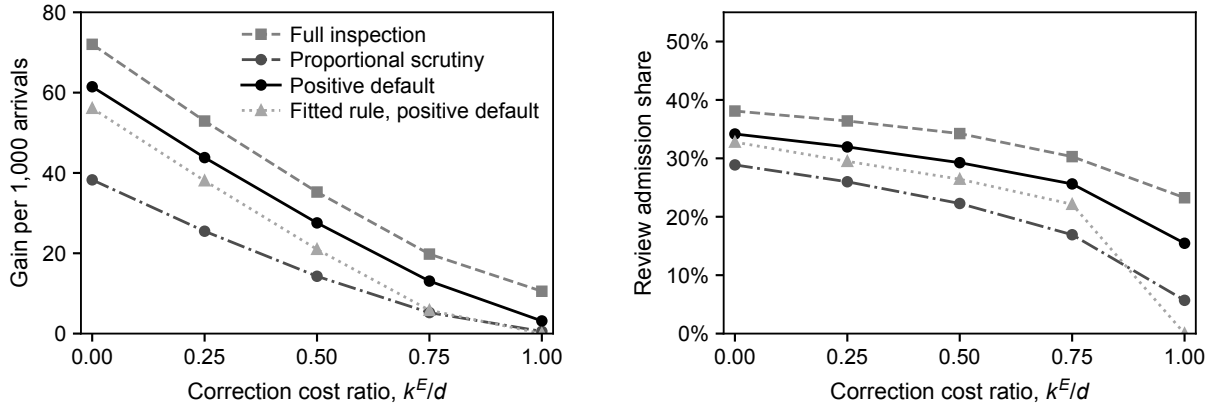
	Full inspection		Proportional		Positive default	
Policy	Gain	Review	Gain	Review	Gain	Review
Net-value admission	35.2	34.2	14.3	22.3	27.6	29.3
Net value, uncalibrated	28.9	29.1	13.4	19.4	25.0	25.3
Fitted monetary interval	16.9	29.9	1.1	15.3	6.4	25.0
Fitted confidence interval	23.2	30.3	6.2	14.9	20.4	25.2
Fitted amount and confidence	26.4	32.1	8.2	17.3	21.1	27.3
Fitted risk and confidence	24.3	32.8	9.0	18.2	21.0	26.4
Risk priority, chosen volume	14.0	36.8	0.0	0.0	< 0.1	< 0.1

Note. Gains are per 1,000 evaluation arrivals over the best nonreview policy. Review is the percentage of arrivals admitted to review. Full disclosure applies throughout. Screening is calibrated except in the indicated row. Practical thresholds use construction customers; the optimum and risk priority choose evaluation-population volume under known model probabilities.

Adding confidence to the monetary rule raises its share of optimized review value from 48% to 75% under full inspection and by larger margins under both discretionary responses (Table 1). Confidence helps distinguish proposals with similar amounts but different prospects for an effective intervention. The risk-and-confidence rule improves further under proportional scrutiny, but

does not dominate the amount-and-confidence rule across regimes. Risk priority performs worse, choosing no review under proportional scrutiny and almost none under positive-default scrutiny. It ranks the costliest errors first even when those requests are better handled directly. At the reference inputs, calibration adds about 6.3 under full inspection and less than 3 under either discretionary response, with admissions optimized for each gate.

Correction costs. These comparisons hold correction costs fixed; Figure 3 shows how the work following a block changes the return to review. As k^E/d rises from zero to one, optimized review value falls by 85% under full inspection, 95% with a positive default, and 99% under proportional scrutiny. Once correction costs approach the cost of independent handling, keeping the proposal adds little value.



(a) Value of review **(b)** Review admissions

Figure 3 Review Value and Admissions across Correction-Cost Scenarios

Note. Curves show model expectations on 49,346 evaluation orders.

At reference reliability, the largest saving from review comes from requests that no longer require independent handling (Table 2). Review preserves the agent's work on these requests, although correction, admission, and congestion absorb much of that saving. Under proportional scrutiny, AI execution loss rises although fewer inappropriate proposals execute, so those that do are costlier on average. Payoff gains and protection against execution loss are therefore distinct outcomes.

Capacity. Congestion also changes which otherwise valuable checks are worth performing. A rule admitting every case with positive value before congestion, rationing only at the admission limit, earns 64.0 less than the best nonreview policy under full inspection and 1.83 less under positive-default scrutiny. Under proportional scrutiny, the same rule applied to that regime's values still gains 12.4.

Table 2 Sources of Review Value

	Full inspection	Proportional	Positive default
Avoided independent handling	83.2	43.1	66.4
Corrections and false blocks	−34.7	−21.0	−31.7
Admission	−8.6	−5.6	−7.3
Congestion	−6.8	−1.1	−3.4
Avoided AI execution loss	2.1	−1.1	3.7
Net gain	35.2	14.3	27.6

Note. Per 1,000 arrivals over the best nonreview policy, using Table 1's optimized admissions. Totals use unrounded values.

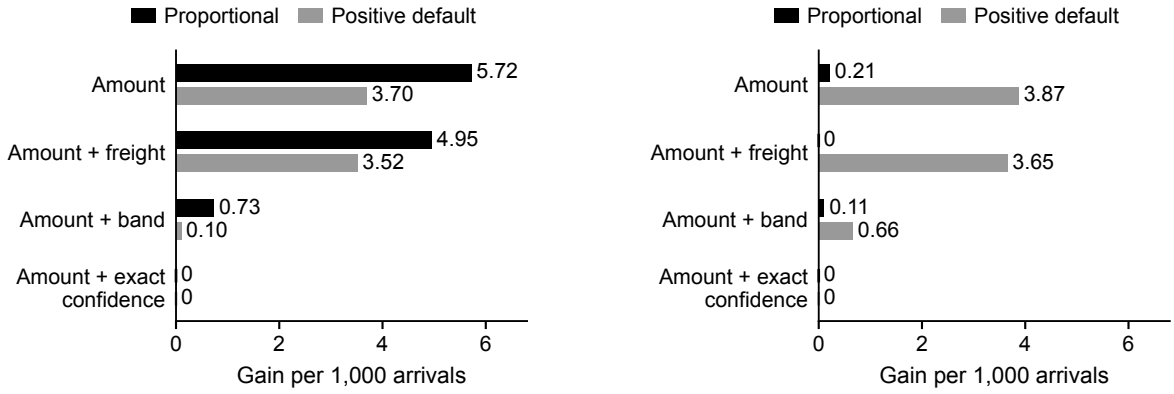
Marginal congestion sets the admission price because the reference admission limit is slack. For the staffing comparison, we represent transaction heterogeneity by 9,600 types using 120 confidence intervals (Appendix C.7). A fourth reviewer then adds 9.55 under full inspection, 1.10 under proportional scrutiny, and 3.91 under positive-default scrutiny, before staffing expense. Staffing adds less value under discretionary scrutiny because extra capacity does not ensure inspection.

Reliability. Higher reliability reduces review's value most sharply when inspection is discretionary. At reliability 0.95, the proportional and positive-default policies retain only 6–24% of optimized full-inspection value, compared with 40–78% at reference reliability. The source of full-inspection value also changes. Raising each action category's reliability to 0.90 and 0.95 lowers its gain to 17.7 and 10.1, respectively, with avoided execution loss accounting for 68% and 91% of the remainder (Appendix C.5). At higher reliability, preventing losses accounts for more of full-inspection value, but discretionary inspection captures a smaller share of that value.

6.4. Attention at a Fixed Workload

We next isolate the effect of admissions on reviewer attention while holding workloads fixed. In Table 1, full disclosure makes scrutiny depend only on a case's own type. With less informative screens, an admission can also affect inspection of other cases that look alike to the reviewer (Proposition 3). A first policy optimizes admissions anticipating scrutiny within each group of requests that share public information. A second keeps every group's workload but reassigns its cases as if scrutiny would stay at the first policy's rates. Each assignment is evaluated at the scrutiny its own composition induces, with identical base charges, congestion, and capacity use.

These comparisons use a finite representation of the transaction heterogeneity: 960 types crossing four amount levels and four freight levels with five action categories and twelve confidence states. Gains on this population therefore differ from the order-level values in Table 1, but all policies within each comparison use the same population. Appendix C.4 gives the construction and bounds.



(a) Admission at common workloads **(b)** Increment over full disclosure

Figure 4 Composition and Disclosure Gains at the Reference Correction Cost

Note. Gains are relative to (a) reassignment at matched group workloads and (b) optimized full disclosure.

At a fixed workload, the value of anticipating attention depends on what the reviewer sees. With only the action and amount visible and proportional scrutiny, choosing admissions to anticipate attention yields 12.3, or 5.72 more than the reassignment. Showing a confidence band as well reduces the difference to 0.735 (Figure 4(a)).

Under positive-default scrutiny, the gain from anticipating attention is smaller at the reference correction ratio but larger at $k^E/d = 0.75$, where reassignment lowers scrutiny from 99.5% to 81.5% and reduces gain from 14.8 to 2.34. This 84% loss occurs at fixed group workloads, because individually attractive replacements weaken the attention that supports other cases.

Under proportional scrutiny, optimized full disclosure gains 13.5, more than anticipatory admission with only the action and amount visible. Under positive-default scrutiny, withholding the remaining assessment instead gains 30.6, against 27.1 for full disclosure, and the ranking is the same at the 0.75 correction ratio. The preferred screen therefore depends on the inspection response, even when admissions are optimized for each screen.

6.5. The Incremental Value of Disclosure

The preceding comparison fixes the screen and lets admissions adjust. We now let the firm choose signals within each public stratum, as in Theorem 1, to measure the gain over optimized full disclosure. Under proportional scrutiny, full disclosure captures over 98% of the best joint-design gain at the reference inputs. With the action and amount visible, optimizing signals and admissions adds 0.215 to the full-disclosure gain of 13.5; a visible confidence band reduces the increment to 0.107 (Figure 4(b)).

Under positive-default scrutiny, pooling can also raise average inspection when a mixture spans the point where the response reaches its cap. For a concave response, pooling weakly raises mean scrutiny relative to revealing its types, although net values and admission sacrifices still determine the firm's gain. At the reference cost, joint design adds 3.87, or 14.3% of full-disclosure review value, and at $k^E/d = 0.75$ it adds 2.56, or 20.3%.

Revealing freight alongside action and amount leaves most of the reference premium, whereas revealing a confidence band removes most of it (Figure 4(b)); visible monetary consequences and interception prospects thus restrict useful mixtures differently. The premium remains near 3.9 as hidden confidence is refined with public information fixed, but falls as more confidence is revealed (Table EC.2). With corrections assigned to a separate resource, making exact confidence public eliminates the gain from selective disclosure.

