# Optimal Monitoring of AI Use with Reputation Concerns*

Natacha Spitzer† Mengjia Xia†

October 10, 2026

The latest version is available here.

**Abstract**

We study how a manager should monitor workers' use of AI when workers have reputation concerns. A worker of privately known ability chooses between costly effort and a tool. The manager can monitor his choice at a cost. Optimally, either she does not monitor and everyone uses the tool, or she monitors just enough that able workers exert effort and others use the tool. Mandatory disclosure is never optimal, monitoring is non-monotonic in tool quality, and adoption can be too high or too low. Outcome-contingent audits target success when able workers are rare and failure when they are common.

**JEL Codes:** D82, D83, M54, O33.

**Keywords:** Reputation, monitoring, disclosure, artificial intelligence, technology adoption.

*We thank George Mailath, Kevin He, Rakesh Vohra and Xiao Lin for their guidance. We also thank Cuimin Ba, Aislinn Bohren, Byunghoon Kim, Alfonso Maselli, Matthew Murphy, Ritika Verma and Alexander Whitefield, as well as the Penn theory group and participants in the Penn Micro Theory Lunch Seminar, for helpful comments and conversations. All remaining errors are our own.

†Department of Economics, University of Pennsylvania. Emails: nspitzer@sas.upenn.edu, xiax@sas.upenn.edu.

# 1 Introduction

In recent years, advanced artificial intelligence (AI) technologies, especially large language models such as ChatGPT, have become widespread at work. A Pew Research Center survey in September 2025 finds that 21% of U.S. workers use AI on the job (Pew Research Center, 2025), and by 2026 nearly all U.S. knowledge workers report using it (Atlassian Teamwork Lab, 2026).

The growing availability of AI changes workplace incentives. Bonuses and promotions depend on a manager's beliefs about a worker's ability, which she infers from performance. While these tools may raise output, they make it harder to distinguish a worker's effort or innate ability: a manager who sees only the finished work cannot tell how much of it the worker did, and may worry that workers rely on the tool excessively. This creates a new informational problem. A natural response is to monitor the tool's use. Yet workers care about how their use is perceived. Nearly half of desk workers say they are uncomfortable disclosing AI use to their managers (Slack, 2024), and the concern is warranted: in a controlled experiment in which evaluators judged identical work, the colleague described as having used AI was rated far lazier and was 24 percentage points less likely to be recommended for a high-visibility project (Atlassian Teamwork Lab, 2026). Reif et al. (2025) document a similar penalty. Firms have responded in strikingly different ways, from blanket bans to open encouragement, and some now mandate use. This raises several questions. How should a manager monitor the use of AI? How does monitoring shape workers' adoption? And does introducing AI improve outcomes even under optimal monitoring?

We answer these questions in a model in which a worker of privately known ability, high or low, is paid by his reputation: the manager's posterior belief that he is of high ability after observing the outcome of a task and any monitoring signal. The worker completes the task either through his own costly effort or via an AI tool, and the manager chooses the probability with which she observes his technology choice. Own effort and AI are substitutes, and AI's accuracy lies between that of the two types: the high type is more accurate on his own, whereas the low type is less accurate. The accuracy-maximizing technology allocation is then for the low type to use the tool and for the high type to exert effort.

First, we show that the optimal monitoring policy is either not to monitor at all or to monitor the minimum amount that sustains separation, so mandatory disclosure of AI use is never optimal, and the optimal monitoring probability is non-monotonic in the quality of AI. Second, relative to a planner the outcome is inefficient in both directions, with too much effort when monitoring is cheap and too much reliance on AI when it is expensive, so that introducing AI can lower accuracy. Third, when monitoring can be conditioned on the outcome, we characterize the optimal audit: it concentrates on success when able workers are rare and on failure when they are common, the

classic rule of inspecting failures (Dye, 1986) can never be used in full because it invites the able worker to manipulate the outcome, and the manager sometimes prefers partial to full separation. The first two results hold whether effort and AI are substitutes or complements.

We begin by characterizing the equilibria for every monitoring probability. Pooling on AI always exists. Separation, in which the high type exerts effort and the low type uses AI, exists for intermediate monitoring: enough that the high type prefers costly effort, since monitoring reveals him, but not so much that the low type imitates him. Every equilibrium other than pooling on AI requires the manager to monitor at least as much as to sustain separation. This observation greatly simplifies her problem: she either monitors just enough to sustain separation or not at all and accepts pooling on AI. When effort is cheap, separation is self-enforcing and no monitoring is needed. As effort becomes costlier, strictly positive monitoring is required, and beyond a threshold it is too expensive and everyone uses AI. Mandatory disclosure of AI use is never optimal.

Two consequences follow. First, optimal monitoring is non-monotonic in the quality of AI. When the tool is barely better than a low-ability worker, separation is free. As it improves, the high type is increasingly tempted by it and the manager must monitor more, and as it approaches the high type's own accuracy, the gain from separation vanishes and she stops monitoring. Second, relative to a planner who could assign technologies directly, the outcome is distorted in both directions: because the manager cares about accuracy but not about the worker's effort cost, she sustains effort the planner would not, and because she must pay for monitoring, she abandons effort the planner would preserve. In the latter case introducing AI can lower accuracy relative to a world without it.

We then extend the model to allow the worker to combine his own effort with the tool, which is more accurate than either alone. The gain is larger for the high type. The main insights persist: reputation alone sustains the combined technology over a range of costs, and beyond it the baseline trade-off returns.

Finally, we allow monitoring to depend on the outcome, which the manager must then account for. This creates a new margin of manipulation, since a worker can misreport a correct signal to trigger or avoid an audit. The optimal audit targets the outcome that provides separating incentives most cheaply. It targets success when high-ability workers are rare, since a correct outcome then carries little reputational value on its own and an audit rewards effort, and failure when they are common, since a failure is then rare and an audit punishes reliance on the tool. The other outcome is audited only if effort is costly, and only as much as needed to deter misreporting. The one-sided audit is cheaper than symmetric monitoring, so the manager sustains separation on a larger set of parameters. The rule familiar from costly-state-verification models, inspect failures and trust successes, can never be used in full when incentives are reputational, because this generates

incentives to misreport. When failures are audited and effort is costly, the manager sometimes does better by tolerating a little effort from the low type, as failures can then be audited more intensely.

Two forces drive all of these results: the worker's technology choice is a signal of his ability, and monitoring is an instrument that helps one type and hurts the other. The same forces operate in both extensions, and the prescription is identical throughout: monitor minimally or not at all.

Section 2 introduces the model and the efficient benchmark, Section 3 characterizes equilibria for a given monitoring policy, Section 4 studies optimal monitoring, Section 5 extends the model to complements and Section 6 to outcome-contingent monitoring, and Section 7 concludes. Proofs are in Appendix A.

**Related literature.** The worker's payoff is his reputation, as in the career-concerns literature following Holmström (1999) and Dewatripont et al. (1999). A related strand studies how reputational concerns distort decisions that change the distribution of outcomes: conservatism and excessive risk-taking in Holmström and Ricart i Costa (1986), Zwiebel (1995) and Chevalier and Ellison (1999), reputational herding in Scharfstein and Stein (1990) and Prendergast and Stole (1996), and reputational cheap talk in Ottaviani and Sørensen (2006). Delegating a task to the tool is a decision of the same kind: output produced with the tool is equally accurate whatever the worker's ability, so what is informative is the choice of technology itself.

The environment closest to ours is that of Weitzner (2024), who shows in a cheap-talk model that a worker who cares about his reputation will sometimes override a more accurate algorithm, because following it signals low skill. We share the reputational payoff and the ordering of accuracies, but our worker's action is costly and unobserved, the manager chooses how often to observe it, and our object is her monitoring policy. The precise relation is discussed after Theorem 1. Baranchuk and Rivera (2026) also give the worker private information about AI, in their case about whether it can perform his job at all: he conceals successful automation to avoid dismissal, and the firm answers with bonuses for truthful reports and the threat of termination. There concealment protects the job. Here it protects how the worker is judged, and the firm's only instrument is the frequency with which it monitors.

Our work ties in with the general literature on principal-agent problems with costly monitoring. In Dye (1986), the principal commits in advance to a rule that lets her pay to investigate the agent's action after seeing output, as the manager does here, and the optimal rule is deterministic and lower-tailed, investigating exactly when output falls below a threshold. Baiman and Demski (1980) ask the same question in an accounting setting, and Young (1986) shows that investigating both tails can be optimal there. In costly state verification the agent instead reports a state the principal

can verify at a cost (Townsend, 1979; Border and Sobel, 1987; Mookherjee and Png, 1989). In the voluntary-disclosure literature following Grossman (1981) and Milgrom (1981), information is verifiable and unravels, so the receiver reads silence as the worst news. Here the worker cannot prove which technology he used, so the manager learns his choice only by paying to monitor. Inspection schemes have also been studied in computer science, where Ezra et al. (2024) and Fallah and Jordan (2024) characterize the complexity of finding optimal incentive-compatible inspections when the agent is paid transfers. Our worker's incentives are purely reputational, and we characterize the optimal scheme rather than its computational hardness.

The model's two substantive assumptions have direct empirical counterparts. The first is that using AI carries a reputational cost: workers hesitate to use it because of how they expect to be perceived (Almog, 2025), and managers told that GPT was used rate the same work lower and update against the analyst, a difference that disappears without disclosure (Yang et al., 2025; Reif et al., 2025; Atlassian Teamwork Lab, 2026). The second is that AI may improve the productivity of low-ability workers while hindering that of high-ability workers (Brynjolfsson et al., 2025; Dell'Acqua et al., 2023; Noy and Zhang, 2023). Complementarity of the tool and effort depending on ability is documented by Otis et al. (2026) and Agarwal et al. (2023).

In the broader literature on AI and the labor market, the technology's role is assigned: firms or the market determine which tasks it performs and at which layer of the organization it is deployed, and the resulting reorganization is the object of study (Acemoglu and Restrepo, 2018; Agrawal et al., 2019; Athey et al., 2020; Ide and Talamás, 2025, 2024; Xu et al., 2025). Here no one assigns anything: the worker chooses unobserved, and what drives his choice is what he expects the firm to infer from it rather than what the tool can do. Oversight appears in Xu et al. (2025) as a check on the tool's errors, whereas here monitoring is a way of learning about the worker.

# 2 Baseline Model and Efficient Benchmark

## 2.1 Model

A manager assigns a task to a worker whose ability is denoted by $\theta \in \{H, L\}$, where $H$ represents high ability and $L$ low ability. We denote by $\mu$ the prior probability that a worker is of high ability. Ability is perfectly observed by the worker but not by the manager.

To complete the task, the worker can either rely on his own (effort-based) technology, denoted by $O$, or use an artificial intelligence technology, denoted by $AI$. The worker's choice is $a \in \{O, AI\}$. We assume that worker effort and AI are substitutes.[1] The cost of using a technology does not

[1] One justification is that the worker faces a fixed time budget and can devote it to only one method. In Section 5,

depend on worker type. Using his own technology costs the worker $c \in (0,1)$, whereas using AI is free.[2]

After choosing a technology, the probability that the worker completes the task correctly depends on both his ability and his technology choice. Let $q_\theta$ denote the probability that a type-$\theta$ worker completes the task correctly with his own technology, and let $q_{AI}$ denote the accuracy of AI.[3] We assume:

$$0.5 < q_L < q_{AI} < q_H < 1,$$

so that high-ability workers outperform AI, while AI outperforms low-ability workers.[4]

A worker's strategy is a mapping

$$\sigma : \{H, L\} \to \Delta(\{O, AI\}).$$

For each type, $\sigma(\theta)$ specifies a probability measure over the two technologies. For any strategy $\sigma$, denote by

$$\alpha_\theta(\sigma) \equiv \Pr_\sigma(a = AI \mid \theta)$$

the probability that a type-$\theta$ worker chooses AI under $\sigma$. For a profile $\alpha = (\alpha_H, \alpha_L)$, expected accuracy is

$$W(\alpha) \coloneqq \mu\Big(\alpha_H q_{AI} + (1-\alpha_H)q_H\Big) + (1-\mu)\Big(\alpha_L q_{AI} + (1-\alpha_L)q_L\Big),$$

the ex-ante probability that the task is completed correctly.

The manager commits ex ante to a monitoring technology.[5] With probability $\beta \in [0,1]$, she perfectly observes whether AI was used to complete the task, along with whether the task was completed correctly.[6] Otherwise, when the worker is not monitored, she only observes whether he was correct.

---

we relax this assumption and allow for complementarities between worker effort and AI.

[2] Alternatively, one may interpret $c$ as the cost differential between worker effort and AI.

[3] A microfoundation for this assumption is to view the task as depending on an underlying unknown state $\omega \in \{0,1\}$, and the worker receives a noisy binary signal whose precision depends on ability and technology. For tractability, we work directly with reduced-form accuracy parameters.

[4] This is the case in which the manager faces a screening problem: if $q_{AI} \geq q_H$, every worker should use AI and there is no tension, whereas if $q_{AI} \leq q_L$, no worker should, and monitoring has nothing to protect. It is also the pattern documented in the field, where AI assistance raises the productivity of less skilled workers by the most and of the most skilled by the least (Brynjolfsson et al., 2025; Dell'Acqua et al., 2023). The case in which the skilled worker gains the most by combining effort with the tool is treated in Section 5.

[5] Commitment is essential. Since the manager cares only about accuracy, monitoring has no value to her once the task is complete, so absent commitment she would never monitor and only pooling on AI would survive.

[6] Since the choice is binary, this implies that when she observes that AI was not used, she perfectly infers that the worker exerted effort.

The worker is paid a fixed wage, so the monitoring probability is the manager's only instrument. This is what distinguishes the exercise from standard contracting, where the principal designs transfers: here the worker's reward for a good reputation comes from outside the current arrangement, through promotion, retention, or the labor market, and the manager chooses only how much she learns about how the work was done.