Because the experiment does not establish whether or where inspection reaches a cap, we compare ten responses that agree with $0.25 + p$ through $p = 0.35$ and differ above it. Selective disclosure still adds value when inspection approaches its cap smoothly or levels off below full inspection. Delaying the rise in inspection leaves much smaller gains (Table EC.4). Choosing a policy for the wrong response can reverse this ranking. With the action and amount visible and both policies held at their h_+ designs, joint design loses 2.92 against full disclosure under a smooth continuation without a plateau and earns less than no review under proportional scrutiny (Appendix C.5). The firm should measure scrutiny at the interception frequencies its proposed assignments would create.

Each policy's incremental gain sets a ceiling on its additional implementation cost. On the 960-type population of Section 6.4, optimized full-disclosure admission under positive-default scrutiny gains 6.80 over the fitted risk-and-confidence rule; selective disclosure adds 3.87. Appendix C.6 examines deployment costs and screening-estimation uncertainty.

6.6. When Reviewers Repair Errors

The comparisons so far assign corrections to a separate resource, so a checker can reject a proposal without repairing it. A reviewer who must repair each error she finds anticipates that burden when deciding whether to inspect, and her repairs then consume the review station's time.

On the 9,600-type population at $k^E/d = 0.75$, jointly optimized admissions and disclosure under positive-default scrutiny yield 15.24 with separate correction and 5.13 when reviewers perform repairs. The incentive effect alone removes more than half the gain; with both channels, scrutiny falls from 97% to 77% and admissions roughly halve. Repairs make capacity use depend on scrutiny,

allowing the capacity comparison in Theorem 1 to reverse (Appendix B.4). Appendix C.5 separates the two channels.

Pay offers an alternative to changing what reviewers see. We compare uniform bonuses for verified corrections with selective disclosure at equal operating cost, including payments and workload. With the action and amount visible, full disclosure with a bonus yields 6.81 at the strongest tested pay response, against 5.13 from selective disclosure alone; allowing both raises the gain to 8.96. When reviewers respond less strongly to pay, bonuses neither outperform selective disclosure nor add much to it. At the strongest tested pay response, compensated full disclosure also outperforms selective disclosure alone at reliability 0.90 and 0.95 (Table EC.5). Appendix C.5 describes the assumed pay responses and bonus levels.

6.7. Alternative Inputs

The reference environment makes quality independent of order attributes. We now vary their association to ask whether monetary rules work better when larger orders are more error-prone, holding the loss mapping and marginal quality fixed and refitting rules on construction customers. Under strong positive dependence, the amount rule's share of full-inspection value rises from 48% to 72% at the reference correction cost, but it chooses no review under proportional scrutiny and captures 27% with a positive default; adding confidence raises all three shares (Table EC.6). Positive dependence leaves low-amount requests with little incentive for inspection and makes many large, error-prone requests better handled directly; confidence separates worthwhile cases within these ranges. Across the dependence scenarios, the positive-default disclosure premium ranges from 2.13 to 3.87 with action and amount visible.

Alternative reliability and loss inputs change the scale of the gains. A second historical agent profile reduces full-disclosure review gains from 14.3 to 7.8 under proportional scrutiny and from 27.6 to 17.9 with a positive default. In a separate scenario, positive loss–error association reduces the proportional gain to 8.9. Optimized review value stays positive, although the relative performance of fitted rules varies across inputs (Appendix C.7). When correction costs differ among requests that look alike, the comparison in Appendix C.5 gives a larger gain from cost-aware admission than from selective disclosure.

7. Conclusion and Managerial Implications

This paper studies how a firm should allocate execution authority when review preserves an AI agent's proposal, inspection is discretionary, and blocked work requires correction. Under full

inspection, review is worthwhile when the loss it prevents covers correction and admission costs and any advantage of independent handling. Neither confidence nor expected loss alone identifies these requests (Lemma 1 and Proposition 1). When inspection is discretionary, admissions also change attention to cases that look alike. Full disclosure is optimal exactly when no feasible mixture gains more from changing attention than the admission value its cases sacrifice at the full-disclosure admission price (Theorem 1). At a fixed station, expanding access can favor withholding when work per admission is common and correction is separate. Assigning repairs to reviewers can reverse this direction because scrutiny then changes how much capacity an admission consumes.

The results suggest three implications for firms that place human checks on AI agents. First, the decision to offer review should begin with what a correction costs relative to starting over. At reference reliability, most of review's value comes from avoiding independent handling, and little remains when the two costs are nearly equal. With more reliable agents, avoided execution loss accounts for more of full-inspection value. Second, admission rules should account for a request's own net value and its effect on attention to cases the reviewer cannot distinguish from it. Adding confidence improves monetary approval rules across the tested amount–error relationships at reference reliability, while assignments that ignore attention lose much of a fixed workload's value. Third, assigning repairs to the checker discourages inspection and consumes capacity. Bonuses for verified corrections partly restore value at the strongest tested pay response. Separate correction has its own labor cost, which the comparison should include.

For the information decision, optimized full disclosure is the practical benchmark. Withholding adds little in the reference proportional-scrutiny comparisons; the larger gains under the positive-default response depend on inspection at error frequencies beyond the experiment's range. A firm considering withholding should measure inspection near the frequencies its assignments would create and compare withholding with full disclosure and feasible compensation, optimizing admissions for each policy.

The baseline model uses a common scrutiny response and a fixed screening rule, with reviewers who know the committed policies. Future work could let reviewers learn queue composition or let the firm jointly choose screening, compensation, and admissions. Field experiments could vary composition at fixed public information and review volume, repair assignments, and bonuses, while recording correction time. These comparisons could test whether inspection reaches a cap and separate incentive from workload effects. The mechanism is also relevant to maker–checker processes in which inspection is discretionary and assignments shape the errors employees expect

to find. As AI agents take on more of the maker's role, we expect sustaining attention to rare errors to remain an important problem for service operations.

References

Agarwal N, Moehring A, Rajpurkar P, Salz T (2023) Combining human expertise with artificial intelligence: Experimental evidence from radiology. Working Paper 31422, National Bureau of Economic Research.

Aghion P, Tirole J (1997) Formal and real authority in organizations. *Journal of Political Economy* 105(1):1–29.

Anunrojwong J, Iyer K, Manshadi V (2023) Information design for congested social services: Optimal need-based persuasion. *Management Science* 69(7):3778–3796.

Athey SC, Bryan KA, Gans JS (2020) The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings* 110:80–84.

Baker GP (1992) Incentive contracts and performance measurement. *Journal of Political Economy* 100(3):598–614.

Balakrishnan M, Ferreira KJ, Tong J (2026) Human-algorithm collaboration with private information: Naïve advice-weighting behavior and mitigation. *Management Science* 72(1):265–284.

Bastani H, Bastani O, Sinchaisri WP (2026) Improving human sequential decision making with reinforcement learning. *Management Science* 72(1):733–755.

Bastani H, Cachon GP (2025) The human-AI contracting paradox. Available at SSRN 5962739.

Blackwell D (1953) Equivalent comparisons of experiments. *The Annals of Mathematical Statistics* 24(2):265–272.

Bondi E, Koster R, Sheahan H, Chadwick M, Bachrach Y, Cemgil T, Paquet U, Dvijotham K (2022) Role of human–AI interaction in selective prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(5):5286–5294.

Boyacı T, Canyakmaz C, de Véricourt F (2024) Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science* 70(2):1258–1275.

Boyacı T, Canyakmaz C, de Véricourt F (2026) Beyond the black box: Unraveling the role of explainability in human-artificial intelligence collaboration. *Management Science* Articles in Advance.

Brynjolfsson E, Li D, Raymond L (2025) Generative AI at work. *The Quarterly Journal of Economics* 140(2):889–942.

Dai T, Singh S (2025) Using artificial intelligence as gatekeeper or second opinion: Designing patient pathways for artificial intelligence augmented healthcare. *Production and Operations Management* OnlineFirst.

Dai T, Taylor T (2026) Designing enterprise AI systems: Hallucination, creativity, and moral hazard. Available at SSRN 5996714.

de Véricourt F, Gurkan H (2026) Is your machine better than you? You may never know. *Management Science* 72(1):558–574.

Dell'Acqua F (2022) Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters. Working paper.

Dessein W (2002) Authority and communication in organizations. *The Review of Economic Studies* 69(4):811–838.

Dreyfuss B, Hoong R (2026) Calibrated coarsening: Designing information for AI-assisted decisions. Available at SSRN 5286198.

Epping GP, Caplin A, Duhaime E, Holmes WR, Martin D, Trueblood JS (2026) Managing cognitive bias in human labeling operations for rare-event AI: Evidence from a field experiment. arXiv preprint arXiv:2603.11511.

European Union (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Friedman HL (2026) Adding an AI agent to a principal-agent model: Implications for task assignment, compensation, and review. Available at SSRN 7321300.

Grand-Clément J, Pauphilet J (2026) The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science* 72(1):667–692.

Guan X, Qi A, Wang S (2025) Why the best machine may not be the best: Incentivizing human-machine collaboration. Available at SSRN 5283571.

Hathaway BA, Kagan E, Dada M (2023) The gatekeeper's dilemma: "When should I transfer this customer?". *Operations Research* 71(3):843–859.

Holmström B, Milgrom P (1991) Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7:24–52.

Ibanez MR, Toffel MW (2020) How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management Science* 66(6):2396–2416.

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Kamenica E, Gentzkow M (2011) Bayesian persuasion. *American Economic Review* 101(6):2590–2615.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

Kolotilin A, Corrao R, Wolitzky A (2025) Persuasion and matching: Optimal productive transport. *Journal of Political Economy* 133(4):1334–1381.

Legros B, de Véricourt F (2026) Human-AI interaction in congested service systems: When it improves performance, and when it doesn't. Available at SSRN 5401875.

Lin R, Zhang J (2026) When to screen, when to bypass: LLM-judges in resource-scarce AI-human workflow. arXiv preprint arXiv:2603.13870.

Lindsay GF, Bishop AB (1964) Allocation of screening inspection effort: A dynamic-programming approach. *Management Science* 10(2):342–352.

Lingenbrink D, Iyer K (2019) Optimal signaling mechanisms in unobservable queues. *Operations Research* 67(5):1397–1416.

Lipnowski E, Mathevet L, Wei D (2020) Attention management. *American Economic Review: Insights* 2(1):17–32.

Liu Y, Sinchaisri WP (2026) Strategic advice in the age of personal AI. arXiv preprint arXiv:2603.02055.

Lu T, Tomlin B (2026) Augmenting the operations manager with a prediction machine. Available at SSRN 5282711.

Moffatt v Air Canada (2024) 2024 BCCRT 149. Civil Resolution Tribunal of British Columbia.

Mozannar H, Sontag D (2020) Consistent estimators for learning to defer to an expert. *International Conference on Machine Learning*, 7076–7087 (PMLR).

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Noti G, Donahue K, Kleinberg J, Oren S (2025) AI-assisted decision making with human learning. *Proceedings of the 26th ACM Conference on Economics and Computation*, 38–39.

Olist (2018) Brazilian e-commerce public dataset by Olist. Kaggle.

Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: An attentional integration. *Human Factors* 52(3):381–410.

Rayo L, Segal I (2010) Optimal information disclosure. *Journal of Political Economy* 118(5):949–987.

Shumsky RA, Pinker EJ (2003) Gatekeepers and referrals in services. *Management Science* 49(7):839–856.

Snyder C, Keppler S, Leider S (2026) Algorithm reliance: Fast and slow. *Management Science* 72(1):368–385.

Vaccaro M, Almaatouq A, Malone T (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8(12):2293–2303.

Wolfe JM, Brunelli DN, Rubinstein J, Horowitz TS (2013) Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of Vision* 13(3):33.

Wolfe JM, Horowitz TS, Kenner NM (2005) Rare items often missed in visual searches. *Nature* 435(7041):439–440.

Wolfe JM, Van Wert MJ (2010) Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology* 20(2):121–124.

Yang G, Li C, Roet-Green R, Xu F (2026) When better AI hurts: Verification, compliance, and AI-human handoffs in service systems. Available at SSRN 6975018.

Yang W (2026) Algorithmic advice, human review, and shared liability. Available at SSRN 6457618.

Yao S, Shinn N, Razavi P, Narasimhan K (2025) τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *International Conference on Learning Representations*.

Electronic Companion to "Act, Ask, or Defer? Designing Human Oversight for Agentic AI"

Appendix A: Proofs for Section 4

Proof of Lemma 1. Subtracting the best nonreview payoff, $R - \min\{r_i, d_i\}$, from the full-inspection payoff gives $v_{q,i} - c_q - (r_i - d_i)^+$. For $r_i > 0$, substitute $v_{q,i} = r_i\eta_{q,i} - \kappa_{q,i}$ and divide by r_i. At $r_i = 0$, interception is impossible, $v_{q,i} = -k_i^V u_{q,i} \le 0$, and autonomous execution strictly dominates the two costly human modes. Equality elsewhere gives a tie with the best nonreview mode. □

For constant expected correction cost, put $a = c_q + \bar{\kappa}_q$. The frontier is a/r below d and $1 - (d-a)/r$ above d. If $0 < a < d$, the two derivatives have opposite signs, giving a unique minimum a/d. With constant effectiveness $\beta \in (0,1)$, review is strictly preferred on $(a/\beta, (d-a)/(1-\beta))$, which is nonempty exactly when $a < \beta d$. At $\beta = 0$ there is no strict review region; at $\beta = 1$ and $a < d$ it is $r > a$. Fixed scrutiny e replaces $v_{q,i}$ by $ev_{q,i}$ in the payoff comparison. Replacing request-specific costs by population averages does not justify these reductions.

Proof of Proposition 1. Fix g. Strict monotonicity of confidence makes conditioning on x equivalent to conditioning on s_A. Conditional independence of other routing attributes gives $\rho_q(x, L; g) = r\psi_q(x; g)$, where $\psi_q = \Pr(\widehat{\theta}_q \ge 0 \mid \theta < 0, x, g)$. Let $p_q(\theta, s) = \Pr(\widehat{\theta}_q(s, s_H; g) \ge 0 \mid \theta, s, g)$. The additive signal and monotone gate make this probability nondecreasing in both arguments. The monotone likelihood ratio order of $\theta \mid s, g$ survives restriction to $\theta < 0$ and implies stochastic dominance there. Thus, for $s' > s$,

$$\begin{aligned}
&\mathbb{E}[p_q(\theta, s') \mid \theta < 0, s', g] \\
&\quad \ge \mathbb{E}[p_q(\theta, s) \mid \theta < 0, s', g] \\
&\quad \ge \mathbb{E}[p_q(\theta, s) \mid \theta < 0, s, g].
\end{aligned}$$

Hence $\eta_q = 1 - \psi_q$ is nonincreasing. At common r, d, the difference in net review value is $r(\eta_{q,2} - \eta_{q,1}) - (\kappa_{q,2} - \kappa_{q,1})$. Its sign proves the equivalence and the sufficient cost condition. □

Proof of Corollary 1. At fixed x, correction cost is independent of L. Net review value is $r\eta_q(x) - \kappa_q(x)$ for $r \le d$ and $d - r(1 - \eta_q(x)) - \kappa_q(x)$ for $r \ge d$. Since $0 < \eta_q(x) < 1$, its unique maximum is $H_q(x)$ at $r = d$. Solving the two strict inequalities against c_q gives the stated interval and its existence condition. A continuous strict downward crossing of H_q gives the ceiling. If effectiveness is nonincreasing and correction cost nondecreasing, the lower endpoint is nondecreasing and the upper endpoint nonincreasing on the eligible region. At the crossing both equal d. Intersecting with attainable risks proves the support qualification. □

Confidence-band example. Let $d = k^E = 0.3$, $c_q = 0.03$, $k^V = 0$, and $\eta_q(x) = 1 - x$. Then $\kappa_q(x) = 0.3(1-x)^2$ and $H_q(x) = 0.3x(1-x)$, so review is feasible exactly for $(1 - \sqrt{0.6})/2 < x < (1 + \sqrt{0.6})/2$. At $r = d$, net review value rises on $(0, 1/2)$ although effectiveness falls. This curve satisfies the anchored-gate assumptions: take equiprobable $\theta \in \{-1, 1\}$ and independent standard normal signal errors. Then $x(s_A) = (1 + e^{-2s_A})^{-1}$. Approve when $s_H \ge \Phi^{-1}(1 - x(s_A)) - 1$; this gate is increasing in both signals and blocks an erroneous proposal with probability $1 - x$. Along r fixed, $L = r/(1-x)$ adds no information beyond s_A. With $k^V = 0.3$, net review value still tends to zero at both confidence endpoints and exceeds c_q at $x = 1/2$, where it equals $0.3\{1/4 - \Phi(-2)/2\}$.

Stakes-dependent gates. Along $L = r/(1-x)$, define $\mathcal{D}_r f = \partial_x f + L\partial_L f/(1-x)$. Differentiating at fixed r, d gives the exact condition $r\mathcal{D}_r\eta_q - \mathcal{D}_r\kappa_q \le 0$ for nonincreasing net review value. With a stakes-blind gate and constant event charges, the derivative is $\{r - k^E(1-x)\}\eta_q'(x) + k^E\eta_q(x) - k^V u_q'(x)$. Lower error incidence can therefore offset declining effectiveness; unconditional false blocking can also rise as appropriate proposals become more frequent.

Calibration. The assessments $(1-w)s_A + w(s_H - \mu)$ and $(1-w)s_A + ws_H$ have mean-squared errors $(1-w)^2\tau^2 + w^2\sigma^2$ and $(1-w)^2\tau^2 + w^2(\sigma^2+\mu^2)$. Their minimizing human weights are $w_q = \tau^2/(\tau^2+\sigma^2)$ and $w_{q'} = \tau^2/(\tau^2 + \sigma^2+\mu^2)$. These minimize assessment error, not the firm's economic loss. Approval thresholds are $T_q = \mu - \sigma^2 s_A/\tau^2$ and $T_{q'} = -(\sigma^2+\mu^2)s_A/\tau^2$. Thus $T_q \ge T_{q'}$ exactly when $\mu + \mu^2 s_A/\tau^2 \ge 0$; a higher threshold weakly increases both interception and false blocking. Subtracting gate payoffs gives (21). At fixed scrutiny the difference in net activation values is multiplied by e; with endogenous scrutiny it becomes $e_q v_q - e_{q'} v_{q'} - (c_q - c_{q'})$.

Appendix B: Proofs and Extensions for Section 5

Proof of Proposition 2. Continuity on the compact feasible set gives existence. At fixed volume, exchanging a lower-score admitted type for a higher-score available type increases value. Maximum gross value is therefore concave and piecewise linear in volume; strictly convex congestion gives a unique optimal volume. The concave program's first-order conditions give the price, sorting, and complementarity conditions. At zero the right derivative is $\max_i a_i^I - C'(0)$, proving the no-review criterion. Ties can yield several allocations at the same volume. □

Proof of Proposition 3. At zero, every noncongestion term is bounded by a constant times admitted mass, so the objective is continuous on the compact feasible set. For $M > 0$, $\partial p(y)/\partial y_i = (\pi_i - p(y))/M$. Differentiation gives $\partial_i F = \widetilde{a}_i(y) - C'(M)$; the box and capacity normal-cone conditions give the stated admission price. When $\omega = 1$ or h is constant, the objective is linear minus convex congestion, making these conditions sufficient. □

Sufficiency fails in general. Take equally common types with $\pi = (0.3, 0.9)$, $L_i = 1$, perfect conditional detection, $d = 0.3$, $c_q = 0.05$, $k^E = 0.01$, $k^V = 0$, $h(p) = p$, $\omega = 0$, $\bar{M} = 0.5$, and $C(M) = 0.01M^2$. At the low-error pure pool the gradient is $(0.0291, -0.2145)$, strictly descending in every nonzero feasible direction $(-s, t)$ with $s \ge t \ge 0$. Its gain, approximately 0.0171, is below the high-error pure pool's 0.0735.

B.1. The admission-price certificate

Any feasible posterior collection is implementable: admit type i with probability y_i/w_i and, conditional on admission, send s with probability $m_s\vartheta_{s,i}/y_i$. Zero-admission types require no signal. Permissible posteriors stay within one public stratum.

Pairwise representation and existence. Within a public stratum, fix a posterior mean p. The polytope

$$\{\vartheta \ge 0 : \mathbf{1}^\top\vartheta = 1,\ \pi^\top\vartheta = p\} \tag{EC.1}$$

has vertices supported on at most two types. On this polytope, $h(p)$ is constant and $\Gamma(\vartheta)$ is linear. Decomposing a posterior into vertices therefore preserves each type's admitted mass and total value. This argument permits signed v_i and arbitrary continuous h.

The decomposition shows that the cone $\mathcal{K}$ generated by $(\vartheta, \Gamma(\vartheta))$ for permissible posteriors is unchanged when their support is restricted to at most two types. These permissible pair generators form a compact set, and each has unit admission mass. A convergent sequence in the cone has bounded total generator mass, so compactness of the generators' convex hull proves that the cone is closed. With $y_i \le w_i$, feasible aggregate vectors are bounded and hence compact. The joint optimum exists. Conic Carathéodory in $\mathbb{R}^{n+1}$ then represents an optimal aggregate vector with at most $n+1$ pair generators, giving some optimal policy with that many signals.