Fixing the profile of probabilities that AI is chosen by each type, $\alpha = (\alpha_H, \alpha_L)$, the manager's posterior beliefs are denoted by $b_r(\alpha)$ when the task is completed correctly and by $b_w(\alpha)$ when the task is completed incorrectly in cases where the worker is not monitored, and we write $\Delta b(\alpha) \coloneqq b_r(\alpha) - b_w(\alpha)$ for the reputational return to a correct prediction. Explicitly, with $x_\theta(\alpha) \coloneqq \alpha_\theta q_{AI} + (1 - \alpha_\theta) q_\theta$ the accuracy of type $\theta$ under $\alpha$,

$$b_r(\alpha) = \frac{\mu\, x_H(\alpha)}{\mu\, x_H(\alpha) + (1-\mu)\, x_L(\alpha)}, \qquad b_w(\alpha) = \frac{\mu\,(1 - x_H(\alpha))}{\mu\,(1 - x_H(\alpha)) + (1-\mu)\,(1 - x_L(\alpha))}.$$

When monitoring occurs, the manager updates her beliefs based on the observed technology choice: $b'_{a,r}(\alpha)$ and $b'_{a,w}(\alpha)$ denote the posteriors when technology $a \in \{O, AI\}$ is observed, conditional on the task being completed correctly or incorrectly, respectively. If the manager monitors and observes a technology that no type uses under $\alpha$, we adopt the pessimistic off-path belief that the worker is of low ability.[7] [8]

The worker seeks to maximize the manager's interim expected posterior belief that he is of high ability, net of the cost of using his own technology.[9] For each type and action pair $(\theta, a)$, define

$$q(\theta, a) \coloneqq \begin{cases} q_{AI} & \text{if } a = AI, \\ q_\theta & \text{if } a = O. \end{cases}$$

[7] A less pessimistic off-path belief would not change the results qualitatively: it raises the payoff to the off-path technology, so the manager needs a higher monitoring probability to give the same incentives.

[8] Standard belief-based refinements have no bite here. In the pooling equilibria the off-path technology yields the same payoff to both types for every off-path belief. Under pooling on AI, output is uninformative, so the unmonitored posterior is the prior whatever the deviator's accuracy, and the monitored part of the deviation payoff is $\beta\, b'_O - c$ for either type. Under pooling on effort the deviation is to AI, whose accuracy is $q_{AI}$ for both. The set of beliefs that would make the deviation profitable therefore coincides across types and neither the Intuitive Criterion nor D1 restricts the belief. Under separation both technologies are used on path, and in the complements environment of Section 5 the off-path technology is dominated for every belief. Multiplicity of equilibria is instead resolved, as is standard in mechanism design, by selecting the equilibrium the manager prefers (Section 4).

[9] This is the standard reduced form of career concerns: if the worker is paid his expected ability in a competitive labor market after the task, as in Holmström (1999), his expected future wage equals the manager's posterior, and the analysis below is unchanged. The planner's benchmark of Section 2 is then the usual one, since wages are transfers.

Given a monitoring probability $\beta$ and fixing beliefs, let

$$\rho_a^\theta(\beta) \;\coloneqq\; \beta\Big(q(\theta,a)\, b'_{a,r} + \big(1-q(\theta,a)\big)\, b'_{a,w}\Big) + (1-\beta)\Big(q(\theta,a)\, b_r + \big(1-q(\theta,a)\big)\, b_w\Big)$$

denote the *expected reputation* of a type-$\theta$ worker who chooses action $a$. With probability $\beta$ he is monitored, the manager sees his technology and whether he was right, and forms $b'_{a,r}$ or $b'_{a,w}$. With probability $1-\beta$ she sees only whether he was right and forms $b_r$ or $b_w$. He is right with probability $q(\theta,a)$, which depends on his ability only if he exerts effort. His payoff is his expected reputation net of the cost of effort,

$$U^\theta(a,\beta) \;=\; \rho_a^\theta(\beta) - \mathbf{1}\{a=O\}\, c.$$

Two objects summarize the reputational incentives at any profile $\alpha$. The first is the return to a correct prediction when the worker is not monitored, $\Delta b(\alpha)$. The second is the *effort premium* of type $\theta$,

$$\Pi_\theta(\alpha) \coloneqq q_\theta\, b'_{O,r}(\alpha) + (1-q_\theta)\, b'_{O,w}(\alpha) - b'_{AI}(\alpha),$$

the expected gain in reputation, when monitored, from being seen to exert effort rather than to use AI. Here $b'_{AI}(\alpha) \coloneqq b'_{AI,r}(\alpha) = b'_{AI,w}(\alpha)$, since correctness carries no information about ability once AI is observed. Every incentive constraint in the paper is an instance of the following comparison.

**Lemma 1** (Effort incentive)**.** *Fix a profile $\alpha$ and a monitoring probability $\beta$. Type $\theta$ weakly prefers his own technology to AI if and only if*

$$c \;\le\; (1-\beta)\,(q_\theta - q_{AI})\,\Delta b(\alpha) \;+\; \beta\,\Pi_\theta(\alpha). \tag{1}$$

The lemma says that effort is worth its cost when the extra accuracy it buys is rewarded in reputation (the first term, which scales with how informative correctness is about ability) or when being seen to exert effort is itself rewarded (the second term, which monitoring makes available). In the profiles that matter below, $\Pi_\theta$ takes only a few values: it equals 1 when effort identifies the high type and AI the low type, $-\mu$ when everyone uses AI and effort is attributed to the low type, and $q_\theta b_r + (1-q_\theta)b_w$ when everyone exerts effort so that being observed reveals nothing beyond correctness.

## 2.2 Efficient Benchmark

Suppose first that there exists a frictionless social planner who wishes to maximize expected accuracy net of effort costs. She chooses a profile of AI-usage probabilities $\alpha = (\alpha_H, \alpha_L)$ and solves

$$\max_{\alpha \in [0,1]^2} W(\alpha) - c\Big[\mu(1-\alpha_H) + (1-\mu)(1-\alpha_L)\Big],$$

where the second term is the expected cost of effort.

**Proposition 1.** *The efficient allocation $\alpha^* = (\alpha_H^*, \alpha_L^*)$ is given by*

$$\alpha_H^* = \begin{cases} 0 & \text{if } c \le q_H - q_{AI}, \\ 1 & \text{if } c > q_H - q_{AI}, \end{cases} \qquad \alpha_L^* = 1.$$

In the efficient allocation, the low-ability worker always uses AI. The efficient technology choice of the high-ability worker depends on whether the gain in precision from own effort relative to AI, $q_H - q_{AI}$, is large enough to justify the cost $c$: for $c \le q_H - q_{AI}$, the high type uses own effort, and otherwise he efficiently uses AI.

In the following section, we solve for the set of worker equilibria as a function of the manager's monitoring technology. Then, in Section 4, we turn to the optimal monitoring scheme of the manager.

# 3 Equilibrium for a Given Monitoring Policy

In this section, we focus on the game played by the worker, taking the monitoring policy $\beta \in [0, 1]$ as exogenously given. We characterize the set of Perfect Bayesian Equilibria under arbitrary monitoring policies. This characterization is used to determine the optimal policy in Section 4.

Applying Lemma 1 profile by profile yields the set of equilibria as a function of the monitoring probability. Two facts organize it. First, the two types cannot both be indifferent between the technologies, since they differ in accuracy, so at most one type mixes in any equilibrium. Second, the profile in which the high type uses AI and the low type exerts effort is never an equilibrium, since monitoring would then identify types in reverse. The remaining profiles can therefore be ordered by the amount of monitoring each requires.

**Proposition 2** (Equilibrium as a function of monitoring)**.** *Fix $c \in (0, 1)$. There exist thresholds $0 \le \underline{\beta}^{sep} < \bar{\beta}^{sep} < \underline{\beta}^{pool}$ and $\bar{\beta}^{HM} \in (\underline{\beta}^{sep}, 1]$ such that, for every $\beta \in [0, 1]$, the perfect Bayesian equilibria are exactly the following.*

*(i) Pooling on AI: for every $\beta$.*

*(ii) Separation, high type exerts effort and low type uses AI: iff $\beta \in [\underline{\beta}^{sep}, \bar{\beta}^{sep}]$.*

*(iii) Pooling on effort: iff $\beta \geq \underline{\beta}^{pool}$.*

*(iv) High type mixes, low type uses AI: iff $\beta \in [\underline{\beta}^{sep}, \bar{\beta}^{HM}]$.*

*(v) Low type mixes, high type exerts effort: iff $\beta \in [\bar{\beta}^{sep}, \min\{1, \underline{\beta}^{pool}\}]$.*

The thresholds have closed forms, given in Appendix A, from which the following properties follow.

**Corollary 1** (Properties of the thresholds)**.** *The thresholds of Proposition 2 have the following properties.*

*(i) $\underline{\beta}^{sep} = 0$ iff $c \leq c^* \coloneqq (q_H - q_{AI})\,\Delta b(0,1)$, the reputational return to effort for the high type under separation. Where positive, $\underline{\beta}^{sep}$ is strictly increasing in $c$ and in $q_{AI}$.*

*(ii) There is a $\bar{c} \in (0,1)$ such that $\underline{\beta}^{pool} \leq 1$, so that pooling on effort exists for some $\beta$, iff $c \leq \bar{c}$.*

*(iii) In the mixed-strategy equilibrium (iv), the high type's probability of using AI, $\hat{\alpha}_H(\beta)$, is continuous and strictly increasing on $[\underline{\beta}^{sep}, \bar{\beta}^{HM}]$. In that of (v), the low type's, $\hat{\alpha}_L(\beta)$, is continuous and strictly decreasing on $[\bar{\beta}^{sep}, \min\{1, \underline{\beta}^{pool}\}]$.*

The proposition describes the equilibrium set as monitoring rises from zero to one. Pooling on AI is always available, since output is then uninformative and effort would be attributed to the low type. Separation requires intermediate monitoring: more monitoring makes effort more attractive to the high type, who is identified when monitored, but also to the low type, who is identified too. The high type's constraint sets the lower threshold and the low type's the upper. Pooling on effort requires more monitoring still, since only the threat of being seen using AI holds the low type to costly effort, and requires effort to be cheap enough, $c \leq \bar{c}$. In between the thresholds, mixed-strategy equilibria exist, in which the mixing probabilities move continuously from one pure-strategy equilibrium to another. Every equilibrium other than pooling on AI therefore requires at least $\underline{\beta}^{sep}$, so the manager's problem in the next section is a choice between separation and no monitoring. Figure 1 collects these regions for a cost $c < \bar{c}$, at which all of them are nonempty.

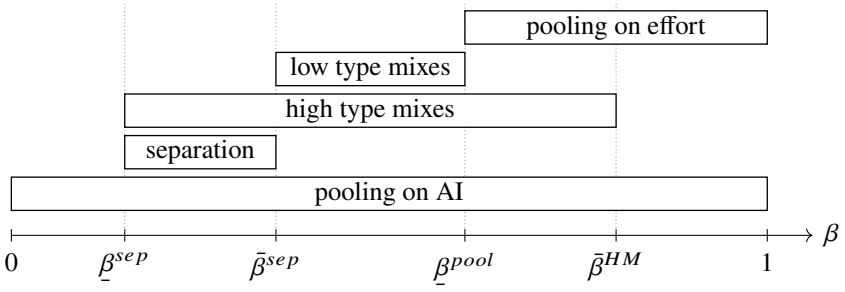



Figure 1: The equilibrium set as a function of the monitoring probability $\beta$ (Proposition 2), drawn for $c < \bar{c}$ so that pooling on effort exists. The position of $\bar{\beta}^{HM} = \min\{1, c/(1-\mu)\}$ relative to $\bar{\beta}^{sep}$ and $\underline{\beta}^{pool}$ is not determined in general: it always exceeds $\underline{\beta}^{sep}$ and increases with the prior $\mu$, and the figure is drawn for $\mu$ large enough that it exceeds $\underline{\beta}^{pool}$.

# 4 Optimal Costly Monitoring

## 4.1 Manager's Problem and Optimal Monitoring Policy

Now suppose that the monitoring probability is chosen by the manager and comes at a cost. When the manager monitors, she must pay a fee $k > 0$. The manager chooses $\beta$ to maximize the ex-ante probability that the task is completed correctly minus the expected cost of monitoring, taking into account the workers' equilibrium behavior.

Let $\Sigma^{PBE}(\beta)$ denote the set of Perfect Bayesian Equilibrium strategy profiles of the worker's game when the monitoring probability is fixed at $\beta$. If multiple equilibria exist for a given $\beta$, we assume that the manager-optimal equilibrium is selected.

Let

$$A(\beta) \coloneqq \max_{\sigma \in \Sigma^{PBE}(\beta)} W\big(\alpha(\sigma)\big)$$

denote the accuracy of the manager-optimal equilibrium at monitoring probability $\beta$. The manager's problem is

$$\max_{\beta \in [0,1]} \; A(\beta) - k\beta.$$

In the absence of monitoring costs the manager implements separation, since it maximizes expected accuracy. Even at $k = 0$, however, she does not monitor with certainty: full monitoring, $\beta = 1$, is never optimal, because it gives the low type reason to exert effort as well. The following result shows that when monitoring is cheap enough relative to the accuracy gain from separation, the manager monitors just enough to sustain it, and that the amount required is zero for every $c \leq c^*$. Otherwise she does not monitor and accepts pooling on AI.

**Theorem 1** (Optimal Costly Monitoring)**.** *The manager optimally sets*

$$\beta^{opt} = \begin{cases} \underline{\beta}^{sep} & \text{if } k\,\underline{\beta}^{sep} \leq \mu(q_H - q_{AI}), \\ 0 & \text{otherwise.} \end{cases}$$

*In the manager-optimal equilibrium, in the first case, the high type uses his own technology and the low type uses AI. In the second case, both types use AI.*[10]

Note that the manager can obtain her preferred separating equilibrium without any monitoring, i.e. with $\beta^{opt} = 0$, if and only if $c \leq c^*$, the threshold of Corollary 1.

When the cost for the worker of choosing his own technology is sufficiently low, the increased reputation effect from getting the answer correct is enough to incentivize the high type to choose his own technology when the low type chooses AI, and thus there exists a separating equilibrium when $\beta = 0$. Since this is the preferred equilibrium of the manager and monitoring is costly, she optimally sets the monitoring probability to zero.

However, if this cost for the worker is sufficiently high, the manager must monitor with strictly positive probability to induce separation. Moreover, to induce either a mixed-strategy equilibrium or the non-trivial pooling equilibrium in which both types use their own technology, the manager must set the monitoring probability to be strictly greater than $\underline{\beta}^{sep}$. If $\beta < \underline{\beta}^{sep}$, then there is a unique equilibrium in which both types use AI. The manager therefore either monitors with just enough probability to induce separation, or does not monitor at all and induces the pooling equilibrium.

**Monitoring as a disclosure policy.** The monitoring probability has a natural institutional counterpart. A firm that requires workers to log and report the tools they used is choosing a $\beta$ close to one. A firm that reads the finished work and asks nothing about how it was produced is choosing $\beta = 0$. A firm that spot-checks a fraction of assignments, or that looks into an assignment when something about the outcome prompts a review, is choosing an interior $\beta$, or the outcome-contingent version of Section 6. Read this way, Theorem 1 is a statement about how much a firm should ask. Asking always is never optimal, because it takes away the low type's reason to use the tool that serves him better. Asking nothing is optimal both when effort is cheap enough for reputation alone to keep the high type working and when checking costs more than the accuracy it buys. In between, the firm should ask as little as separation requires.