Proof of Theorem 1. The full-disclosure problem maximizes $(a^F)^\top y - C(\mathbf{1}^\top y)$ over the availability and admission constraints. Its optimal mass M_F is unique. A compatible price satisfies $\varpi = C'(M_F) + \nu$, volume complementarity, and the sorting conditions

$$\begin{aligned} a_i^F &\geq \varpi \quad \text{if } y_i = w_i, \\ a_i^F &= \varpi \quad \text{if } 0 < y_i < w_i, \\ a_i^F &\leq \varpi \quad \text{if } y_i = 0. \end{aligned}$$

At a binding limit, some type is admitted and some type is not fully admitted because $0 < \bar{M} < 1$. The compatible prices form a nonempty bounded closed interval. At a slack limit, the only price is $C'(M_F)$. Thus ϖ_F^+ exists, including when $M_F = 0$.

For sufficiency, take a compatible price ϖ satisfying the pair inequalities and set $\lambda_i = (a_i^F - \varpi)^+$. Since

$$a_i^F + (\varpi - a_i^F)^+ = \varpi + (a_i^F - \varpi)^+,$$

the inequalities imply $\Gamma(\vartheta) \leq \varpi + \lambda^\top \vartheta$ for pair posteriors and, by decomposition, every permissible posterior. Any joint policy consequently has value at most

$$\varpi M + \lambda^\top w - C(M).$$

Convexity and volume complementarity give

$$\varpi M - C(M) \leq \varpi M_F - C(M_F).$$

The full-disclosure allocation attains the resulting bound: $\lambda_i > 0$ only for a fully admitted type, and each positively admitted type satisfies $a_i^F = \varpi + \lambda_i$. This proves sufficiency at ϖ_F^+.

For necessity, write the joint problem as a concave maximization of $G - C(\mathbf{1}^\top y)$ over $(y, G) \in \mathcal{K}$ and the linear availability and admission constraints. A relative-interior point of the cone with positive mass in every type can be scaled to make these constraints strict. Such a point exists because all singleton generators are available and every w_i is positive. Relative Slater duality gives multipliers $\lambda_i \geq 0$ and $\nu \geq 0$. If full disclosure is optimal, the supporting inequality is

$$\Gamma(\vartheta) \leq \varpi + \lambda^\top \vartheta, \qquad \varpi = C'(M_F) + \nu.$$

Availability complementarity and equality on admitted singleton signals imply $\lambda_i = (a_i^F - \varpi)^+$. Indeed, a type with unused availability has zero multiplier, and an admitted singleton must attain equality. Hence this price is compatible with the full-disclosure optimum. Applying the inequality to each pair gives the certificate at ϖ. Its right side is nondecreasing in price, so it also holds at ϖ_F^+. Failure of any pair inequality there therefore implies a strict improvement over full disclosure.

Finally, order the full-disclosure scores. While the admission limit binds below the unconstrained optimal volume, the highest compatible price is the score of the last admitted type and weakly decreases with the limit. At an exact availability breakpoint it is the higher of the adjacent marginal scores. Once the limit is slack, optimal volume and price remain fixed. At the transition, marginal congestion lies between the relevant adjacent scores, so the price cannot jump upward. Pair premiums do not change with the limit, while each admission shortfall weakly falls. The limits satisfying all pair inequalities form a lower interval. The value difference itself may be nonmonotone in the admission limit.

The price also weakly falls when $\widetilde{C}' \le C'$ at every feasible volume. Optimal full-disclosure volume weakly rises because $C - \widetilde{C}$ is nondecreasing. If volume rises, a type whose admission increases has a score no greater than the old price and no less than the new one. At unchanged volume, slackness gives the derivative ordering; at a binding limit, the highest compatible price is the same marginal score, including availability breakpoints. The pair premiums are unchanged, so the certificate has the same direction. For the reference Erlang-C station, $C_k(M) = 0.016 N_k(5M)$, where $N_3(a) = a^4/[(3-a)(a^2+4a+6)]$ and $N_4(a) = a^5/[(4-a)(a^3+6a^2+18a+24)]$. Differentiation gives $N_4'(a) \le N_3'(a)$ for $0 \le a < 3$. □

Proof of Corollary 2. For proportional scrutiny, $\Delta_{ij}(\alpha) = J_{ij}\alpha(1-\alpha)$. Pairs with $J_{ij} \le 0$ always satisfy the certificate. For $J_{ij} > 0$, dividing by $\alpha(1-\alpha)$ and minimizing over the interior gives

$$J_{ij} \le \{\sqrt{\delta_i(\varpi)} + \sqrt{\delta_j(\varpi)}\}^2, \tag{EC.2}$$

using $\inf_{0<\alpha<1}\{\delta_i/\alpha + \delta_j/(1-\alpha)\} = (\sqrt{\delta_i} + \sqrt{\delta_j})^2$. Zero shortfalls follow by limits. Write $\bar{a} = \max\{a_i^F, a_j^F\}$ and $\underline{a} = \min\{a_i^F, a_j^F\}$. For $\underline{a} \le \varpi \le \bar{a}$, the right side is $\varpi - \underline{a}$. Thus the threshold is $\underline{a} + J_{ij}$ if $J_{ij} \le \bar{a} - \underline{a}$. Otherwise both shortfalls are positive at the threshold. Solving

$$\sqrt{\varpi - \bar{a}} + \sqrt{\varpi - \underline{a}} = \sqrt{J_{ij}}$$

gives $(\bar{a} + \underline{a})/2 + J_{ij}/4 + (\bar{a} - \underline{a})^2/(4J_{ij})$. The expression is valid for signed admission scores. Taking the largest threshold over permissible positive-premium pairs proves the corollary. □

B.2. The three-type example

Figure 2 uses $w = (0.45, 0.45, 0.10)$, $\pi = (0.20, 0.60, 0.40)$, $v = (0.20, 0.08, 0.175)$, effectiveness 0.8, $c_q = 0.025$, $d = 0.30$, $k^E = 0.15$, $k^V = 0$, $h(p) = p$, and $C(M) = 0.005M^2$. Set $x_i = 1 - \pi_i/0.8$ and $L_i = 0.15 + v_i/\pi_i$. This gives $r = (0.2875, 0.2125, 0.29375) < d$, $t_i = -0.025$, and $a^F = (0.015, 0.023, 0.045)$. Only types 1 and 2 share public information. Since $C'(M) < 0.015$ for $M < 1$, full disclosure fills its limit in order 3, 2, 1. Its highest price is 0.045 through $\bar{M} = 0.10$, then 0.023 through 0.55, and 0.015 thereafter. The feasible pair has $J_{12} = 0.048$ and critical price $0.019 + 0.012 + 0.008^2/(4 \cdot 0.048) = 0.031\overline{3}$. Full disclosure is therefore optimal exactly through $\bar{M} = 0.10$, including the boundary.

In panel (b), varying k^E at fixed losses and probabilities gives $a^F = (0.021 - 0.04k^E, 0.077 - 0.36k^E, 0.069 - 0.16k^E)$ and $J_{12} = 0.024 + 0.16k^E$. The same highest-price test gives each regime. At $k^E = 0$, types 2 and 3 precede type 1, and the full-disclosure region ends at their combined share 0.55.

B.3. Further disclosure results

Fix positive admitted masses. Write $p_{\mathcal{S}} = \mathbb{E}_y[\pi_i \mid Z, \mathcal{S}]$ and $n_{\mathcal{S}} = \mathbb{E}_y[v_i \mid Z, \mathcal{S}]$. The disclosure-dependent payoff is $(1-\omega)M\mathbb{E}_y[h(p_{\mathcal{S}})n_{\mathcal{S}}]$. A refinement is a finite randomized experiment from which the coarser signal is obtained by type-independent garbling within public strata (Blackwell 1953). The characterization below uses convexity in posterior beliefs.

PROPOSITION EC.1 **(Information, alignment, and response curvature).** *Suppose $0 \le \omega < 1$ and h is twice continuously differentiable with $h' > 0$ on each stratum's attainable interception interval. Every randomized refinement weakly raises fixed-pool firm value if and only if, in each stratum with nonconstant π_i, constants ξ_0, ξ_1 satisfy*

$$v_i = \xi_0 + \xi_1 \pi_i, \qquad (\xi_0 + \xi_1 p)h''(p) + 2\xi_1 h'(p) \ge 0 \tag{EC.3}$$

for all types and attainable p. Constant π_i makes disclosure irrelevant, as does $\omega = 1$.

Proof. For $f(\vartheta) = h(\pi^\top\vartheta)v^\top\vartheta$, convexity is sufficient by conditional Jensen. It is necessary at a prior positive on the admitted support: any interior posterior can occur with small positive probability and then split into a nearby mean-preserving pair. For $\mathbf{1}^\top a = 0$, $p = \pi^\top\vartheta$, and $n = v^\top\vartheta$,

$$D^2 f[a,a] = nh''(p)(\pi^\top a)^2 + 2h'(p)(\pi^\top a)(v^\top a).$$

If the tangent-space projections of π and v are independent, choose $\pi^\top a = 1$ and $v^\top a = -T$; large T violates convexity. Thus $v = \xi_0\mathbf{1} + \xi_1\pi$, reducing convexity to (EC.3). Continuity handles endpoints. Constant π makes f linear. □

The condition is sufficient, but need not be necessary, for deterministic partitions alone. More generally, a refinement changes payoff divided by $(1-\omega)M$ by

$$\mathbb{E}_y[\mathrm{Cov}_y(h(p_{\mathcal{S}'}), n_{\mathcal{S}'} \mid Z, \mathcal{S})] + \mathbb{E}_y[n_{\mathcal{S}}\{\mathbb{E}_y[h(p_{\mathcal{S}'}) \mid Z, \mathcal{S}] - h(p_{\mathcal{S}})\}].$$

This follows by adding and subtracting the product of the conditional means and using iterated expectations. Proportional scrutiny eliminates the second term. Concavity makes its bracket nonpositive; the covariance and the sign of net value still determine the firm's overall comparison.

Full disclosure under $h(p) = a + bp$, $b \geq 0$, gains $b(1-\omega)M\mathbb{E}_y[\mathrm{Cov}_y(\pi_i, v_i \mid Z, \mathcal{S})]$ over any signal. Same ordering within strata makes this nonnegative because $2\,\mathrm{Cov}(\pi, v) = \mathbb{E}[(\pi_i - \pi_j)(v_i - v_j)]$ for independent posterior draws. This proves the full-disclosure comparison for each admission rule and after optimization. If π_i is public, all posteriors share that interception probability, and disclosure has no effect for any continuous h. Exact (x_i, g_i) identifies π_i under the stakes-blind gate. Public loss alone does not: with $L = 1$, equiprobable $\pi = (0.2, 0.8)$, $k^E = (0, 1)$, and no false blocks, proportional scrutiny gives benefits 0.02 under revelation and 0.05 under pooling.

COROLLARY EC.1 **(Net-value-aligned scrutiny).** *Suppose the reviewer values net avoided loss and has effort cost $\iota_H e^2/2$, where $\iota_H > 0$ and $\iota_H \geq \max_i v_i^+$. Every randomized refinement weakly raises fixed-pool value, allowing negative v_i. If all admitted values are nonnegative, its gain is $(1-\omega)M\mathbb{E}_y[\mathrm{Var}_y(n_{\mathcal{S}'} \mid Z, \mathcal{S})]/\iota_H$. The optimized comparison holds when experiments and their garbling relation are fixed across feasible admissions.*

Proof. The reviewer chooses $[n/\iota_H]_{[0,1]}$. Excluding upper saturation makes the firm's benefit $(n^+)^2/\iota_H$, convex even at negative means. Conditional Jensen proves the comparison; for nonnegative values the conditional-variance identity gives the formula. Pointwise dominance survives admission optimization. □

Without the upper-saturation restriction, equiprobable net values $(0.5, 1.5)$ and $\iota_H = 1$ give pooling benefit 1 versus 0.875 under revelation. Positive correction costs can implement these values using perfect conditional detection, $\pi = (0.5, 0.75)$, $k^E = 0.02$, and $r_i = v_i + 0.02\pi_i$.

Positive default. Maximizing $(a_H + b_H p)e - \iota_H e^2/2$ over $[0,1]$ gives (14). The continuous capped response satisfies Theorem 1; Proposition 3 applies away from its kink. Corollary EC.1 uses its own reviewer objective, with no standing checking benefit.

B.4. When reviewers also perform corrections

A reviewer who must repair the errors she finds has a different incentive to inspect. Let $a_i^E, a_i^V \geq 0$ be her burdens from correcting an error and resolving a false block. Quadratic effort cost gives $e(\vartheta) = [a_H/\iota_H + v^\top\vartheta]_{[0,1]}$, where $v_i = \{(b_H - a_i^E)\pi_i - a_i^V u_i\}/\iota_H$. The firm observes the routing type, and the burden schedule is common knowledge.

Attributes visible before scrutiny belong to Z. A larger interception probability need not raise scrutiny when finding an error creates more work than its reward compensates.

An admission uses $s_0 > 0$ units of reviewer time; inspection adds $s_I \geq 0$ and the two blocking events add $s_i^E, s_i^V \geq 0$. Put $\tau_i = s_I + s_i^E \pi_i + s_i^V u_i$. A posterior has value and work $\Gamma_H(\vartheta) = t^\top \vartheta + e(\vartheta) v^\top \vartheta$ and $w_H(\vartheta) = s_0 + e(\vartheta)\tau^\top \vartheta$. The firm maximizes $\sum_s m_s \Gamma_H(\vartheta_s) - C(T)$ subject to availability and $T = \sum_s m_s w_H(\vartheta_s) \leq \bar{T}$, with convex differentiable nondecreasing C. Monetary event charges exclude this workload penalty; employee burdens are not charged again to the firm. Here $\omega = 0$ and $\bar{T} > 0$.

PROPOSITION EC.2 (**The price of shared reviewer time**). *Let $A_i = t_i + e_i^F v_i$, $W_i = s_0 + e_i^F \tau_i$, and $e_i^F = e(\mathbf{e}_i)$. Full disclosure ranks A_i/W_i. It is jointly optimal if and only if some compatible work price λ satisfies, for every permissible pair posterior,*

$$\Gamma_H(\vartheta) - \lambda w_H(\vartheta) \leq \sum_i \vartheta_i (A_i - \lambda W_i)^+. \tag{EC.4}$$

Compatibility requires $\lambda = C'(T_F) + \eta$, $\eta \geq 0$, $\eta(\bar{T} - T_F) = 0$, and the full-disclosure sorting conditions. Public knowledge of v_i eliminates disclosure value.

Proof. At fixed $v^\top \vartheta$, both value and work are affine in the posterior. Its decomposition into vertices supported on at most two types preserves admitted masses, value, and work. The cone generated by $(\vartheta, \Gamma_H, w_H)$ is closed by compactness and unit admission mass. Feasible aggregate vectors are compact. A relative-interior point containing every singleton can be scaled to satisfy availability and work constraints strictly. Concave-program duality therefore applies. At a full-disclosure optimum, stationarity gives the compatible work price; availability complementarity and equality on admitted singletons give multipliers $\mu_i = (A_i - \lambda W_i)^+$. The dual posterior inequality is (EC.4), proving necessity. Conversely, summing that inequality bounds every policy by $\mu^\top w + \lambda T - C(T)$. Compatibility makes T_F maximize the last two terms over $[0, \bar{T}]$, and full disclosure attains the bound. If v_i is public, scrutiny is constant within every permissible posterior, so revelation preserves value and work. □

Unlike Theorem 1, the highest compatible price need not certify optimality. For equiprobable types, set $\pi = \nu = (0.2, 0.8)$, $v = (0.49, 0.50)$, $t_i = -0.001$, $\tau = (0, 3)$, $s_0 = 1$, $a_H = 0$, and $C(T) = 10^{-4} T^2$. At $\bar{T} = 2.2$, full disclosure admits both types and is optimal: $\lambda = 0.00044$ satisfies the certificate. The highest compatible price, 0.097, fails it at an equal mixture. At $\bar{T} = 0.54$, an equal mixture strictly outperforms full disclosure. Thus increasing capacity can restore full disclosure.

Appendix C: Numerical Design and Evaluation

C.1. Requests and screening

Requiring a matched positive item value excludes 775 of 99,441 orders, including 164 cancellations. The retained population contains 461 cancellations and has complete customer identifiers. Construction data determine 1st/99th-percentile winsorization and normalization of item value V and freight F, whose medians are 86.90 and 17.16 reais. Write the normalized values as V_n, F_n. Independent categories (account update, refund, delivery promise, replacement, escalation) have shares $(0.30, 0.22, 0.20, 0.15, 0.13)$ and loss multipliers $m = (0.40, 1.20, 1.00, 1.10, 1.40)$. The loss index is

$$L_i = a_L m_{g_i} (0.05 + 0.55 V_{n,i} + 0.15 F_{n,i}), \tag{EC.5}$$

Table EC.1 Sources and Roles of the Evaluation Inputs

Input	Source	Role in the comparison
Order characteristics	Olist transactions	Experimental case records; monetary heterogeneity and public states
Inspection choices	Refund-desk experiment	Evidence on response to assigned error frequencies
Agent reliability	Historical task success; assumed sparse-category and higher-reliability targets	Reference quality levels
Confidence and screening	Gaussian specification	Conditional correctness, interception, and false blocks
Losses, costs, and service	Specified mappings and operating scenarios	Consequences, correction economics, and congestion

where a_L sets construction median loss to 0.60. Evaluation uses the same transformation. The weights are assumed consequences, not measured error costs. Ratings and delivery outcomes are excluded from this routing-time index.

For category g, let $\theta \sim N(\Phi^{-1}(\bar{a}_g), 1)$, $\tau^2 = 1$, $\sigma^2 = 0.5$, and $\mu = 0.8$. Conditional on s_A, its mean is m_A and variance $v_A = 0.5$, giving $x = \Phi(m_A/\sqrt{v_A})$. Calibrated and uncalibrated gates use human weights $2/3$ and $1/2.14$, respectively, with only the former removing the shift. If the conditional assessment has mean m_q and standard deviation s_q, and $\chi = \sqrt{v_A/(v_A+\sigma^2)}$, then

$$\begin{aligned} \pi_q &= \Phi_2\left(-m_A/\sqrt{v_A}, -m_q/s_q; \chi\right), \\ u_q &= \Phi(-m_q/s_q) - \pi_q. \end{aligned} \tag{EC.6}$$

Deterministic quadrature evaluates the bivariate probability. Net values may be negative.

The reference category targets are $(0.467, 0.632, 0.500, 0.620, 0.400)$, from historical GPT-4o terminal retail-task success across 432 runs of 108 tasks. Targets 0.500 and 0.400 are assumed for categories with fewer than twenty runs. Each task is classified by its final required state-changing action; repeated runs of a task do not independently verify action correctness. Uniform targets 0.90 and 0.95 supply the high-reliability scenarios.

C.2. Accounting and admission rules

For admission probability z_i and scrutiny e_i, define $M = \mathbb{E}[z_i]$, $I = \mathbb{E}[z_i e_i]$, $B_E = \mathbb{E}[z_i e_i \pi_i]$, $B_V = \mathbb{E}[z_i e_i u_i]$, and $D = \mathbb{E}[(1-z_i)\mathbf{1}\{r_i > d\}]$. Expected AI execution loss and total payoff are

$$\begin{aligned} Q &= \mathbb{E}[(1-z_i)\mathbf{1}\{r_i \le d\} r_i + z_i(r_i - e_i L_i \pi_i)], \\ U &= 1 - Q - dD - c_q M - k^E B_E - k^V B_V - C(M). \end{aligned} \tag{EC.7}$$

To obtain inappropriate-execution frequency from Q, hold the authority decisions and routing indicators fixed and replace the loss terms r_i and $L_i\pi_i$ by $1 - x_i$ and π_i, respectively. Losses that remain after direct handling or correction are included in those costs. For the $M/M/3$ station, independent admission marks, exponential slot times, and delay charge 0.08 give

$$C(M) = \frac{0.016a^4}{(3-a)(a^2+4a+6)}, \qquad a = 5M. \tag{EC.8}$$

The formula holds on $0 \le M \le 0.54$ and equals $0.08MW_q(M)$ for the Erlang-C delay W_q. Base and event costs exclude station congestion.

For fixed scrutiny, sort $a_i = t_i + e_i v_i$ and maximize cumulative value less congestion, permitting fractional marginal admission and zero review. Risk priority instead sorts r_i and evaluates all prefix endpoints and stationary points. Amount

and confidence intervals end at construction 5th percentiles or are unbounded; the combined rule searches all pairs. The risk-and-confidence rule optimizes confidence prefixes at 60 risk quantiles from the fifth to hundredth percentile, the direct-handling boundary, and an unrestricted cutoff, permitting fractional marginal mass. Thresholds stay fixed in evaluation, and excess workload is rationed uniformly. The confidence-only guardrail minimizes construction handling and execution cost over confidence cutoffs. Every review rule retains the risk-informed nonreview fallback.