[10] When the manager is indifferent, we select the separating equilibrium.

**Relation to reputational algorithm aversion.** In Weitzner (2024) a worker chooses between his own private signal and an algorithm's signal that the manager also observes, and the choice is costless. Because the manager observes the algorithm's signal, she can infer when it was followed. Following it therefore conveys no information about the worker, and the low type overrides it to raise his perceived ability. In our model the technology choice is unobserved unless the manager monitors, so an unmonitored low type uses AI: only the outcome is informative, and AI makes a correct outcome more likely. The reputational cost of using the tool is thus created by monitoring itself, which is why full disclosure is never optimal. Its failure is not a matter of expense: even at $k = 0$, monitoring with certainty leads the low type to bear the cost of effort rather than be seen using AI, so separation is no longer an equilibrium and every equilibrium that remains is less accurate. This is clearest at $c = 0$, where separation is obtained at $\beta = 0$. A strictly positive $k$ only adds to the loss, since full disclosure is also the most expensive policy she could choose.

## 4.2 Shape of Optimal Monitoring in AI Quality

We now study how the manager's optimal monitoring intensity responds to improvements in AI accuracy. Fix the effort cost $c$ and a strictly positive monitoring cost $k > 0$, and consider $\beta^{opt}$ as a function of $q_{AI} \in (q_L, q_H)$. Two thresholds organize the result. The first is the cost below which separation is free at a given AI accuracy, $c^*(q_{AI}) = (q_H - q_{AI})\Delta b(0,1)$ from Corollary 1, which is strictly decreasing in $q_{AI}$ and vanishes at $q_H$. The second matters when $c > c^*(q_L)$, so that monitoring is needed at every AI quality. The manager monitors when the accuracy gain from separation covers its cost, $k\,\underline{\beta}^{sep}(q_{AI}) \leq \mu(q_H - q_{AI})$. As AI improves the gain on the right falls and the monitoring on the left rises, so the condition is easiest to meet at the lowest AI quality. Evaluating it there gives $\bar{k}(c) \coloneqq \mu(q_H - q_L)/\underline{\beta}^{sep}(q_L)$, the monitoring cost above which the manager never monitors, whatever the quality of AI.

**Proposition 3** (Shape of Optimal Monitoring in AI Accuracy)**.** *Either the manager never monitors, or there exist $q_L \leq q_- < q_+ < q_H$ such that she monitors if and only if $q_{AI} \in (q_-, q_+]$, and $\beta^{opt}(q_{AI}) = \underline{\beta}^{sep}(q_{AI})$ is strictly increasing on that interval. Which case obtains depends on the cost of effort $c$ and the cost of monitoring $k$:*

*(i) If $c \leq c^*(q_L)$, the manager monitors on a nonempty interval for every $k > 0$, with $q_- > q_L$.*

*(ii) If $c > c^*(q_L)$, she monitors on a nonempty interval if and only if $k \leq \bar{k}(c)$, with $q_- = q_L$.*

Proposition 3 shows that, as the quality of AI increases between $q_L$ and $q_H$, the manager either never monitors, or monitors only on a single strict subinterval, and it says which. When effort is

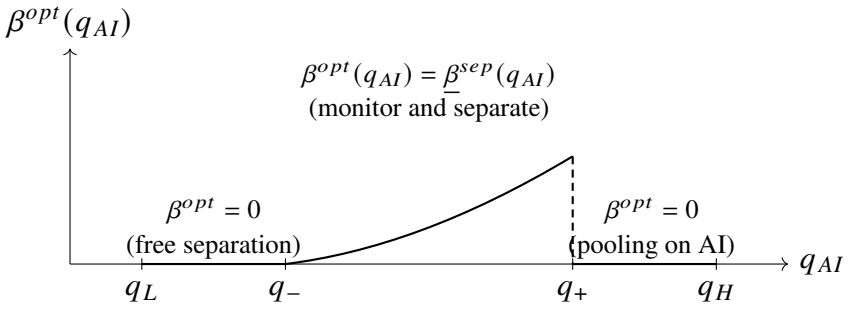



Figure 2: Optimal monitoring $\beta^{opt}(q_{AI})$ as AI accuracy increases, in case 1 of Proposition 3 ($c \leq c^*(q_L)$): free separation below $q_-$, monitoring on $(q_-, q_+)$, pooling on AI above $q_+$.

cheap enough that separation is free for AI barely better than the low type, there is always such an interval, whatever the cost of monitoring. When effort is expensive, there is one if and only if monitoring is cheap enough. Monitoring is beneficial for the manager when three conditions hold simultaneously: separation is more accurate than pooling on AI, reputational incentives alone are not strong enough to sustain separation, and the accuracy gain from separation is large enough to justify the monitoring cost $k$.

Figure 2 illustrates the first case, in which $q_L < q_- < q_+ < q_H$. When the AI precision is close to that of the low type and the cost of effort is sufficiently low, the reputational gain for the high type from separating is sufficiently high such that no monitoring is required to achieve separation and the manager optimally monitors with probability zero. As the AI precision increases, AI becomes more and more attractive for the high type and thus the manager must increase the probability of monitoring to achieve separation. Eventually, as separation becomes more and more costly for the manager and the gain in accuracy diminishes with AI quality, the manager optimally stops monitoring and accepts that all workers use AI.

## 4.3 Comparison to the Efficient Benchmark

The efficient benchmark in Proposition 1 abstracts from both reputation and monitoring: a frictionless planner directly assigns technologies to types to maximize expected accuracy net of effort costs. In the baseline model, this yields a simple target allocation. The low-ability worker should always use AI, while the high-ability worker should use own effort if and only if the gain in precision relative to AI exceeds the private cost of effort, $q_H - q_{AI} \geq c$. For low effort costs, the efficient allocation is therefore separating: the high type exerts own effort and the low type uses AI. For high effort costs, the efficient allocation instead assigns AI to both types.

By contrast, the manager in our model cannot directly choose technologies. She chooses a

monitoring intensity $\beta$, and technologies are chosen endogenously through the worker's reputational incentives. For low values of $c$ such that separation can be sustained without monitoring, $c \leq c^*$, the manager-optimal equilibrium coincides with the efficient allocation. Once separation requires strictly positive monitoring, however, two kinds of departures from efficiency arise. For intermediate effort costs with $q_H - q_{AI} < c$, the planner would already prefer the high type to use AI, but if the monitoring cost $k$ is sufficiently small the manager still prefers to sustain the separating equilibrium, since she does not internalize the worker's private cost of effort. In this region the high type uses *too little* AI relative to the benchmark. For higher monitoring costs, separation is superior in accuracy but too expensive to sustain: even when the planner would like the high type to use own effort and the low type to use AI (i.e., $c < q_H - q_{AI}$), the manager optimally sets $\beta^{opt} = 0$ and accepts pooling on AI instead. In this region the high type uses *too much* AI relative to the efficient benchmark.

The analysis so far treats effort and AI as pure substitutes: using AI crowds out own effort. In many applications, however, AI and human effort are complementary, since workers can combine their own analysis with AI assistance to achieve higher productivity than either technology alone. In the next section, we extend the model to allow for such complementarities.

# 5 Complementarities Between AI and Worker Effort

## 5.1 Model

Now suppose that the worker can choose to use either his own technology, AI, or both technologies simultaneously. The action set becomes

$$a \in \{O, AI, B\},$$

where $B$ denotes using both own effort and AI.

As before, if he uses only one technology, the low-ability worker is less precise than AI and the high-ability worker is more precise than AI, while using both is most precise for each type: $0.5 < q_L < q_{AI} < q_{LB} < 1$ and $0.5 < q_{AI} < q_H < q_{HB} < 1$. In this sense, AI and worker effort are complements.

**Assumption 1** (Ability-Dependent Complementarity)**.** *Compared to the worker's own technology alone, the high-ability worker's gain in precision from using both technologies is higher than that of the low-ability worker:*

$$q_{LB} - q_L < q_{HB} - q_H.$$

Exerting own effort entails a private cost $c \in (0,1)$, regardless of whether the worker uses effort alone or in combination with AI. Using only AI is free.

Monitoring is as in the baseline model, except that the manager cannot distinguish $AI$ from $B$. With probability $\beta$, she observes whether AI was used (i.e., whether $a \in \{AI, B\}$), but not whether the worker additionally exerted own effort.

We extend the definition of signal precision to

$$q(\theta, a) \coloneqq \begin{cases} q_{AI} & \text{if } a = AI, \\ q_\theta & \text{if } a = O, \\ q_{\theta B} & \text{if } a = B, \end{cases}$$

so that $q_{LB}$ and $q_{HB}$ denote the precisions when the low- and high-ability workers use both technologies.

Given beliefs as in Section 2, with the same pessimistic off-path belief upon observing a technology no type uses, the expected reputation $\rho_a^\theta(\beta)$ of a type-$\theta$ worker who chooses $a \in \{O, AI, B\}$ is defined exactly as before, with $q(\theta, a)$ extended as above, and his payoff is

$$U^\theta(a, \beta) \;=\; \rho_a^\theta(\beta) - \mathbf{1}\{a \in \{O, B\}\}\, c.$$

The only change from the baseline is that effort may be combined with AI, at the same cost and with higher accuracy. Since the manager cannot tell $B$ from $AI$, the monitored beliefs after either action coincide, $b'_{B,\cdot} = b'_{AI,\cdot}$, so $\rho_B^\theta$ and $\rho_{AI}^\theta$ differ only through the probability of being correct.

## 5.2 Efficient Benchmark

Conditional on exerting effort, both types can costlessly increase their precision by also using AI. Hence it is never efficient to use own effort without AI, and the relevant choice for each type is between AI only and the combined technology $B$. As in the baseline, the planner's problem separates across types, and for each it compares the accuracy that effort adds on top of AI to the cost of effort.

For a profile $\sigma$ over $\{O, AI, B\}$, let $\alpha_\theta^a(\sigma)$ denote the probability that type $\theta$ chooses action $a$, and let

$$W_C(\sigma) \coloneqq \mu\big(\alpha_H^O q_H + \alpha_H^{AI} q_{AI} + \alpha_H^B q_{HB}\big) + (1-\mu)\big(\alpha_L^O q_L + \alpha_L^{AI} q_{AI} + \alpha_L^B q_{LB}\big)$$

denote expected accuracy. The planner therefore chooses, for each type $\theta \in \{H, L\}$, a probability

$\alpha_\theta$ of using AI only (and $1-\alpha_\theta$ of using $B$), solving

$$\max_{\alpha_H,\alpha_L\in[0,1]} W_C(\sigma) - c\Big[\mu(1-\alpha_H) + (1-\mu)(1-\alpha_L)\Big]$$

over profiles with $\alpha_\theta^{AI} = \alpha_\theta$ and $\alpha_\theta^{B} = 1-\alpha_\theta$.

**Proposition 4.** *The allocation $\alpha^* = (\alpha_H^*, \alpha_L^*)$ is efficient, where for each $\theta \in \{H, L\}$,*

$$\alpha_\theta^* = \begin{cases} 1 & \text{if } q_{\theta B} - q_{AI} \le c, \\ 0 & \text{if } q_{\theta B} - q_{AI} > c. \end{cases}$$

In the efficient allocation, AI is always used. For each type, own effort is used (via the combined technology $B$) if and only if the gain in precision $q_{\theta B} - q_{AI}$ is large enough to offset the private cost of effort $c$.

## 5.3 Manager's Problem and Optimal Monitoring Policy

As in the baseline model, the manager cares only about prediction accuracy and incurs a linear monitoring cost $k\beta$. For a given monitoring probability $\beta \in [0,1]$, let $\Sigma_C^{PBE}(\beta)$ denote the set of Perfect Bayesian equilibria of the worker game in the environment with actions $\{O, AI, B\}$.

Writing $A_C(\beta) \coloneqq \max_{\sigma\in\Sigma_C^{PBE}(\beta)} W_C(\sigma)$ for the accuracy of the manager-optimal equilibrium, the manager's problem in the complements environment is

$$\max_{\beta\in[0,1]} A_C(\beta) - k\beta.$$

The manager-optimal input choice is now for both types to pool on using both technologies, since this yields the highest precision for both. The following proposition implies that whether this equilibrium exists is completely independent of the monitoring probability. Since the manager cannot distinguish $AI$ from $B$, in any profile in which both types use AI (alone or combined) monitoring is uninformative, and the only reputational return is to a correct prediction. With a slight abuse of notation, for accuracies $x_H, x_L \in (0,1)$ let $\Delta b(x_H, x_L) \coloneqq b_r - b_w$ denote this return when the high type is correct with probability $x_H$ and the low type with probability $x_L$, computed as in Section 2. Note that $\Delta b(x_H, x_L)$ is strictly decreasing in $x_L$ and strictly increasing in $x_H$. The counterpart of Lemma 1 is then immediate.

**Lemma 2** (Effort incentive with complementarities)**.** *Fix a profile in which both types use AI, alone or combined, inducing accuracies $(x_H, x_L)$, and any $\beta \in [0,1]$. Own technology alone is*

*dominated by the combined technology for both types, and type $\theta$ weakly prefers the combined technology to AI alone if and only if*

$$c \;\leq\; (q_{\theta B} - q_{AI})\,\Delta b(x_H, x_L).$$

**Proposition 5** (Equilibria with the combined technology)**.** *Suppose Assumption 1 holds. There exist thresholds $0 < \tilde{c} < c^{\dagger} < c^{\ddagger} < 1$, independent of $\beta$ and given in closed form in Appendix A, such that for every $\beta \in [0,1]$ the perfect Bayesian equilibria in which some type uses the combined technology are exactly the following.*

*(i) Pooling on the combined technology: iff $c \leq \tilde{c}$.*

*(ii) High type uses the combined technology, low type mixes between it and AI: iff $c \in [\tilde{c}, c^{\dagger}]$.*

*(iii) Separation, high type uses the combined technology and low type uses AI: iff $c \in [c^{\dagger}, c^{\ddagger}]$.*

*(iv) High type mixes between the combined technology and AI, low type uses AI: iff $c < c^{\ddagger}$.*

*In (ii) the low type's probability of using the combined technology is continuous and strictly decreasing in c, and in (iv) the high type's is continuous and strictly increasing in c.*

Intuitively, when $c$ is very small, both types are willing to pay the effort cost in order to enjoy the reputational benefits from the more precise signal $q_{\theta B}$. This yields a pooling equilibrium in which both types always choose both technologies. As $c$ increases beyond $\tilde{c}$, the low type is the first to become marginal: he starts mixing between both technologies and AI only, while the high type still strictly prefers to use both technologies because his gain from using both technologies versus either own effort only or AI only is larger. As $c$ increases beyond $c^{\dagger}$, the low type finds effort too costly and switches to AI only, whereas the high type still uses both technologies, generating a separating equilibrium.

Alongside these, for every $c < c^{\ddagger}$ there is also an equilibrium in which the high type randomizes between both technologies and AI while the low type uses AI. Here the high type is indifferent, so the reputational gain from the combined technology must equal $c$ exactly, and his mixing probability adjusts to make it so: using the combined technology more often raises his accuracy and separates him further from the low type, which raises the gain, so a higher $c$ is sustained by more use of the combined technology, not less. The equilibrium is dominated in accuracy by those in which the high type uses both technologies for sure. Once $c > c^{\ddagger}$, even the high type is no longer willing to bear the effort cost given the reputational benefits, and no equilibrium where at least one type uses both technologies survives. In that region, the analysis reverts to the baseline environment

with actions $\{O, AI\}$ only. A key feature of Proposition 5 is that the thresholds $\tilde{c}$, $c^{\dagger}$, and $c^{\ddagger}$ do not depend on the monitoring intensity $\beta$: because the manager observes only whether AI was used (and thus cannot distinguish $AI$ from $B$), monitoring affects incentives to use AI versus own effort, but not the structure of equilibria in which at least one type adopts both technologies.

**Theorem 2** (Optimal Monitoring With Complementarities)**.** *Suppose Assumption 1 holds. The optimal monitoring policy is characterized according to the cost of effort $c$:*

1. *If $c \leq c^{\ddagger}$, the manager optimally sets $\beta_C^{opt} = 0$. In every manager-optimal equilibrium for such $c$, the high type always uses the combined technology. The low type uses the combined technology or AI, possibly mixing between them.*

2. *If $c > c^{\ddagger}$, there is no equilibrium in which the combined technology is chosen by either type, for any $\beta \in [0, 1]$. In this region, the manager optimally sets*

$$\beta_C^{opt} = \begin{cases} \underline{\beta}^{sep} & \text{if } k\,\underline{\beta}^{sep} \leq \mu\big(q_H - q_{AI}\big), \\ 0 & \text{otherwise.} \end{cases}$$

   *In the manager-optimal equilibrium, in the first case, the high type uses his own technology and the low type uses AI. In the second case, both types use AI.*

When effort is cheap, complementarities between AI and effort are strong enough that the high type is willing to use the combined technology $B$ without any help from monitoring. In this region, giving the manager an extra monitoring instrument $\beta$ does not change behavior at all: the manager can already induce her preferred use of $B$ purely through reputational incentives, so any positive monitoring only adds cost $k\beta$ without improving accuracy. Hence she optimally sets $\beta_C^{opt} = 0$.

Once effort becomes sufficiently costly, equilibria in which at least one type adopts both technologies disappear altogether: neither type is willing to bear the cost of effort when it must be combined with AI. The problem reduces to the baseline environment with actions $\{O, AI\}$ only. In this region, the logic of Theorem 1 reappears: the manager either finds it worthwhile to pay for just enough monitoring to sustain separation $(O, AI)$, or, when doing so is too costly, she gives up on separation and accepts pooling on AI. Therefore, complementarities expand the parameter region in which the manager can implement her preferred outcome with no monitoring.

## 5.4 Comparison to the Efficient Benchmark

We now compare the manager–optimal outcome in Theorem 2 to the efficient benchmark in Proposition 4. Recall that the planner always uses AI and asks type $\theta$ to exert effort if and only if

the gain in precision $q_{\theta B} - q_{AI}$ exceeds the cost $c$. Let $c^e_\theta \coloneqq q_{\theta B} - q_{AI}$, and note that $\tilde{c} \leq c^e_L$ since $\Delta b \leq 1$. In what follows we describe the case $c^e_L < c^\ddagger$. The remaining case is analogous, with the interval $[c^e_L, c^\ddagger)$ empty.

When effort is cheap, complementarities partially eliminate distortions. For $c \leq \tilde{c}$ the manager–optimal equilibrium coincides with the efficient allocation. As $c$ increases such that $c \in (\tilde{c}, c^e_L)$, the planner would still like both types to use both technologies, but the manager cares only about accuracy and not about the private cost $c$. She is therefore willing to sustain equilibria in which the high type uses both technologies while the low type either mixes between both technologies and AI or uses AI only (Proposition 5). In this intermediate region, the low-ability worker exerts too little effort relative to the efficient benchmark.

Once $c$ is large enough such that $c \in [c^e_L, c^\ddagger)$, both the planner and the manager want only the high type to exert effort. The efficient allocation is that the high type uses both technologies and the low type only AI. The manager–optimal equilibrium also implements this allocation with $\beta_C^{opt} = 0$. Hence, over this range, the use of both AI and effort is again undistorted.

For $c > c^\ddagger$ no equilibrium where at least one type uses both technologies exists, while the planner would still like the high type to use $B$ as long as $c \leq c^e_H$. The manager is constrained to the baseline action set and faces the same trade-off as in Theorem 1: she can pay for monitoring to induce separation, or, if monitoring is too costly, set $\beta_C^{opt} = 0$ and accept pooling on AI. In this high-cost region the high-ability worker exerts too little effort relative to the efficient benchmark, and the manager may rationally choose an outcome with strictly more AI use and lower accuracy than a frictionless planner would implement.

# 6 Outcome-Contingent Monitoring

In this section, we consider a more general monitoring policy in which monitoring depends on whether the worker's output is correct. Unlike continuous supervision, this form of monitoring resembles an audit: the manager observes the final outcome and then retrospectively investigates the worker's actions. Let $\beta_T$ denote the probability of monitoring when the task is completed correctly, and $\beta_F$ the probability when it is completed incorrectly. When these probabilities differ, the worker may have an incentive to misreport his signal, producing an incorrect outcome deliberately.[11] A strategy therefore specifies a technology and a report as a function of the worker's signal, and

[11] A microfoundation for this idea is to return to the state-based formulation of the task. Suppose the true state is $\omega \in \{0, 1\}$ and the technology chosen by the worker provides a noisy signal $s \in \{0, 1\}$ that matches the true state with a type- and technology-dependent precision $q(\theta, a)$. If the worker reports $1 - s$ instead of $s$, he can deliberately flip the correctness of the outcome.

we restrict attention throughout to equilibria in which the worker reports truthfully. Part (ii) of Proposition 6 below is exactly what makes truthful reporting a best response. The question is what the optimal audit looks like: whether the manager should audit after success, after failure, or both, and whether doing so improves on the symmetric policy of Theorem 1. As before, the manager's problem is to choose the cheapest policy that sustains her preferred equilibrium. We first characterize the policies that sustain separation and then select the cheapest among them.[12]

## 6.1 Sustaining Separation

Consider the separating profile in which the high type exerts effort, the low type uses AI, and both report truthfully. Beliefs are as in the baseline: $b_r$ and $b_w$ when unmonitored and, when monitored, 1 after effort and 0 after AI. Unless stated otherwise, $\Delta b$, $b_r$ and $b_w$ are evaluated at this profile and we suppress the argument.

The result features two key objects. The first is the *effective audit probability* of a type, $\bar{\beta}^{\theta}(\beta_T, \beta_F)$: the symmetric monitoring probability that would give type $\theta$ the same incentive to exert effort as the policy $(\beta_T, \beta_F)$. It is a weighted average of $\beta_T$ and $\beta_F$, with weights, given in Appendix A, that measure how much an audit after each outcome raises type $\theta$'s return to effort. Under a symmetric policy, this is equal for both types. The second is the *value of being right* to a worker using technology $a$, the difference in expected reputation from being right versus wrong.

$$V^{O}(\beta_T, \beta_F) \coloneqq \Delta b + \beta_T(1 - b_r) - \beta_F(1 - b_w), \qquad V^{AI}(\beta_T, \beta_F) \coloneqq \Delta b - \beta_T\, b_r + \beta_F\, b_w.$$

An audit after success raises the value of being right for a worker exerting effort and lowers it for one using AI, and an audit after failure does the reverse. Under a symmetric policy both values are positive.

**Proposition 6** (Separation under outcome-contingent monitoring)**.** *The separating equilibrium with truthful reporting exists under $(\beta_T, \beta_F)$ if and only if*

*(i) the high type's effective audit probability is at least $\underline{\beta}^{sep}$ and the low type's is at most $\bar{\beta}^{sep}$: $\bar{\beta}^{H}(\beta_T, \beta_F) \geq \underline{\beta}^{sep}$ and $\bar{\beta}^{L}(\beta_T, \beta_F) \leq \bar{\beta}^{sep}$, and*

*(ii) being right is worth more than being wrong under either technology: $V^{O}(\beta_T, \beta_F) \geq 0$ and $V^{AI}(\beta_T, \beta_F) \geq 0$.*

[12] Pooling on effort, by contrast, is never implemented: it is dominated in accuracy by separation, and the constraints that sustain it are recorded in Appendix A, where it is shown never to be chosen.

Condition (i) is that of separation in Proposition 2 with each type's effective audit probability in place of the common one: enough auditing for the high type to exert effort, not so much that the low type does too. Condition (ii) is new and bounds the asymmetry of the policy. If failures are audited much more often than successes, a high type is better off misreporting a correct signal so as to be audited and seen exerting effort. If successes are audited much more often than failures, a low type is better off misreporting a correct signal so as not to be audited and seen using AI. Note that under a symmetric policy both values in (ii) are positive with room to spare, so condition (ii) is an additional constraint that binds only under an asymmetric policy.

**Corollary 2.** *Auditing only failures sustains separation only if $\beta_F \leq \Delta b/(1 - b_w)$, and auditing only successes only if $\beta_T \leq \Delta b/b_r$. Both bounds are below one. In particular, the rule that audits every failure and no success never sustains separation. More generally, no audit that sustains separation audits either outcome with certainty.*

Note that $c^*$ is the threshold of Theorem 1: when effort is cheap enough that reputation alone sustains separation, the manager audits nothing, and the extra freedom to condition on the outcome is of no use to her.

This contrasts with the standard prescription in the literature on costly monitoring. When the worker is paid through transfers, an audit is an instrument of detection, and it is optimal to concentrate it where a problem is most likely: Dye (1986) shows that even when randomized policies are available, the optimal one is deterministic and lower-tailed, investigating the worker exactly when output falls below a threshold. With two outcomes that rule is simply to audit after failure. Two-tailed investigations can be optimal in that literature as well (Young, 1986), but always for the same reason: an audit is worth running where it is most likely to uncover a problem. Under reputational incentives, an audit is instead an instrument of revelation. An audit after failure reveals the high type and so rewards him, and a policy that inspects failures too often invites him to produce one. Failures can therefore be audited only so often, and never with certainty. The same logic appears in Theorem 1. There, the natural policy of disclosing all AI use is never optimal, because separation requires far less than full disclosure for the high type to exert effort, while full disclosure gives the low type enough reason to pool with him. Here, the natural policy of auditing every failure is not even feasible, because it makes failure something the high type wants. In both cases, the instrument that would be ideal if workers were passive is the one they manipulate. Each condition in Proposition 6 is linear in $(\beta_T, \beta_F)$, so the set of policies that sustain separation is an intersection of half-planes with the unit square, and hence a convex polygon. The manager's cost is linear too, so it is minimized at one of the polygon's vertices, and the cheapest separating audit sets at least one of the two probabilities at a bound. Figure 3 below shows the polygon, and the audit that minimizes

cost on it, for three sets of parameters.

## 6.2 The Optimal Audit

Return now to the manager's problem. Under an outcome-contingent policy the expected cost of monitoring is $k\,[\,P_r\,\beta_T + P_w\,\beta_F\,]$, where $P_r$ is the probability of a correct outcome in the equilibrium implemented and $P_w = 1 - P_r$. Two channels compete in the choice of which outcome to audit. An audit after success *rewards* effort: it lifts the reputation of a correct outcome from $b_r$ to one, which is worth more the rarer high types are. An audit after failure *punishes* AI: it drops the reputation of an incorrect outcome from $b_w$ to zero, which is worth more the more common high types are, and is cheap because failures are then rare. Let $\hat{\mu} \in (0,1)$ be the prior at which the two channels are equally cost-effective under separation.[13]

Besides pooling on AI, the equilibria the manager might implement all have the high type exerting effort. They differ in how often the low type uses AI. Write $\sigma_\alpha$ for the profile in which the high type exerts effort, the low type uses AI with probability $\alpha \in [0,1]$, and both report truthfully, so that $\sigma_1$ is separation, and write $C(\alpha)$ for the lowest expected monitoring cost at which $\sigma_\alpha$ can be sustained. By *partial separation* we mean a $\sigma_\alpha$ with $\alpha \in (0,1)$: the high type exerts effort and the low type mixes, using AI with probability $\alpha$ and exerting effort otherwise, at which he is indifferent between the two. Three thresholds on the effort cost also appear below, all given in closed form in Appendix A: $c^*$, below which separation requires no auditing at all, and $c_S$ and $c_F$, below which auditing successes alone, respectively failures alone, is enough to sustain it.

**Theorem 3** (Optimal Audits)**.** *The manager optimally sets*

$$\left(\beta_T^{opt}, \beta_F^{opt}\right) = \begin{cases} (\beta_T^S, \beta_F^S) & \text{if } k\,C(\alpha^{opt}) \le W(\alpha^{opt}) - q_{AI} \text{ and } k \le \underline{k}, \\ (\beta_T^P, \beta_F^P) & \text{if } k\,C(\alpha^{opt}) \le W(\alpha^{opt}) - q_{AI} \text{ and } k > \underline{k}, \\ (0,0) & \text{otherwise,} \end{cases}$$

*where $\alpha^{opt}$ maximizes $W(\alpha) - k\,C(\alpha)$ over $\alpha \in [0,1]$ and equals one exactly when $k \le \underline{k}$, and $\underline{k}$ and the audit policies $(\beta_T^S, \beta_F^S)$ and $(\beta_T^P, \beta_F^P)$ are defined in Appendix A.*[14] *In the manager-optimal equilibrium, in the first case, the high type uses his own technology and the low type uses AI. In the*

[13] Equivalently, $\hat{\mu}$ is the prior at which the odds of a correct outcome equal the geometric mean of the two types' odds of being right, $P_r/P_w = \sqrt{q_H q_{AI}/\big((1-q_H)(1-q_{AI})\big)}$. See the proof of Theorem 3.

[14] The threshold $\underline{k}$ is the monitoring cost above which some profile in which the low type mixes is cheap enough to sustain to outweigh the accuracy it gives up.