The gain over the best nonreview policy decomposes as

$$U - U_0 = d(D_0 - D) + (Q_0 - Q) \\ - c_q M - k^E B_E - k^V B_V - C(M),$$

where D_0, Q_0 are that policy's direct-handling share and AI execution loss. The decomposition uses the 49,346 evaluation orders at every reliability target. Direct-handling cost includes labor, delay, and residual error cost, so its saving is not a separate estimate of labor productivity. At reliability 0.90 and 0.95, avoided execution loss accounts for 24% and 45% of positive-default review value, and 68% and 91% under full inspection; under proportional scrutiny the share is negative.

C.3. Public information and numerical bounds

Construction quartiles define four mean-valued levels each of item value and freight; evaluation orders determine their joint frequencies. Crossing these states with five independent action categories and twelve equiprobable AI-signal intervals yields 960 types. Conditional x, π, u are integrated separately within each signal interval, and losses use (EC.5). Each amount recurs across quality and freight states and therefore cannot identify confidence. The no-review payoff is 0.79655. Action-and-amount signals remain within those public levels; adding freight or a confidence band further restricts them. Six public bands each combine two hidden confidence states.

To separate approximation from observability, retain the four amount levels, four freight levels, their joint frequencies, and the five categories. Increase the number of equiprobable signal intervals from 12 to 24 and 48, and to 96 at reference reliability and $k^E/d = 0.5$. Conditional x, π, u are integrated within each interval. The public screen always reveals action and amount; it either adds no confidence information or retains the same six signal-quantile bands. Thus hidden refinement changes neither visible monetary states nor public confidence thresholds. Both correction ratios and all three reliability profiles are evaluated.

Public-resolution comparisons instead fix 48 hidden intervals and vary the public confidence partition over 1, 2, 6, 12, 24, and 48 nested bands. One band adds no confidence information; 48 reveals the discrete confidence state. Type values, availability and the full-disclosure benchmark are identical across these comparisons. Finer public information restricts feasible signals, so the optimized joint value weakly decreases. Table EC.2 reports the reference-cost results. The bounds below certify optimization on each finite population; they do not bound approximation error relative to continuous types.

At reliability 0.90 and 0.95, uniform refinement to 96 and 192 states still changes the positive-default premium. To resolve rare low-confidence cases, augment the 48-interval grid with quantiles $2^{-j}/48$, $j = 1, \ldots, 12$, giving 60 intervals, then bisect each interval to obtain 120. Interval weights and conditional screening probabilities are integrated under the same signal model; public thresholds remain fixed. Table EC.3 shows that the twelve-state grid greatly understates these premiums. At reference correction cost, the 120-interval positive-default gains are 0.300 at reliability

Table EC.2 Hidden-State Refinement and Public Information

	Proportional		Positive default	
	No band	Six bands	No band	Six bands
Hidden confidence states; public screens fixed				
12	0.215	0.107	3.872	0.658
24	0.251	0.155	3.847	0.632
48	0.268	0.174	3.918	0.702
96	0.271	0.178	3.917	0.701

Public confidence bands; 48 hidden states fixed		
	Proportional	Positive default
1	0.268	3.918
2	0.265	3.594
6	0.174	0.702
12	0.107	0.260
24	0.052	0.089
48	0.000	0.000

Note. Increment over optimized full disclosure, per 1,000 arrivals. Reference reliability and $k^E/d = 0.5$; action and amount remain public.

Table EC.3 Disclosure Value and Rare Confidence States

Reliability	k^E/d	12 uniform	48 uniform	192 uniform	120 tail-refined
0.90	0.50	0.004	0.040	0.302	0.300
0.90	0.75	0.007	0.046	0.259	0.251
0.95	0.50	0.000	0.006	0.020	0.074
0.95	0.75	0.000	0.009	0.021	0.067

Note. Positive-default increment over optimized full disclosure, per 1,000 arrivals. Action and amount remain public; no confidence band is visible.

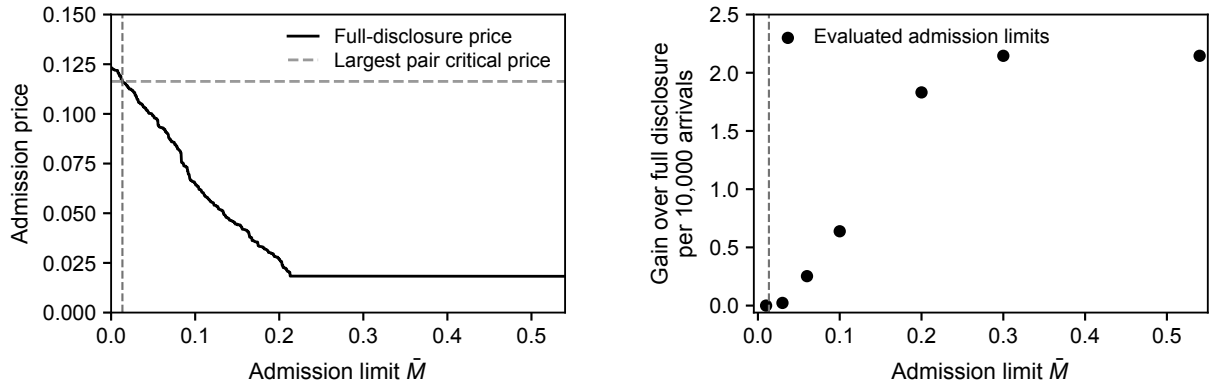
0.90 and 0.074 at 0.95, compared with 0.283 and 0.069 on the 60-interval grid. Six public confidence bands leave nearly the same gains because the relevant rare states remain in their lowest band. These gains are smaller than at reference reliability but remain positive as the grid is refined.

Theorem 1 permits pair signals. For general continuous scrutiny, retain separate posteriors on each pair. Let U_{LP} solve a restricted signal-mass linear program with supporting tangents below C, and let ε bound every positive reduced posterior value under its dual prices. Then

$$U^* \le U_{\mathrm{LP}} + \bar{M}\varepsilon. \tag{EC.9}$$

Adding ε per admission covers every signal in the dual; a feasible policy evaluated under exact congestion gives a lower bound. For piecewise-quadratic h, reduced pair value is cubic on each response segment; endpoints and feasible real roots of its quadratic derivative give exact pricing, including signed values and equal interception probabilities. Subdivision at response breakpoints covers the whole pair. The largest gap in the tail comparisons is below 5×10^{-8} per arrival.

With public action and amount under proportional scrutiny, the largest pair critical price is 0.11635. The full-disclosure price remains at or above it through admission limit 0.01345 and then falls below it. Figure EC.1 holds the station fixed and distinguishes this optimality boundary from the resulting premium.



(a) Full-disclosure admission price **(b)** Disclosure premium

Figure EC.1 Admission Prices and Disclosure Premiums as Access to the Fixed Station Expands

C.4. Admission-induced attention

For public group G, write $M_G = \sum_{i\in G} y_i$, $P_G = \sum_{i\in G} \pi_i y_i$, and $V_G = \sum_{i\in G} v_i y_i$. With one signal per group, the objective is $F_{\mathcal{G}}(y) = \sum_i t_i y_i + \sum_G h(P_G/M_G)V_G - C(\sum_G M_G)$, with zero contribution from empty groups. Let y^A maximize this over availability and total admission constraints. The matched allocation y^N maximizes $\sum_G \sum_{i\in G}(t_i + e_G^A v_i)y_i$ subject to $\sum_{i\in G} y_i = M_G^A$, where $e_G^A = h(P_G^A/M_G^A)$. It sorts own contributions within each group; evaluation uses its induced e_G^N. Equal group masses cancel admission and congestion costs. Since y^N maximizes own contributions at fixed scrutiny, any advantage of y^A comes from the scrutiny its composition sustains. If π_i is constant within groups, the policies have equal value. The comparison holds workloads fixed and does not model repeated adjustment.

For global bounds, let $Q_G(\lambda)$ be each group's optimal value after charging λ per admission. Summing these values and $\max_{0\le M\le \bar{M}}\{\lambda M - C(M)\}$ gives an upper bound. Linear upper envelopes for $P_G V_G/M_G$ on posterior-mean intervals yield group bounds; subdivision at 0.75 handles h_+. Refinement gives gaps below 10^{-6} per arrival. With 4,320 types under proportional scrutiny, composition gains are 5.70 without a band and 0.670 with one (5.72 and 0.735 on 960 types). At reliability 0.95, refinement raises the increment from zero to 0.036.

C.5. Correction costs and response uncertainty

Across $c_q \in \{0.01, 0.025, 0.05\}$, $k^E/d \in \{0, 0.25, 0.5, 1\}$, and $k^V/d \in \{0, 0.1, 0.5\}$, proportional review gains range from zero to 43.4 per 1,000 arrivals. Limiting corrections to five per 100 arrivals reduces the reference gain to 6.74. At reliabilities 0.90 and 0.95, proportional gains are 1.50 and 0.581, and positive-default gains are 5.43 and 2.45.

Costs within a public screen. Each of the 960 types is split into two cost types with charges $k_i^{E,\pm} = \bar{k}^E(1 \pm \Delta)$, where $\Delta \in \{0.25, 0.50\}$ and $\bar{k}^E/d \in \{0.50, 0.75\}$. Within each action–amount–confidence-band stratum H, let $\bar{\pi}_H$ be the availability-weighted mean and $D_H = \max_{j\in H} |\pi_j - \bar{\pi}_H|$. Set

$$\Pr(+ \mid i) = \tfrac{1}{2}\{1 + \zeta(\pi_i - \bar{\pi}_H)/D_H\},$$

$$\zeta \in \{-0.8, 0, 0.8\},$$

using $1/2$ if $D_H = 0$. Both charges have equal aggregate mass within H. This preserves the mean charge and all original request characteristics under both public screens; association changes event-weighted costs. The firm knows cost type; the reviewer initially does not. Screening, scrutiny, and separate correction resources are fixed, with $k^V = 0.03$.

At $\Delta = 0.50$ and the 0.75 ratio under h_+, cost-aware admission gains 7.51–8.90 across associations; disclosure adds 2.28–3.43 without a band and 0.113–0.597 with one. The independent split gives 8.22, 2.93, and 0.494 per 1,000 arrivals.

On the twelve-state confidence grid used for this cost-heterogeneity comparison, premiums at reliability 0.90 and 0.95 stay below 0.031 per 1,000 arrivals across the tested spreads, associations, responses, and screens. Revealing cost category at reference reliability and $\Delta = 0.50$ weakly reduces every premium. Exact confidence eliminates it throughout.

The tail and compensation comparisons use the 120-interval population in Appendix C.3, with public thresholds fixed. The ten response scenarios equal $0.25 + p$ through $a = 0.35$, covering the experimentally observed interception range. Above a, the cap-location family is $\min\{1, 0.60 + 0.40(p - a)/(s - a)\}$ for $s \in \{0.60, 0.75, 0.90, 1, 1.25\}$; changing s changes slope as well as cap location. The rounded family retains $0.25 + p$ below $p_- = 0.75 - w$, uses $1 - (p_+ - p)^2/(4w)$ between p_- and $p_+ = 0.75 + w$, and equals one above p_+, for $w \in \{0.05, 0.10, 0.20\}$. Two smooth continuations use $z = p - a$: $0.60 + z - z^2/1.60$ throughout $p > a$, or $0.60 + z - z^2$ through 0.85 followed by ceiling 0.85. Both are continuously differentiable and concave; neither reaches full inspection on $[0, 1]$.