*second case, the high type uses his own technology and the low type mixes. In the third case, both types use AI.*

**Corollary 3.** *The separating audit* $(\beta_T^S, \beta_F^S)$ *has the following properties.*

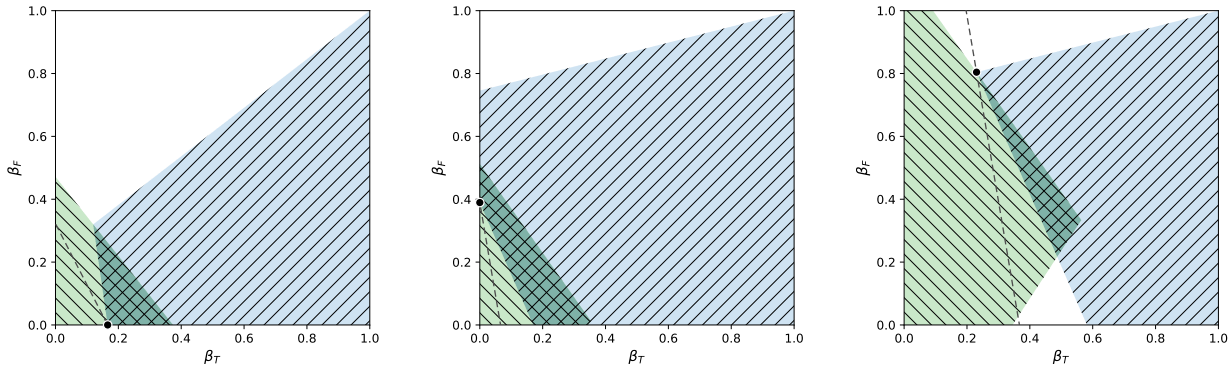
*(i) It targets successes if* $\mu \leq \hat{\mu}$ *and failures if* $\mu > \hat{\mu}$*, and is one-sided whenever that outcome suffices, which it does if* $c \leq c_S$ *(respectively* $c \leq c_F$*). Otherwise it audits the other outcome as well, and only to the point at which the type tempted to misreport is indifferent. Both components are zero if and only if* $c \leq c^*$*.*

*(ii) Both components are nondecreasing in the effort cost, both are strictly below one, and the targeted component strictly exceeds the other.*

*(iii) It is strictly cheaper than symmetric monitoring whenever* $c > c^*$ *and* $\mu \neq \hat{\mu}$*, so the set of parameters on which the manager monitors rather than accepting pooling on AI is strictly larger than in Theorem 1. The partial-separation audit tilts further toward failures,* $\beta_F^P > \beta_F^S$ *and* $\beta_T^P < \beta_T^S$*, and the manager never implements pooling on effort or an equilibrium in which the high type mixes.*

The optimal audit targets successes when $\mu \leq \hat{\mu}$ and failures when $\mu > \hat{\mu}$, and the effort cost decides how much of it is needed: it is zero below $c^*$, one-sided up to $c_S$ (respectively $c_F$), and two-sided beyond. It is in the last of these cases that mixing can arise: the second case of Theorem 3 is nonempty whenever $\mu > \hat{\mu}$ and $c > \max\{c_F, \hat{c}\}$ for a threshold $\hat{c}$ given in Appendix A, that is, when high-ability workers are common and effort is costly enough that the audit is capped by the high type's incentive to misreport. The one-sided audits of part (i) are $(\beta_T^S, \beta_F^S) = ((c - c^*)/w_T^H,\ 0)$ and $(0,\ (c - c^*)/w_F^H)$, with $w_T^H = q_H q_{AI}/P_r$ and $w_F^H = (1 - q_H)(1 - q_{AI})/P_w$. Its cost $C(1)$, which enters the manager's decision whether to monitor at all, is given in closed form in each case in Appendix A. Figure 3 shows the cheapest separating audit in each of these cases. The partial-separation audit $(\beta_T^P, \beta_F^P)$ solves the low type's indifference condition together with the high type's truthful-reporting condition at the beliefs of $\sigma_\alpha$.

The outcome optimally targeted by the manager for auditing reverses across the prior. When high-ability workers are rare, the optimal audit resembles a promotion review, triggered by success and rewarding those who earned it on their own. When they are common, it resembles the costly-state-verification rule, triggered by failure and catching those who leaned on the tool, though never with certainty, since past a certain point the high type would rather misreport a correct signal and be audited than report it and be trusted. The type tempted to misreport is always the one the audit does not target. Under failure audits it is the high type, for whom an audit after failure is a reward, so he




(a) $\mu = 0.2 < \hat{\mu}$, $c = 0.2$ (b) $\mu = 0.85 > \hat{\mu}$, $c = 0.2$ (c) $\mu = 0.85 > \hat{\mu}$, $c = 0.46$

Figure 3: The cheapest separating audit. The region hatched with forward slashes is where the high type exerts effort and reports truthfully, the region hatched with backslashes is where the low type uses AI and reports truthfully, and their intersection is the set of audits that sustain separation. The dashed line is one of the manager's isocost lines, along which her expected monitoring cost is constant, and the large dot marks the cheapest audit on that set. Throughout, $q_H = 0.9$, $q_{AI} = 0.6$ and $q_L = 0.55$, so that $\hat{\mu} = 0.62$. In panel (a) high-ability workers are rare and the manager audits successes only. In panel (b) they are common and she audits failures only. In panel (c) they are common and effort is costly enough that failures alone cannot sustain separation, so she adds success audits until the high type is just willing to report truthfully.

would misreport a correct signal to obtain one. Under success audits it is the low type, for whom an audit after success is a punishment, so he would misreport a correct signal to avoid one. When the temptation binds, the manager does not audit the targeted outcome less. Instead she adds audits of the other outcome, just enough that being right is again worth as much as being wrong to the tempted type, and the audit becomes targeted towards one side, with a smaller chance of auditing the other outcome. And because the concentrated audit is cheaper, the manager sustains separation at monitoring costs that would have made her accept pooling under symmetric monitoring.

Partial separation is the one addition to the set of implemented profiles from the baseline case. For $\alpha < 1$ the low type must be exactly indifferent between AI and effort, so tolerating a little AI avoidance costs strictly more than separation unless the cheapest separating audit already leaves him indifferent, which happens only when the high type's truthful-reporting constraint binds, $\mu > \hat{\mu}$ and $c > c_F$. The reason is that an audit after failure certifies a high type, so the manager cannot audit failures so often that he would rather misreport a correct signal than report it. If the low type exerts effort with small probability, an audit after effort no longer certifies anyone: it finds a high type most of the time, but not always. Misreporting then pays the high type less, his truthful-reporting constraint slackens, and failures can be audited harder at a lower total cost, since failures are rare. Against that saving, $C'(1^-)$ per unit, the manager weighs the accuracy lost, $(1-\mu)(q_{AI} - q_L)$ per

unit of the low type's effort. The saving is positive if and only if $c > \hat{c}$. When it is, $\underline{k}$ is finite and the manager tolerates mixing once $k$ exceeds it. This is the only channel through which mixing can pay: in every other case the low type strictly prefers AI at the cheapest separating audit, and making him indifferent requires more monitoring, not less.

# 7 Conclusion

This paper studies how an accuracy-seeking manager can optimally monitor AI use when workers care about their reputations. A worker of privately known ability chooses between his own costly effort and a free AI tool. The manager observes outcomes and can invest in monitoring that sometimes reveals whether AI was used. Because the worker's payoff is purely reputational, the manager implements behavior indirectly through monitoring and beliefs.

A natural benchmark is a world without AI, in which output reflects only human effort. Introducing AI expands the set of technologies, so a planner who can dictate behavior is never made worse off by it: at worst she ignores the tool. The manager cannot dictate behavior. She faces two constraints the planner does not, reputational incentives and costly monitoring, and what she obtains depends on both. When effort is cheap or AI is not too accurate, reputation alone sustains separation, and AI improves accuracy relative both to the no-AI world and to a ban. When separation needs monitoring, she buys the least that sustains it. When monitoring is too expensive, she gives up on separation and lets everyone use AI. If high-ability workers are common, accuracy is then lower than it would have been without AI, because the effort that revealed them is no longer worth paying for.

The same pattern appears in both extensions. When effort and AI are technological complements, a planner would always want AI in the background and would add human effort whenever the accuracy gain justifies the cost. Yet in the principal–agent setting, a combined-technology equilibrium often exists in the absence of monitoring, relying purely on reputational incentives. Only when effort is sufficiently costly does the manager resort to monitoring, and again it is used as sparingly as possible. Allowing monitoring probabilities to differ after correct versus incorrect outcomes introduces scope for strategic misreporting and more intricate incentive constraints, but the basic message remains: with reputational concerns and costly, outcome-dependent monitoring, equilibrium behavior can involve both too much and too little effort, and too much reliance on AI, relative to what would be chosen under direct control of technologies.

A natural next step is to let the worker care about output as well as reputation, moving from the pure career-concerns payoff studied here to a mixture of the two, and to trace how the distortions

in AI adoption change as reputational concerns become less important. The incentive structure of Lemma 1 carries over with an additional term for the direct value of accuracy.

Overall, the analysis suggests that the key policy question is not simply whether to adopt AI or ban it, but how AI, monitoring, and reputation interact. Our results indicate that a well-designed policy often features no monitoring of technology use, and when it is present, it is deployed in a focused way to support desirable behavior. Although a planner with direct control would always benefit from AI, a manager who must work through reputational incentives may sometimes prefer the discipline of a no-AI world.

# References


Acemoglu, D. and Restrepo, P. (2018). Artificial intelligence, automation, and work. In Agrawal, A., Gans, J., and Goldfarb, A., editors, *The Economics of Artificial Intelligence: An Agenda*, pages 197–236. University of Chicago Press.

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

Agrawal, A., Gans, J. S., and Goldfarb, A. (2019). Artificial intelligence: The ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, 33(2):31–50.

Almog, D. (2025). Barriers to AI adoption: Image concerns at work. arXiv:2511.18582.

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020). The allocation of decision authority to human and artificial intelligence. *AEA Papers and Proceedings*, 110:80–84.

Atlassian Teamwork Lab (2026). New research shows honesty about AI use at work is backfiring. Atlassian Teamwork Lab report, June 2026. https://www.atlassian.com/blog/ai-at-work/new-research-shows-honesty-about-ai-use-at-work-is-backfiring.

Baiman, S. and Demski, J. S. (1980). Economically optimal performance evaluation and control systems. *Journal of Accounting Research*, 18:184–220.

Baranchuk, N. and Rivera, A. (2026). Successfully fired: The unique incentives of agentic-AI adoption. Working Paper 00205-00, Finance Theory Group.

Border, K. C. and Sobel, J. (1987). Samurai accountant: A theory of auditing and plunder. *Review of Economic Studies*, 54(4):525–540.

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

Chevalier, J. and Ellison, G. (1999). Career concerns of mutual fund managers. *Quarterly Journal of Economics*, 114(2):389–432.

Dell'Acqua, F., McFowland III, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School.

Dewatripont, M., Jewitt, I., and Tirole, J. (1999). The economics of career concerns, part I: Comparing information structures. *Review of Economic Studies*, 66(1):183–198.

Dye, R. A. (1986). Optimal monitoring policies in agencies. *RAND Journal of Economics*, 17(3):339–350.

Ezra, T., Leonardi, S., and Russo, M. (2024). Contracts with inspections. arXiv:2402.16553.

Fallah, A. and Jordan, M. (2024). Contract design with safety inspections. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 616–638.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24(3):461–483.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *Review of Economic Studies*, 66(1):169–182.

Holmström, B. and Ricart i Costa, J. (1986). Managerial incentives and capital management. *Quarterly Journal of Economics*, 101(4):835–860.

Ide, E. and Talamás, E. (2024). The turing valley: How AI capabilities shape labor income. arXiv:2408.16443.

Ide, E. and Talamás, E. (2025). Artificial intelligence in the knowledge economy. *Journal of Political Economy*, 133(12).

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12(2):380–391.

Mookherjee, D. and Png, I. (1989). Optimal auditing, insurance, and redistribution. *Quarterly Journal of Economics*, 104(2):399–415.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Otis, N. G., Clarke, R., Delecourt, S., Holtz, D., and Koning, R. (2026). The uneven impact of generative AI on entrepreneurial performance: Evidence from a field experiment in Kenya. *Management Science*. Forthcoming; doi:10.1287/mnsc.2024.06909.

Ottaviani, M. and Sørensen, P. N. (2006). Reputational cheap talk. *RAND Journal of Economics*, 37(1):155–175.

Pew Research Center (2025). About 1 in 5 U.S. workers now use AI in their job, up since last year.

Prendergast, C. and Stole, L. (1996). Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning. *Journal of Political Economy*, 104(6):1105–1134.

Reif, J. A., Larrick, R. P., and Soll, J. B. (2025). Evidence of a social evaluation penalty for using AI. *Proceedings of the National Academy of Sciences*, 122(19):e2426766122.

Scharfstein, D. S. and Stein, J. C. (1990). Herd behavior and investment. *American Economic Review*, 80(3):465–479.

Slack (2024). The fall 2024 workforce index shows executives and employees investing in ai, but uncertainty holding back adoption.

Townsend, R. M. (1979). Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory*, 21(2):265–293.

Weitzner, G. (2024). Reputational algorithm aversion. Working paper, McGill University, arXiv:2402.15418.

Xu, F., Hou, J., Chen, W., and Xie, K. (2025). Generative AI and organizational structure in the knowledge economy. arXiv:2506.00532.

Yang, C. L., Restrepo-Amariles, D., Allen, L., and Troussel, A. (2025). Gpt adoption dilemma and the impact of disclosure policies. Technical report, HEC Paris.

Young, R. A. (1986). A note on "economically optimal performance evaluation and control systems": The optimality of two-tailed investigations. *Journal of Accounting Research*, 24(1):231–240.

Zwiebel, J. (1995). Corporate conservatism and relative compensation. *Journal of Political Economy*, 103(1):1–25.

# A Proofs

## Proof of Lemma 1

*Proof.* $U^\theta(O,\beta) - U^\theta(AI,\beta) = \rho^\theta_O(\beta) - \rho^\theta_{AI}(\beta) - c$, and collecting the unmonitored and monitored terms of the difference in expected reputations gives $(1-\beta)(q_\theta - q_{AI})(b_r - b_w) + \beta\,\Pi_\theta(\alpha)$. □

## Proof of Proposition 1

*Proof.* Recall that $\alpha_\theta$ denotes the probability that type $\theta$ uses AI. The planner maximizes expected accuracy net of effort costs by choosing $\alpha_H, \alpha_L \in [0,1]$ to solve

$$\max_{\alpha_H,\alpha_L\in[0,1]} \mu\Big(\alpha_H q_{AI} + (1-\alpha_H)(q_H - c)\Big) + (1-\mu)\Big(\alpha_L q_{AI} + (1-\alpha_L)(q_L - c)\Big).$$

The objective is additively separable in $(\alpha_H, \alpha_L)$, so we can optimize separately for each type.

**Low-ability type.** The term involving $\alpha_L$ is

$$(1-\mu)\Big(\alpha_L q_{AI} + (1-\alpha_L)(q_L - c)\Big) = (1-\mu)\Big[(q_L - c) + \alpha_L\big(q_{AI} - (q_L - c)\big)\Big].$$

The coefficient of $\alpha_L$ is

$$q_{AI} - (q_L - c) = (q_{AI} - q_L) + c > 0,$$

since $q_{AI} > q_L$ and $c > 0$. Thus the objective is strictly increasing in $\alpha_L$ and is maximized by $\alpha_L^* = 1$.

**High-ability type.** The term involving $\alpha_H$ is

$$\mu\Big(\alpha_H q_{AI} + (1-\alpha_H)(q_H - c)\Big) = \mu\Big[(q_H - c) + \alpha_H\big(q_{AI} - (q_H - c)\big)\Big].$$

The coefficient of $\alpha_H$ is

$$q_{AI} - (q_H - c) = q_{AI} - q_H + c.$$

If $c \leq q_H - q_{AI}$, then $q_{AI} - q_H + c \leq 0$, so the objective is weakly decreasing in $\alpha_H$ and is maximized at $\alpha_H^* = 0$ (the planner optimally assigns the high type to own effort).

If $c > q_H - q_{AI}$, then $q_{AI} - q_H + c > 0$, so the objective is strictly increasing in $\alpha_H$ and is maximized at $\alpha_H^* = 1$ (the planner optimally assigns the high type to AI).

At $c = q_H - q_{AI}$, both $\alpha_H = 0$ and $\alpha_H = 1$ yield the same value, and we adopt the convention $\alpha_H^* = 0$ in this knife-edge case, as in the statement of the proposition.

Putting the two types together yields

$$\alpha_H^* = \begin{cases} 0 & \text{if } c \leq q_H - q_{AI}, \\ 1 & \text{if } c > q_H - q_{AI}, \end{cases} \qquad \alpha_L^* = 1,$$

which proves the result. □

## Proof of Proposition 2: Pure-Strategy Profiles

Throughout the appendix, profiles are written as $\alpha = (\alpha_H, \alpha_L)$ and beliefs are evaluated at the candidate profile. Each pure-strategy profile is checked by evaluating $\Delta b(\alpha)$ and the effort premia $\Pi_\theta(\alpha)$ and applying Lemma 1: type $\theta$ plays $O$ in equilibrium iff (1) holds and plays $AI$ iff it fails. We use throughout that, writing $x$ for the low type's accuracy at fixed high-type accuracy, $b_r = q_H\mu/(q_H\mu + x(1-\mu))$ is strictly decreasing and $b_w = (1-q_H)\mu/((1-q_H)\mu + (1-x)(1-\mu))$ strictly increasing in $x$, so that $\Delta b$ is strictly decreasing in the low type's accuracy. Symmetrically, it is strictly increasing in the high type's accuracy.

### Separating profile $(1,0)$: high type uses AI, low type uses own technology

Monitoring identifies types, but in reverse: $b'_{AI} = 1$ and $b'_{O,r} = b'_{O,w} = 0$, so $\Pi_\theta = -1$ for both types, while $\Delta b(1,0) > 0$. For the low type, (1) requires $c \leq (1-\beta)(q_L - q_{AI})\Delta b - \beta < 0$, which fails. Hence $(1,0)$ is never an equilibrium.

### Separating profile $(0,1)$: high type uses own technology, low type uses AI

**Lemma 3.** *$(\alpha_H, \alpha_L) = (0,1)$ is a perfect Bayesian equilibrium if and only if $\beta \in [\underline{\beta}^{sep}, \bar{\beta}^{sep}]$, where $0 \leq \underline{\beta}^{sep} < \bar{\beta}^{sep} < 1$.*

*Proof.* Monitoring identifies types: $b'_{O,r} = b'_{O,w} = 1$ and $b'_{AI} = 0$, so $\Pi_H = \Pi_L = 1$. Write $\Delta b \coloneqq \Delta b(0,1) = \frac{\mu(1-\mu)(q_H - q_{AI})}{[q_{AI}(1-\mu)+q_H\mu][(1-q_H)\mu+(1-q_{AI})(1-\mu)]} \in (0,1)$. Note that $(q_H - q_{AI})\Delta b < 1$.

For the high type, (1) reads $c \leq (1-\beta)(q_H - q_{AI})\Delta b + \beta$, i.e. $\beta[1 - (q_H - q_{AI})\Delta b] \geq c - (q_H - q_{AI})\Delta b$. This holds for all $\beta \in [0,1]$ if $c \leq (q_H - q_{AI})\Delta b$, and otherwise iff $\beta \geq \underline{\beta}^{sep}$, where

$$\underline{\beta}^{sep} \equiv \max\left\{0, \frac{c - (q_H - q_{AI})\Delta b}{1 - (q_H - q_{AI})\Delta b}\right\} < 1. \tag{2}$$

For the low type, (1) must fail: $c > (1-\beta)(q_L - q_{AI})\Delta b + \beta$, i.e. $\beta[1 + (q_{AI} - q_L)\Delta b] < c + (q_{AI} - q_L)\Delta b$. Allowing indifference, this is

$$\beta \leq \frac{c + (q_{AI} - q_L)\Delta b}{1 + (q_{AI} - q_L)\Delta b} \equiv \bar{\beta}^{sep} < 1. \tag{3}$$

Finally $\bar{\beta}^{sep} > \underline{\beta}^{sep}$: if $\underline{\beta}^{sep} = 0$ this is immediate, and otherwise it follows from $c < 1$ since $(c-x)/(1-x) < c < (c+y)/(1+y)$ for $x \in [0,1)$, $y > 0$. □

**Pooling profile $(1,1)$: both types use AI**

**Lemma 4.** *$(\alpha_H, \alpha_L) = (1,1)$ is a perfect Bayesian equilibrium for every $\beta \in [0,1]$.*

*Proof.* Correctness is uninformative, $\Delta b(1,1) = 0$, and $b'_{AI} = \mu$. By the pessimistic off-path belief $b'_{O,r} = b'_{O,w} = 0$, so $\Pi_\theta = -\mu$. Then (1) reads $c \leq -\beta\mu$, which fails for both types. □

**Pooling profile $(0,0)$: both types use own technology**

**Lemma 5.** *There exists $\bar{c} \in (0,1)$ such that $(\alpha_H, \alpha_L) = (0,0)$ is a perfect Bayesian equilibrium if and only if $c \leq \bar{c}$ and $\beta \in [\underline{\beta}^{pool}, 1]$, where $\underline{\beta}^{pool} \in (0,1]$.*

*Proof.* Observing $O$ reveals nothing beyond correctness, so $b'_{O,r} = b_r$ and $b'_{O,w} = b_w$ with $(b_r, b_w)$ evaluated at $(0,0)$, and by the pessimistic off-path belief $b'_{AI} = 0$. Hence $\Pi_\theta = q_\theta b_r + (1-q_\theta)b_w$ and, writing $\Delta b \coloneqq \Delta b(0,0)$, (1) for type $\theta$ reads

$$c \leq (1-\beta)(q_\theta - q_{AI})\Delta b + \beta\big[q_\theta b_r + (1-q_\theta)b_w\big] = (q_\theta - q_{AI})\Delta b + \beta\big[q_{AI} b_r + (1-q_{AI})b_w\big],$$

using $q_\theta b_r + (1-q_\theta)b_w = q_{AI} b_r + (1-q_{AI})b_w + (q_\theta - q_{AI})\Delta b$. The right-hand side is increasing in $q_\theta$, so the low type's constraint binds, and it holds iff

$$\beta \geq \frac{c + (q_{AI} - q_L)\Delta b}{q_{AI} b_r + (1-q_{AI})b_w} \equiv \underline{\beta}^{pool} > 0.$$

Since $\beta \leq 1$, an equilibrium exists iff $\underline{\beta}^{pool} \leq 1$, i.e. iff $c \leq q_{AI} b_r + (1 - q_{AI}) b_w - (q_{AI} - q_L)\Delta b = b_w + q_L \Delta b \equiv \bar{c}$. □

### Comparison of thresholds

Write $\delta^{sep} := \Delta b(0,1)$ and $\delta^{pool} := \Delta b(0,0)$. Since the low type is less accurate at $(0,0)$ than at $(0,1)$ ($q_L < q_{AI}$), $\delta^{pool} > \delta^{sep}$. Hence, using $q_{AI} b_r + (1 - q_{AI}) b_w \leq 1$,

$$\bar{\beta}^{sep} = \frac{c + (q_{AI} - q_L)\delta^{sep}}{1 + (q_{AI} - q_L)\delta^{sep}} < \frac{c + (q_{AI} - q_L)\delta^{pool}}{1 + (q_{AI} - q_L)\delta^{pool}} \leq \underline{\beta}^{pool},$$

which gives the ranking of thresholds in the proposition. This proves items 1–3 of the proposition and the ordering of thresholds. For Corollary 1(i)–(ii): item (ii) is the condition $\underline{\beta}^{pool} \leq 1$ of Lemma 5. For (i), $\underline{\beta}^{sep} = 0$ iff $c \leq (q_H - q_{AI})\delta^{sep} = c^*$, and where positive, $\underline{\beta}^{sep} = (c - x)/(1 - x)$ with $x = (q_H - q_{AI})\delta^{sep}$ is strictly increasing in $c$ and, since $x$ is strictly decreasing in $q_{AI}$ (both factors are), strictly increasing in $q_{AI}$.

## Proof of Proposition 2: Mixed-Strategy Profiles, and of Corollary 1(iii)

For mixed-strategy equilibria, both types cannot randomize between AI and their own technology. The indifference condition would require that the payoff from using AI equals the payoff from using one's own technology for both types simultaneously. This, in turn, implies that the payoff from using the worker's own technology must be the same across types. However, since the two types differ in their ability to predict the correct state, this cannot hold.

Therefore, we characterize two possible cases:

1. an equilibrium in which the high type mixes while the low type always chooses AI, and
2. an equilibrium in which the high type always uses his own technology while the low type mixes.

### High Type Mixes

**Lemma 6.** *There exist thresholds* $\underline{\beta}^{HM}$, $\bar{\beta}^{HM} \in [0,1]$ *such that* $(\alpha_H, \alpha_L) = (\alpha_H^*, 1)$ *with* $\alpha_H^* \in (0,1)$ *constitutes a Perfect Bayesian Equilibrium if and only if* $\beta \in [\underline{\beta}^{HM}, \bar{\beta}^{HM}]$.

*Proof.* Suppose the high type uses a mixed strategy while the low type only uses AI, i.e., $\alpha_H \in (0,1)$ and $\alpha_L = 1$. The manager's beliefs when she monitors the worker are $b'_{AI} = \frac{\mu\alpha_H}{\mu\alpha_H + 1 - \mu}$ and

$b'_{O,r} = b'_{O,w} = 1$. And her beliefs when she does not monitor the worker are

$$b_r = \frac{\mu[\alpha_H q_{AI} + (1-\alpha_H)q_H]}{\mu[\alpha_H q_{AI} + (1-\alpha_H)q_H] + (1-\mu)q_{AI}}$$

and

$$b_w = \frac{\mu[\alpha_H(1-q_{AI}) + (1-\alpha_H)(1-q_H)]}{\mu[\alpha_H(1-q_{AI}) + (1-\alpha_H)(1-q_H)] + (1-\mu)(1-q_{AI})}.$$

Here $\Pi_H = 1 - b'_{AI}$, so by Lemma 1 the high type is indifferent iff

$$c = \beta(1-b'_{AI}) + (1-\beta)(q_H - q_{AI})(b_r - b_w). \tag{4}$$

For any $\alpha_H \in (0,1)$, we solve $\beta^*$ from (4). When the high type is indifferent, the low type strictly prefers AI: his right-hand side in (1) is smaller, since $q_L < q_H$ and $\Pi_L = \Pi_H$.

Let the right-hand side of (4) be $F(\beta, \alpha_H)$.

$$\begin{aligned}
\frac{\partial F}{\partial \alpha_H} &= -\beta\frac{\partial b'_{AI}}{\partial \alpha_H} + (1-\beta)(q_H - q_{AI})\left(\frac{\partial b_r}{\partial \alpha_H} - \frac{\partial b_w}{\partial \alpha_H}\right) \\
&= -\beta\frac{\mu(1-\mu)}{(\mu\alpha_H + 1 - \mu)^2} + (1-\beta)(q_H - q_{AI})\frac{\mu(q_{AI} - q_H)(1-\mu)q_{AI}}{\{\mu[\alpha_H q_{AI} + (1-\alpha_H)q_H] + (1-\mu)q_{AI}\}^2} \\
&\quad - (1-\beta)(q_H - q_{AI})\frac{\mu(q_H - q_{AI})(1-\mu)(1-q_{AI})}{\{\mu[\alpha_H(1-q_{AI}) + (1-\alpha_H)(1-q_H)] + (1-\mu)(1-q_{AI})\}^2} \\
&< 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial F}{\partial \beta} &= 1 - b'_{AI} - (q_H - q_{AI})(b_r - b_w) \\
&= \frac{1-\mu}{\mu\alpha_H + 1 - \mu} - \frac{(q_H - q_{AI})^2(1-\alpha_H)\mu(1-\mu)}{D_r\, D_w} \\
&= \frac{(1-\mu)\left[q_{AI}(1-q_{AI}) + \mu(1-\alpha_H)(q_H - q_{AI})(1 - q_H - q_{AI})\right]}{(\mu\alpha_H + 1 - \mu)\, D_r\, D_w} \\
&> 0,
\end{aligned}$$

where $D_r \coloneqq \mu[\alpha_H q_{AI} + (1-\alpha_H)q_H] + (1-\mu)q_{AI}$ and $D_w \coloneqq \mu[\alpha_H(1-q_{AI}) + (1-\alpha_H)(1-q_H)] + (1-\mu)(1-q_{AI})$ are the two denominators appearing above, and the last inequality holds because $q_{AI}(1-q_{AI}) - (q_H - q_{AI})(q_H + q_{AI} - 1) = q_H(1-q_H) > 0$, so the numerator is at least $(1-\mu)\, q_H(1-q_H) > 0$.