Table EC.4 Disclosure Gains across Response Tails

	$k^E/d = 0.50$		$k^E/d = 0.75$	
Response	Amount	With band	Amount	With band
Full scrutiny at 0.60	4.305	0.241	3.978	0.183
Reference cap at 0.75	3.917	0.701	2.638	0.476
Full scrutiny at 0.90	1.611	0.871	1.016	0.532
Full scrutiny at 1.00	0.229	0.145	0.190	0.105
Linear continuation, $s = 1.25$	0.171	0.081	0.151	0.062
Rounded, $w = 0.05$	3.111	0.680	2.025	0.431
Rounded, $w = 0.10$	2.447	0.643	1.520	0.392
Rounded, $w = 0.20$	1.325	0.478	0.834	0.302
Smooth without a plateau	0.409	0.185	0.303	0.131
Smooth ceiling at 0.85	0.961	0.156	0.460	0.082

Note. Increment over reoptimized full disclosure per 1,000 arrivals, on 9,600 types. Action and amount are public; With band adds public confidence bands.

With both policies reoptimized across the ten tails, two correction ratios, and two public screens, increments range from 0.061 to 1.051 at reliability 0.90 and from 0.027 to 0.295 at 0.95.

Response misspecification. Both policies designed under h_+ are frozen and evaluated under each alternative with correct beliefs about their assignment. At the reference correction ratio, the smooth continuation without a plateau gives joint-design increments of −2.92 without a band and −0.98 with one against frozen full disclosure, versus −4.54 and −2.60 against reoptimized full disclosure. A 0.85 ceiling raises the latter losses to 10.33 and 10.76; proportional scrutiny, to 19.26 and 11.39. Without a public confidence band, the proportional evaluation earns less than no review.

Shared reviewer time. Set $b_H = \iota_H = s_0 = 1$, $s_I = 0$, $a^E = \beta$, $a^V = 0.2\beta$, $s^E = \ell$, and $s^V = 0.2\ell$. Work replaces admissions in the limit 0.54 and in (EC.8), now a specified convex penalty rather than an exact queueing formula for variable service times. At reference reliability, $k^E/d = 0.75$, and $a_H = 0.25$, review gains are 6.94 with burden alone ($\beta = 0.25$), 9.13 with time alone ($\ell = 1$), and 5.13 with both. Scrutiny falls from 97% with separate resources

to 77% with both; admissions fall from 27% to 14%. Higher-burden and duty-heterogeneity variants use the original twelve-state grid: at $\beta = 1.25$, no review is optimal; an independent 50% cost split scales burdens and durations by 0.5 or 1.5, so premiums with exact confidence and hidden duties stay below 0.011; revealing duties removes them.

Compensation at equal operating cost. An additional uniform bonus b per verified error correction changes the scrutiny driver to $\nu_i(b) = \nu_i + \upsilon b \pi_i$, where υ converts firm-value units to normalized employee utility. Verification is assumed feasible within event costs; screening and repair durations stay fixed. Let I_i be inspected mass and $f_i(b) = (k_i^E + b)\pi_i + k_i^V u_i$. With M, D as defined above, operating cost is

$$K = dD + c_q M + C(T) + \sum_i I_i f_i(b). \tag{EC.10}$$

We minimize execution loss with the nonreview fallback fixed, subject to $K \le \mathcal{B}$, availability, public information, and $T \le 0.54$. Each screen's $\mathcal{B}$ equals the no-bonus joint optimum's expenditure; neither system incurs a fixed deployment fee. The reported policies exhaust this budget. At fixed b, pair decomposition preserves costs, work, and loss; posterior pricing includes budget and work prices.

Table EC.5 Compensation and Disclosure at Equal Cost

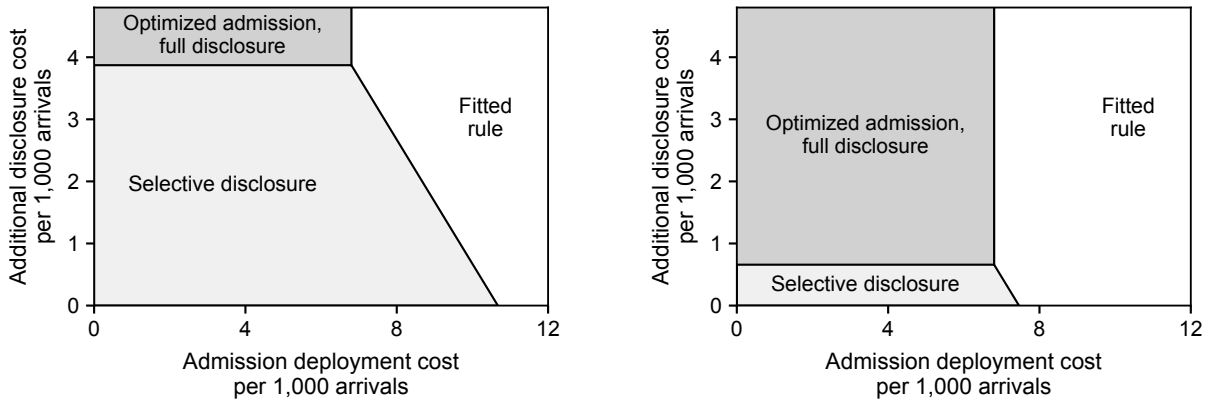
		Action and amount			Add confidence band		
Reliability	υ	Selective	Full + bonus	Both	Selective	Full + bonus	Both
Reference	1	5.130	4.993	5.130	5.058	4.993	5.058
Reference	5	5.130	5.003	5.244	5.058	5.003	5.174
Reference	20	5.130	6.812	8.964	5.058	6.811	7.123
0.90	20	2.085	5.539	8.067	2.085	5.539	7.389
0.95	20	1.114	3.562	5.054	1.114	3.562	4.771

Note. Gains over no review per 1,000 arrivals; $k^E/d = 0.75$, $\beta = 0.25$, $\ell = 1$, and $a_H = 0.25$. Expenditure is matched within each screen and reliability level. Selective uses no bonus; Both allows disclosure and a bonus.

We test $b = r\beta/\upsilon$ for $r \in \{0, 1/8, \ldots, 1, 1.25, 1.5, 2, 3, 4, 6, 8, 16\}$ and $\upsilon \in \{1, 5, 20\}$, at both correction ratios and all reliability levels, with $\beta = 0.25$ and $\ell = 1$. The conversion rates are scenario inputs; reported bonuses maximize this finite menu. At reference reliability, full disclosure with the best tested bonus overtakes selective disclosure alone at a conversion between 6.474 and 6.483 with action and amount public, and between 5.546 and 5.555 when a confidence band is also public. At reliability 0.95 and conversion 20, the best joint bonus is the largest tested value; the table does not establish an interior optimum.

C.6. Deployment and screening estimates

Let G_P, G_F, and $G_J(s)$ denote gains over the best nonreview policy from the fitted risk-and-confidence rule, optimized full disclosure, and joint design, respectively. Deployment chooses the largest of $0, G_P, G_F - f_A, G_J(s) - f_A - f_S$, where f_A is the added deployment cost of optimized admission over the fitted rule and f_S is the added cost of selective disclosure, both per arrival. A per-review charge s replaces t_i by $t_i - s$ in joint design. At $s = 0$, the three boundaries are $f_A = G_F - G_P$, $f_S = G_J - G_F$, and $f_A + f_S = G_J - G_P$. These give Figure EC.2. Reoptimizing at each s yields positive-default break-even charges per review of 0.01281 without a band and 0.00230 with one, falling to 0.00965 and 0.00155 at $k^E/d = 0.75$. The costs exclude extra service time.



(a) Public action and amount **(b)** A confidence band is also public

Figure EC.2 Policy Regions after Deployment Costs

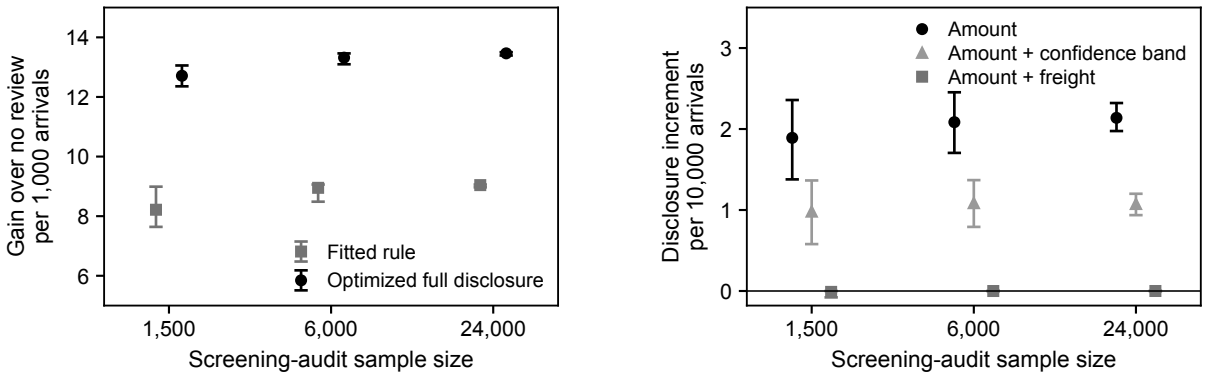
Note. Positive-default scrutiny on 960 types at reference correction costs.

Screening estimation fixes $h(p) = p$, with known confidence, losses, type frequencies, and costs. In each of 60 action/confidence strata, $n \in \{25, 100, 400\}$ independent audits have four outcome probabilities $(\pi_i, 1 - x_i - \pi_i, u_i, x_i - u_i)$. If N_E, N_V count invalid/valid proposals and B_E, B_V their blocked counts, use $\widehat{\pi}_i = (1-x_i)(B_E + 1/2)/(N_E + 1)$ and $\widehat{u}_i = x_i(B_V + 1/2)/(N_V + 1)$. The conditional blocking rates have independent Beta$(1/2, 1/2)$ priors; these expressions scale their posterior means by the known probabilities of invalid and valid proposals. Policies share the estimates, reselect admissions and signals, and are evaluated by (31) with $\omega = 0$ under true probabilities. The fitted risk-and-confidence rule is refitted on the same risk grid and confidence cutoffs. Only the designer's screening estimates are uncertain; h and reviewer beliefs are known.