Hence,

$$\frac{d\beta^*}{d\alpha_H} = -\frac{\partial F/\partial \alpha_H}{\partial F/\partial \beta} > 0.$$

This implies that for each $\alpha_H \in (0,1)$, there exists a unique $\beta^*$ solving (4), and moreover, $\beta^*$ is increasing in $\alpha_H$.

The limits of $\beta^*$ are obtained by evaluating (4) at $\alpha_H = 0$, where $b'_{AI} = 0$ and $(b_r, b_w)$ are the beliefs at the separating profile $(0,1)$, and at $\alpha_H = 1$, where $b'_{AI} = \mu$ and $b_r = b_w$:

$$\lim_{\alpha_H \to 0} \beta^* = \frac{c - (q_H - q_{AI})\Delta b(0,1)}{1 - (q_H - q_{AI})\Delta b(0,1)}$$

and

$$\lim_{\alpha_H \to 1} \beta^* = \frac{c}{1-\mu} > 0.$$

Since $\beta$ denotes the monitoring probability, it must lie within the interval $\beta \in [0,1]$. The equilibrium values of $\beta$ may not reach the theoretical limits if they fall outside this range. Therefore, define

$$\underline{\beta}^{HM} = \max\left\{0, \frac{c - (q_H - q_{AI})\Delta b(0,1)}{1 - (q_H - q_{AI})\Delta b(0,1)}\right\}$$

and

$$\bar{\beta}^{HM} = \min\left\{1, \frac{c}{1-\mu}\right\}.$$

For any $\beta \in [\underline{\beta}^{HM}, \bar{\beta}^{HM}]$, there exists a unique equilibrium in which the high type adopts AI with probability $\alpha_H$, while the low type always uses AI.

Note that $\underline{\beta}^{HM} = \underline{\beta}^{sep}$. □

**Low Type Mixes**

**Lemma 7.** *There exist thresholds $\underline{\beta}^{LM}$, $\bar{\beta}^{LM} \in [0,1]$ such that $(\alpha_H, \alpha_L) = (0, \alpha_L^*)$ with $\alpha_L^* \in (0,1)$ constitutes a Perfect Bayesian Equilibrium if and only if $\beta \in [\underline{\beta}^{LM}, \bar{\beta}^{LM}]$.*

*Proof.* Suppose the low type uses a mixed strategy while the high type only uses own technology, i.e., $\alpha_H = 0$ and $\alpha_L \in (0,1)$. The manager's belief when she does not monitor the worker are

$$b_r = \frac{\mu q_H}{\mu q_H + (1-\mu)[\alpha_L q_{AI} + (1-\alpha_L) q_L]}$$

and

$$b_w = \frac{\mu(1-q_H)}{\mu(1-q_H) + (1-\mu)[\alpha_L(1-q_{AI}) + (1-\alpha_L)(1-q_L)]}.$$

And her belief when she monitors the worker and the worker uses AI is $b'_{AI} = 0$, while the belief

also depends on the correctness of the prediction when the worker uses own technology, i.e.,

$$b'_{O,r} = \frac{\mu q_H}{\mu q_H + (1-\mu)(1-\alpha_L)q_L}$$

and

$$b'_{O,w} = \frac{\mu(1-q_H)}{\mu(1-q_H) + (1-\mu)(1-\alpha_L)(1-q_L)}.$$

Here $\Pi_L = q_L b'_{O,r} + (1-q_L)b'_{O,w}$, so by Lemma 1 the low type is indifferent iff

$$c = \beta[q_L b'_{O,r} + (1-q_L)b'_{O,w}] - (1-\beta)(q_{AI} - q_L)(b_r - b_w). \tag{5}$$

For any $\alpha_L \in (0,1)$, we solve $\beta^*$ from (5). When the low type is indifferent, the high type strictly prefers own technology: his right-hand side in (1) is larger, since $q_H > q_L$ and $b'_{O,r} > b'_{O,w}$.

Let the right-hand side of (5) be $F(\beta, \alpha_L)$.

$$\begin{aligned}
\frac{\partial F}{\partial \alpha_L} &= \beta q_L \frac{\partial b'_{O,r}}{\partial \alpha_L} + \beta(1-q_L)\frac{\partial b'_{O,w}}{\partial \alpha_L} - (1-\beta)(q_{AI}-q_L)\left(\frac{\partial b_r}{\partial \alpha_L} - \frac{\partial b_w}{\partial \alpha_L}\right) \\
&= \beta q_L \frac{\mu q_H (1-\mu) q_L}{[\mu q_H + (1-\mu)(1-\alpha_L)q_L]^2} + \beta(1-q_L)\frac{\mu(1-q_H)(1-\mu)(1-q_L)}{[\mu(1-q_H) + (1-\mu)(1-\alpha_L)(1-q_L)]^2} \\
&+ (1-\beta)(q_{AI}-q_L)\frac{\mu q_H (q_{AI}-q_L)(1-\mu)}{\{\mu q_H + (1-\mu)[\alpha_L q_{AI} + (1-\alpha_L)q_L]\}^2} \\
&+ (1-\beta)(q_{AI}-q_L)\frac{\mu(q_{AI}-q_L)(1-\mu)(1-q_H)}{\{\mu(1-q_H) + (1-\mu)[\alpha_L(1-q_{AI}) + (1-\alpha_L)(1-q_L)]\}^2} \\
&> 0
\end{aligned}$$

$$\frac{\partial F}{\partial \beta} = q_L b'_{O,r} + (1-q_L)b'_{O,w} + (q_{AI}-q_L)(b_r - b_w) > 0$$

Hence,

$$\frac{d\beta^*}{d\alpha_L} = -\frac{\partial F/\partial \alpha_L}{\partial F/\partial \beta} < 0.$$

This implies that for each $\alpha_L \in (0,1)$, there exists a unique $\beta^*$ solving (5), and moreover, $\beta^*$ is decreasing in $\alpha_L$.

The limits of $\beta^*$ are obtained by evaluating (5) at $\alpha_L = 0$, where $b'_{O,r} = b_r$ and $b'_{O,w} = b_w$ are the beliefs at the pooling profile $(0,0)$, and at $\alpha_L = 1$, where $b'_{O,r} = b'_{O,w} = 1$ and $(b_r, b_w)$ are the

beliefs at the separating profile $(0,1)$:

$$\lim_{\alpha_L \to 0} \beta^* = \frac{c + (q_{AI} - q_L)\Delta b(0,0)}{q_{AI} b_r(0,0) + (1 - q_{AI}) b_w(0,0)}$$

and

$$\lim_{\alpha_L \to 1} \beta^* = \frac{c + (q_{AI} - q_L)\Delta b(0,1)}{1 + (q_{AI} - q_L)\Delta b(0,1)} > 0.$$

Since $\beta$ denotes the monitoring probability, it must lie within the interval $\beta \in [0,1]$. The equilibrium values of $\beta$ may not reach the theoretical limits if they fall outside this range. Therefore, define

$$\underline{\beta}^{LM} = \frac{c + (q_{AI} - q_L)\Delta b(0,1)}{1 + (q_{AI} - q_L)\Delta b(0,1)}$$

and

$$\bar{\beta}^{LM} = \min\left\{1, \frac{c + (q_{AI} - q_L)\Delta b(0,0)}{q_{AI} b_r(0,0) + (1 - q_{AI}) b_w(0,0)}\right\}.$$

For any $\beta \in [\underline{\beta}^{LM}, \bar{\beta}^{LM}]$, there exists a unique equilibrium in which the high type always uses own technology, while the low type adopts AI with probability $\alpha_L$. □

Note that $\underline{\beta}^{LM} = \bar{\beta}^{sep}$ and $\bar{\beta}^{LM} = \min\{1, \underline{\beta}^{pool}\}$.

## Proof of Theorem 1

*Proof.* Since $q_L < q_{AI} < q_H$, $W$ is maximized over $[0,1]^2$ at the separating profile $(0,1)$, with $W(0,1) - W(1,1) = \mu(q_H - q_{AI})$. By Proposition 2, the pooling equilibrium $(1,1)$ is the unique equilibrium for $\beta < \underline{\beta}^{sep}$, and $(0,1)$ is an equilibrium for every $\beta \in [\underline{\beta}^{sep}, \bar{\beta}^{sep}]$. Hence

$$A(\beta) = W(1,1) \ \text{ for } \beta < \underline{\beta}^{sep}, \qquad A(\beta) = W(0,1) \ \text{ for } \beta \in [\underline{\beta}^{sep}, \bar{\beta}^{sep}],$$

and $A(\beta) \le W(0,1)$ for all $\beta$. Since $k > 0$, the objective $A(\beta) - k\beta$ is therefore maximized either at $\beta = 0$, with value $W(1,1)$, or at $\beta = \underline{\beta}^{sep}$, with value $W(0,1) - k\underline{\beta}^{sep}$, and any other $\beta$ is dominated by one of the two. The manager monitors with probability $\underline{\beta}^{sep}$ iff $W(0,1) - k\underline{\beta}^{sep} \ge W(1,1)$, i.e. iff $k\underline{\beta}^{sep} \le \mu(q_H - q_{AI})$, with the separating equilibrium selected at indifference, and otherwise she does not monitor. When $\underline{\beta}^{sep} = 0$ the condition holds trivially and separation is obtained at no cost. □

## Proof of Proposition 3

*Proof.* Fix the effort cost $c$ and the monitoring cost $k > 0$, and regard everything as a function of the quality of AI. Write

$$x(q_{AI}) \;\coloneqq\; c^*(q_{AI}) \;=\; (q_H - q_{AI})\,\Delta b(0,1),$$

the effort cost below which separation is free, where $\Delta b(0,1)$ is evaluated with the low type's accuracy equal to $q_{AI}$. Both factors are continuous and strictly decreasing in $q_{AI}$, the second by the monotonicity of $\Delta b$ in the low type's accuracy established in the proof of Proposition 2, so $x$ is continuous and strictly decreasing, and $x(q_H) = 0$. By Theorem 1 the manager sets $\beta^{opt} = \underline{\beta}^{sep}$ when $k\,\underline{\beta}^{sep} \le \mu(q_H - q_{AI})$ and $\beta^{opt} = 0$ otherwise, where $\underline{\beta}^{sep} = \max\{0,\ (c-x)/(1-x)\}$.

*Step 1: where separation is not free.*

Separation requires monitoring exactly when $x(q_{AI}) < c$. Since $x$ is strictly decreasing, this holds on an interval of high AI qualities, $q_{AI} > q_-$, where $q_- = (c^*)^{-1}(c)$ if $c \le c^*(q_L)$ and $q_- = q_L$ otherwise. On that interval $\underline{\beta}^{sep} = (c-x)/(1-x)$ is continuous and strictly increasing in $q_{AI}$: the map $x \mapsto (c-x)/(1-x)$ has derivative $(c-1)/(1-x)^2 < 0$ and is therefore strictly decreasing, while $x$ itself is strictly decreasing in $q_{AI}$. As $q_{AI}$ approaches $q_H$ the required monitoring tends to $c$.

*Step 2: where monitoring is worth its cost.*

On $(q_-, q_H)$ the manager monitors precisely when the accuracy gain from separation covers its cost, that is when

$$g(q_{AI}) \;\coloneqq\; \mu(q_H - q_{AI}) \;-\; k\,\underline{\beta}^{sep}(q_{AI}) \;\ge\; 0.$$

The first term is strictly decreasing in $q_{AI}$ and, by Step 1, so is the negative of the second, so $g$ is continuous and strictly decreasing. Moreover, $g(q_H) = -kc < 0$. Hence the set on which the manager monitors is an interval $(q_-, q_+]$, with $q_+ < q_H$ the unique zero of $g$ when $g$ is positive just above $q_-$, and is empty otherwise.

*Step 3: the two cases.*

If $c \le c^*(q_L)$, then separation is free at $q_-$ and $\underline{\beta}^{sep} \to 0$ as $q_{AI} \downarrow q_-$, so $g(q_-^+) = \mu(q_H - q_-) > 0$ for every $k > 0$ and the monitoring interval is nonempty. If instead $c > c^*(q_L)$, then $q_- = q_L$ and $g(q_L^+) = \mu(q_H - q_L) - k\,\underline{\beta}^{sep}(q_L)$, which is nonnegative if and only if $k \le \bar{k}(c)$. When $k$ exceeds $\bar{k}(c)$ the manager never monitors.

Collecting the three steps: below $q_-$ separation is free and $\beta^{opt} = 0$, on $(q_-, q_+]$ the manager monitors and $\beta^{opt} = \underline{\beta}^{sep}$ is strictly increasing, and above $q_+$ monitoring no longer pays and

$\beta^{opt} = 0$ again. □

## Proof of Proposition 4

*Proof.* The planner chooses $(\alpha_H, \alpha_L) \in [0,1]^2$ to maximize

$$\mu\Big(\alpha_H q_{AI} + (1-\alpha_H)(q_{HB} - c)\Big) + (1-\mu)\Big(\alpha_L q_{AI} + (1-\alpha_L)(q_{LB} - c)\Big).$$

The objective is additively separable in $(\alpha_H, \alpha_L)$, so we can optimize with respect to each $\alpha_\theta$ separately for $\theta \in \{H, L\}$.

Fix a type $\theta \in \{H, L\}$. The contribution of type $\theta$ to the objective is

$$f_\theta(\alpha_\theta) \coloneqq \alpha_\theta q_{AI} + (1-\alpha_\theta)(q_{\theta B} - c) = (q_{\theta B} - c) + \alpha_\theta\big(q_{AI} - (q_{\theta B} - c)\big).$$

This is an affine function of $\alpha_\theta$ on $[0,1]$. Therefore the maximizer lies at one of the endpoints $\alpha_\theta \in \{0,1\}$.

The coefficient on $\alpha_\theta$ is

$$q_{AI} - (q_{\theta B} - c) = c - (q_{\theta B} - q_{AI}).$$

- If $q_{\theta B} - q_{AI} < c$, then $q_{AI} - (q_{\theta B} - c) > 0$ and $f_\theta$ is strictly increasing in $\alpha_\theta$, so the maximum is attained at $\alpha_\theta = 1$.
- If $q_{\theta B} - q_{AI} > c$, then $q_{AI} - (q_{\theta B} - c) < 0$ and $f_\theta$ is strictly decreasing in $\alpha_\theta$, so the maximum is attained at $\alpha_\theta = 0$.
- If $q_{\theta B} - q_{AI} = c$, then $f_\theta$ is constant in $\alpha_\theta$ on $[0,1]$, so any $\alpha_\theta \in [0,1]$ is optimal, and we select $\alpha_\theta^* = 1$ as in the statement.