Figure EC.3 uses 200 independent synthetic samples per size. With a band and charge $f_S = 0.100$ per 1,000 arrivals, adoption based on $\widehat{G}_J - \widehat{G}_F > f_S$ occurs in 1%, 54%, and 84% of samples; mean net gains are approximately zero, 0.0059, and 0.0068. Exact public confidence reproduces fitted full disclosure in every sample. At $k^E = 0.1875$, reoptimization gains 9.00 versus 8.67 for joint design frozen at $k^E = 0.15$; admission error can exceed the disclosure premium.

C.7. Amount–error dependence and alternative inputs

For the order-level comparison, form normal scores z_i from within-category amount ranks, separately on construction and evaluation customers. For $\rho \in \{-0.8, -0.4, 0.4, 0.8\}$, sort orders by $-\rho z_i + \sqrt{1-\rho^2}\epsilon_i$, with independent standard normal ϵ_i, and assign complete quality tuples in increasing-confidence order. A tuple contains confidence and both event probabilities for every gate. The construction preserves their empirical joint distribution within each category and sample. Zero retains the reference assignment. Positive ρ associates larger amounts with more errors; evaluation rank correlations between amount and error probability range from -0.747 to 0.751 at reference reliability. Correctness and screening outcomes conditional on the routing type are unchanged, so amount adds no further correctness information once that type is disclosed. All loss indices, customer assignments, and category frequencies remain fixed. Practical thresholds are refitted before evaluation.



(a) Admission with estimated screening **(b)** Additional disclosure value

Figure EC.3 Estimated Screening and Disclosure Value by Audit Sample Size

Note. Means and central 95% ranges across 200 synthetic screening-audit samples per size.

Table EC.6 reports captured review value, using each scenario's optimized full-disclosure gain as denominator. Adding confidence to the fitted amount rule raises captured value in every reference-reliability comparison, including $k^E/d = 0.75$. This need not hold out of sample: at reliability 0.95, three proportional comparisons give small negative combined-rule gains where the fitted amount rule selects no review. The largest loss is 0.0011 per 1,000 arrivals.

Table EC.6 Amount–Error Dependence and Fitted Admission Rules

	Full inspection		Proportional		Positive default	
ρ	Amount	Both	Amount	Both	Amount	Both
-0.8	51.2	63.9	38.8	52.6	45.4	63.3
-0.4	44.4	71.4	17.8	56.3	28.5	72.6
0.0	47.9	75.0	7.7	57.1	23.4	76.6
0.4	56.5	77.2	0.2	48.9	23.5	79.6
0.8	72.1	81.1	0.0	30.8	27.0	86.5

Note. Percentage of scenario-specific optimized review value at reference reliability and $k^E/d = 0.5$. Both denotes fitted amount and confidence intervals.

For joint design, a Gaussian copula with correlation $-\rho$ couples the four amount states to the twelve increasing-confidence states. Integrating over their probability intervals changes type availabilities while preserving amount–freight and quality marginals within each category. The public groups remain action and amount, or action, amount, and the existing confidence band. Admission and signals are reoptimized within those groups. At the reference correction cost, proportional premiums range from 0.094 to 0.264 without the band and from 0.053 to 0.137 with it; positive-default premiums are 2.13–3.87 and 0.302–0.939. At the 0.75 correction ratio, positive-default premiums are 1.41–2.56 and 0.232–0.416. On the twelve-state confidence grid, the maximum premium across both costs, screens, and responses is 0.0215 at reliability 0.90 and below numerical tolerance at 0.95. The finer grids in Table EC.3 distinguish such numerical zeros from the zero implied by exact public confidence. With separate correction, exact public confidence

gives zero premium in every reference-cost dependence scenario. The largest upper–lower gap is 1.86×10^{-8} per arrival.

Alternative loss mappings replace the bracket in (EC.5) by $0.05 + 0.70V_n$ or $0.05 + 0.35V_n + 0.35F_n$, retaining category multipliers and rescaling construction median loss to 0.60. A matched-lognormal distribution has the reference construction mean and variance. For positive and negative loss–error association, normal scores of error-probability ranks are combined with independent normal noise using coefficients $(\pm 0.8, 0.6)$; sorted reference losses are reassigned in that order, separately within each sample. Evaluation rank correlations are 0.788 and −0.784. These scenarios reorder losses, whereas the amount–error comparison above preserves the loss mapping. Historical Claude 3.5 Sonnet retail runs provide the second agent profile, with category targets $(0.605, 0.722, 0.667, 0.671, 0.969)$ obtained by the same terminal-task classification. Table EC.7 reports optimized gains. Across these inputs and both correction ratios, adding confidence to the fitted amount rule improves evaluation gain except in one negative loss–error scenario at $k^E/d = 0.75$ under proportional scrutiny: its gain is −0.023 while the amount rule selects no review.

Station and screening. The table's lower panel fixes the 120 hidden confidence intervals and public information. Two, three, and four unit-rate reviewers face arrival rate five and a 90% utilization ceiling; congestion is $\delta MW_q(M)$ for the corresponding Erlang-C delay. Staffing expense is excluded. Screening changes multiply noise standard deviation by 0.75 or 1.25 and recompute calibrated weights, interceptions, and false blocks jointly. Halving or doubling $\delta = 0.08$, both correction ratios, and reliability 0.95 preserve positive optimized review value. Admission limits are varied separately at the fixed three-reviewer station. The screening sensitivity changes private human evidence after inspection, leaving routing information fixed. Noise changes both interception probabilities and net inspection values; the gate's calibrated weights minimize assessment error rather than the firm's loss. More precise screening can therefore lower discretionary review value. At reference costs, increasing noise standard deviation by 25% raises optimized value by 0.25 under proportional scrutiny and 0.54 under positive-default scrutiny, even with full disclosure.

Table EC.7 Review Value under Alternative Inputs

	Full inspection		Proportional		Positive default	
k^E/d	0.50	0.75	0.50	0.75	0.50	0.75
Reference	35.2	19.8	14.3	5.2	27.6	13.1
Amount-focused loss	34.1	18.9	13.3	4.8	26.3	12.3
Freight-focused loss	36.3	20.7	14.8	5.4	28.4	13.7
Matched lognormal	34.8	19.2	13.8	5.0	27.4	12.9
Positive loss–error association	36.2	20.4	8.9	2.8	29.1	13.0
Negative loss–error association	34.6	22.0	13.7	5.6	24.0	12.5
Second agent profile	30.9	20.3	7.8	3.2	17.9	9.7
Refined population; station and screening changes						
Reference	34.9	19.4	13.8	5.0	27.2	12.6
Two reviewers	21.1	12.0	10.8	3.8	19.1	8.0
Four reviewers	44.4	25.3	14.9	5.4	31.1	14.9
Noise standard deviation 0.75σ	36.6	21.8	13.8	5.1	27.3	12.8
Noise standard deviation 1.25σ	34.3	18.3	14.1	5.1	27.8	13.0

Note. Gain per 1,000 arrivals over each scenario's best nonreview policy, with optimized admission and full disclosure. The upper panel uses evaluation orders; the lower panel fixes the 9,600-type population.

Appendix D: Refund-Desk Experiment

Order categories, dates, item prices, and freight charges were sampled from Olist; refund requests, return status, and proposed approvals were constructed. A refund was valid if the return had been received, the request was within 30 days of delivery, and the amount did not exceed payment. In rounds 1–3, participants saw routing information and chose audit coverage before seeing individual proposals, amounts, or records. Audits were selected at random. Selected records showed category, dates, return status, reason, item and freight charges, and the proposed refund. Participants could release or block; unaudited proposals executed. Blocks required no correction. Each round began with 15 points, paid 5 per correct outcome, and charged $0.125k^2$ for k audits. The task counted for 2% of the course grade, 1% for completion and 1% for performance; performance credit depended on one randomly selected round and a final check.

Population invalidity was 20%, with 40% routed for review. Enriched, neutral, and depleted routing selected respectively $(80\%, 30\%)$, $(40\%, 40\%)$, and $(20\%, 45\%)$ of invalid and valid proposals, producing queue invalid rates 40%, 20%, and 10%. Each round drew ten proposals independently. Participants were randomized over the six queue orders crossed with whether queue rates were stated. Beliefs were elicited before audit choices but were not separately rewarded. A final, always-last round offered five small orders labeled 40% invalid and five large orders labeled 10% invalid, with a common audit budget and equal rewards. It jointly varied size and error frequency and does not identify a stakes effect.

Of 155 responses, seven test or unidentified entries, 24 incomplete attempts, and six repeat attempts were excluded. The first complete response, including the task and exit survey, was retained for each of 118 MBA participants; 61 received stated rates. Quiz performance, speed, and choices do not determine inclusion.

Queue comparisons use paired Student intervals; detection intervals bootstrap participants, keeping each one's rounds together. Neutral-queue coverage exceeds interpolation between the other queues by 0.021, with interval $[-0.015, 0.058]$. In routed rounds, 361 of 426 audited invalid and 102 of 1,140 audited valid proposals were blocked: rates 0.847 and 0.089 with participant-bootstrap intervals $[0.790, 0.898]$ and $[0.059, 0.123]$.

For beliefs $(\widetilde{q}, \widetilde{\eta}, \widetilde{\phi})$ about invalidity, detection, and false blocking, the continuous optimal audit count is $k^* = [20\{\widetilde{q}\widetilde{\eta} - (1-\widetilde{q})\widetilde{\phi}\}]_{[0,10]}$; the integer optimum is adjacent. Perfect checking gives eight, four, and two audits. Figure EC.4 compares coverage with h_+ and score benchmarks using objective invalidity and pooled checking performance, not subjective optima. Using pooled detection, the queues span interception probabilities approximately 0.085–0.339. Neither a zero-error queue nor the saturation range was observed; correction costs and response tails remain unidentified.

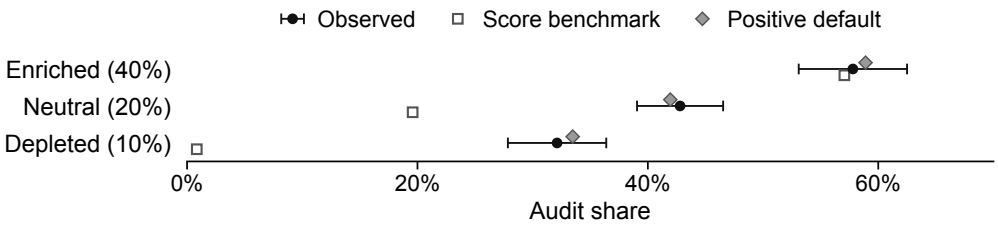


Figure EC.4 Audit Shares by Queue Invalidity

Note. Observed means and pointwise 95% Student intervals use 118 participants; benchmarks use pooled checking performance.