Since the problem is separable across types, the efficient allocation is given by the pair $(\alpha_H^*, \alpha_L^*)$ obtained from this rule for $\theta = H$ and $\theta = L$, which yields

$$\alpha_\theta^* = \begin{cases} 1 & \text{if } q_{\theta B} - q_{AI} \le c, \\ 0 & \text{if } q_{\theta B} - q_{AI} > c. \end{cases}$$

The proposition follows. □

## Proof of Lemma 2

*Proof.* In such a profile monitoring reveals only that AI was used, which is uninformative, so type $\theta$'s payoff from any action with accuracy $q$ is $q\,b_r + (1-q)\,b_w$ less the effort cost if any. The combined technology costs the same as own technology alone, is strictly more accurate, and does not expose the worker to the pessimistic off-path belief. Comparing $B$ with $AI$ gives the displayed inequality. □

## Proof of Proposition 5

*Proof.* Recall that $\Delta b(x_H, x_L)$ is strictly decreasing in $x_L$ and strictly increasing in $x_H$.

The thresholds are

$$\tilde{c} \coloneqq (q_{LB} - q_{AI})\,\Delta b(q_{HB}, q_{LB}), \qquad c^{\dagger} \coloneqq (q_{LB} - q_{AI})\,\Delta b(q_{HB}, q_{AI}),$$
$$c^{\ddagger} \coloneqq (q_{HB} - q_{AI})\,\Delta b(q_{HB}, q_{AI}),$$

ordered as stated since $\Delta b$ is strictly decreasing in the low type's accuracy and $q_{HB} > q_{LB}$. By Lemma 2, in every profile considered only deviations between $B$ and $AI$ need to be checked, and type $\theta$ weakly prefers $B$ to $AI$ if and only if

$$c \le (q_{\theta B} - q_{AI})\,\Delta b(x_H, x_L), \tag{6}$$

where $(x_H, x_L)$ are the accuracies induced by the candidate profile. None of these conditions depends on $\beta$.

**(i) Pooling equilibrium $(B, B)$.** Both types use $B$, so $(x_H, x_L) = (q_{HB}, q_{LB})$. By (6), type $\theta$ has no profitable deviation to $AI$ iff $c \le (q_{\theta B} - q_{AI})\Delta b(q_{HB}, q_{LB})$. Since $q_{HB} > q_{LB}$, the low type's constraint binds, and the profile is an equilibrium iff $c \le (q_{LB} - q_{AI})\Delta b(q_{HB}, q_{LB}) \equiv \tilde{c}$.

**(ii) High type on $B$, low type mixes between $B$ and $AI$.** Let the low type use $AI$ with probability $\alpha_L \in (0,1)$, so $(x_H, x_L) = (q_{HB},\ (1-\alpha_L)q_{LB} + \alpha_L q_{AI})$, and define

$$\Phi_L(\alpha_L; c) \coloneqq (q_{LB} - q_{AI})\,\Delta b\big(q_{HB}, (1-\alpha_L)q_{LB} + \alpha_L q_{AI}\big) - c,$$

the low type's gain from $B$ over $AI$. As $\alpha_L$ rises the low type's accuracy falls, so correctness becomes more informative about the worker's type and $\Delta b$ rises: $\Phi_L$ is continuous and strictly increasing in $\alpha_L$, with $\Phi_L(0; c) = \tilde{c} - c$ and $\Phi_L(1; c) = c^{\dagger} - c$, where $c^{\dagger} \coloneqq (q_{LB} - q_{AI})\Delta b(q_{HB}, q_{AI}) > \tilde{c}$. Hence for $c \in [\tilde{c}, c^{\dagger}]$ there is a unique $\alpha_L(c) \in [0,1]$ with $\Phi_L(\alpha_L(c); c) = 0$, at which the low type

is indifferent and mixes. Since $\Phi_L$ is increasing in $\alpha_L$ and decreasing in $c$, $\alpha_L(c)$ is continuous and strictly increasing in $c$, so the probability $\hat{\alpha}_L^B(c) = 1 - \alpha_L(c)$ of using $B$ is strictly decreasing. For $c < \tilde{c}$ the low type strictly prefers $B$ at every $\alpha_L$ and for $c > c^\dagger$ he strictly prefers $AI$ at every $\alpha_L$, so no mixed-strategy equilibrium exists. There is no profitable deviation for the high type: by Assumption 1, $q_{HB} - q_{AI} > q_{LB} - q_{AI}$, so at the low type's indifference point (6) holds strictly for the high type.

**(iii) Separating equilibrium $(B, AI)$.** Here $(x_H, x_L) = (q_{HB}, q_{AI})$. The high type has no profitable deviation to $AI$ iff $c \leq (q_{HB} - q_{AI})\Delta b(q_{HB}, q_{AI}) \equiv c^\ddagger$, and the low type has none to $B$ iff $c \geq (q_{LB} - q_{AI})\Delta b(q_{HB}, q_{AI}) = c^\dagger$. Hence the profile is an equilibrium iff $c \in [c^\dagger, c^\ddagger]$, an interval since $q_{HB} > q_{LB}$.

**(iv) High type mixes between $B$ and $AI$, low type on $AI$.** Let the high type use $AI$ with probability $\alpha_H \in (0, 1)$, so $(x_H, x_L) = ((1 - \alpha_H)q_{HB} + \alpha_H q_{AI},\ q_{AI})$, and define

$$\Phi_H(\alpha_H; c) \coloneqq (q_{HB} - q_{AI})\,\Delta b\big((1 - \alpha_H)q_{HB} + \alpha_H q_{AI},\ q_{AI}\big) - c.$$

As $\alpha_H$ rises the high type's accuracy falls toward $q_{AI}$ and $\Delta b$ falls to 0: $\Phi_H$ is continuous and strictly decreasing in $\alpha_H$, with $\Phi_H(0; c) = c^\ddagger - c$ and $\Phi_H(1; c) = -c < 0$. Hence for every $c \in (0, c^\ddagger)$ there is a unique $\alpha_H(c) \in (0, 1)$ with $\Phi_H(\alpha_H(c); c) = 0$, continuous and strictly decreasing in $c$, since $\Phi_H$ is decreasing in both arguments, so the probability $\hat{\alpha}_H^B(c) = 1 - \alpha_H(c)$ of using $B$ is strictly increasing. The direction is the opposite of (ii) because the mixing type is the high type: using $B$ more often raises his accuracy and separates him further from the low type, which raises $\Delta b$, whereas the low type's use of $B$ moves his accuracy toward the high type's and lowers it. In either case the mixing probability adjusts so that the reputational gain just covers $c$. The low type has no profitable deviation to $B$: his gain from $B$ is $(q_{LB} - q_{AI})\Delta b - c < (q_{HB} - q_{AI})\Delta b - c = 0$ at the same beliefs. For $c \geq c^\ddagger$ no such equilibrium exists since $\Phi_H < 0$ for all $\alpha_H \in (0, 1)$.

**(v) No other equilibrium with $B$ in support.** Since $O$ is dominated by $B$, any equilibrium in which $B$ is played with positive probability has both types choosing from $\{B, AI\}$. Both types cannot mix: indifference of both would require $c = (q_{HB} - q_{AI})\Delta b = (q_{LB} - q_{AI})\Delta b$ at common beliefs, impossible since $q_{HB} > q_{LB}$ and $\Delta b > 0$. If the low type uses $B$ with positive probability, then by (6) the high type strictly prefers $B$ (Assumption 1) and uses it with probability one, which leaves the pooling profile of (i) or the mixed-strategy profile of (ii). If the low type never uses $B$, he uses $AI$, and the high type either uses $B$ with probability one, the separating profile of (iii), or mixes, the profile of (iv). This exhausts the cases and proves the proposition. □

## Proof of Theorem 2

*Proof.* We proceed in two steps corresponding to the two cost regions identified in Proposition 5.

*Step 1: Region $c \le c^{\ddagger}$.*

By Proposition 5, for any fixed $c \in (0, c^{\ddagger}]$ and any $\beta \in [0,1]$, there exists a PBE in which the high type uses the combined technology $B$ with probability one. More precisely:

- If $c \in (0, \tilde{c}]$, there is a pooling equilibrium $(B, B)$ in which both types always use $B$.
- If $c \in [\tilde{c}, c^{\dagger}]$, there is a mixed-strategy equilibrium in which the high type uses $B$ and the low type mixes between $B$ and $AI$.
- If $c \in [c^{\dagger}, c^{\ddagger}]$, there is a separating equilibrium in which the high type uses $B$ and the low type uses $AI$.

The thresholds $\tilde{c}, c^{\dagger}, c^{\ddagger}$ depend only on the primitives and not on $\beta$, and in each of the three regions the equilibrium action profile does not depend on $\beta$.

Because $q_{HB}$ is the highest attainable precision for the high type and $q_{LB}$ is strictly higher than both $q_L$ and $q_{AI}$, any profile in which the high type does *not* use $B$ with probability one, including the equilibria of Proposition 5(iv) and those involving only $\{O, AI\}$, is strictly dominated in accuracy by one in which he does use $B$. Therefore, for each $\beta \in [0,1]$ and each $c \le c^{\ddagger}$, the manager-optimal equilibrium is one in which the high type chooses $B$ (with the low type using $B$, $AI$, or a mixture depending on $c$).

Fix $c \le c^{\ddagger}$ and consider such a manager-optimal equilibrium $\sigma^{B}(\beta)$ at monitoring intensity $\beta$. Let $\alpha_L(\beta)$ denote the equilibrium probability with which the low type uses $AI$ under $\sigma^{B}(\beta)$. Then expected accuracy in this equilibrium is

$$A^{B}(c) \;=\; \mu q_{HB} + (1-\mu)\big[\alpha_L(\beta) q_{AI} + (1-\alpha_L(\beta)) q_{LB}\big].$$

By Proposition 5, the low type's mixing probability (when he mixes) is a function of $c$ only, not of $\beta$, and in the pure cases it is either 0 or 1, again independent of $\beta$. Hence for each $c \le c^{\ddagger}$, $A^{B}(c)$ does not depend on $\beta$.

The manager's payoff from this equilibrium is therefore

$$V^{B}(\beta; c) \;=\; A^{B}(c) - k\beta,$$

which is strictly decreasing in $\beta$ for $k > 0$. Since the manager-optimal equilibrium at each $\beta$ has the high type using $B$, we conclude that for $c \le c^{\ddagger}$ any $\beta > 0$ strictly reduces the manager's payoff

relative to $\beta = 0$:

$$V(0) \;=\; A^B(c) \;>\; A^B(c) - k\beta \;=\; V(\beta) \quad \text{for all } \beta > 0.$$

Thus the unique optimal monitoring intensity in this region is $\beta_C^{opt} = 0$. In the corresponding manager-optimal equilibria, the high type always uses $B$, and the low type uses $B$ or $AI$ (possibly mixing) as described in Proposition 5, proving part (i) of the theorem.

*Step 2: Region $c > c^{\ddagger}$.*

Again by Proposition 5, for any $c > c^{\ddagger}$ and any $\beta \in [0, 1]$ there is *no* PBE in which the combined technology $B$ is used by either type. In this region, all equilibria involve only the actions $O$ and $AI$. Monitoring then carries the same information as in the baseline: observing $a \in \{AI, B\}$ is observing $AI$ when $B$ is unused, and observing that AI was not used is observing $O$. The worker game therefore coincides with the baseline model without complementarities.

Therefore, for $c > c^{\ddagger}$ the manager's problem is identical to the one analyzed in Theorem 1: given monitoring cost $k$ and a worker game with action set $\{O, AI\}$, she chooses $\beta$ to maximize expected accuracy net of monitoring cost, under the assumption that the accuracy-maximizing PBE is selected at each $\beta$.

Applying Theorem 1 directly, we obtain that for $c > c^{\ddagger}$ the optimal monitoring intensity is

$$\beta_C^{opt} = \begin{cases} \underline{\beta}^{sep} & \text{if } k\,\underline{\beta}^{sep} \le \mu(q_H - q_{AI}), \\ 0 & \text{otherwise,} \end{cases}$$

where $\underline{\beta}^{sep}$ is the minimal monitoring probability that supports the separating equilibrium $(O, AI)$ in the baseline model. In the first case, the manager implements separation: the high type uses his own technology and the low type uses AI. In the second case, monitoring is too costly relative to the accuracy gain from separation, and the manager optimally sets $\beta_C^{opt} = 0$ and accepts pooling on AI.

This proves part (ii) and completes the proof of the theorem. □

## Proof of Proposition 6

The notation in this paragraph is used in this proof and the two that follow. Beliefs are as in Section 2, with the monitoring probability now depending on the outcome: a worker who is correct

is monitored with probability $\beta_T$ and one who is wrong with probability $\beta_F$. Write

$$v_T^a \coloneqq \beta_T\, b'_{a,r} + (1-\beta_T)\, b_r, \qquad v_F^a \coloneqq \beta_F\, b'_{a,w} + (1-\beta_F)\, b_w,$$

so that the expected reputation of Section 2 becomes $q(\theta,a)\, v_T^a + (1-q(\theta,a))\, v_F^a$ under truthful reporting and $(1-q(\theta,a))\, v_T^a + q(\theta,a)\, v_F^a$ under misreporting, less $c$ if $a = O$. At the separating profile $b_r = \mu q_H / P_r$ and $b_w = \mu(1-q_H)/P_w$, and

$$w_T^\theta \coloneqq q_\theta(1-b_r) + q_{AI}\, b_r, \qquad w_F^\theta \coloneqq (1-q_\theta)(1-b_w) + (1-q_{AI})\, b_w,$$

where, at the separating profile,

$$w_T^H = \frac{q_H\, q_{AI}}{P_r}, \qquad w_F^H = \frac{(1-q_H)(1-q_{AI})}{P_w},$$

and $w_T^\theta + w_F^\theta = 1 - (q_\theta - q_{AI})\Delta b$. The effective audit probability of type $\theta$ is $\bar\beta^\theta \coloneqq (\beta_T w_T^\theta + \beta_F w_F^\theta)/(w_T^\theta + w_F^\theta)$.

**Lemma 8** (Truthful reporting)**.** *A type-$\theta$ worker using $a$ prefers to report truthfully if and only if $v_T^a \geq v_F^a$.*

*Proof.* Truthful reporting beats misreporting iff $q\, v_T^a + (1-q)\, v_F^a \geq (1-q)\, v_T^a + q\, v_F^a$ with $q = q(\theta,a) > 1/2$, i.e. iff $(2q-1)(v_T^a - v_F^a) \geq 0$. □

In the separating profile, $b'_{O,r} = b'_{O,w} = 1$ and $b'_{AI,r} = b'_{AI,w} = 0$, so

$$v_T^O = b_r + \beta_T(1-b_r), \quad v_F^O = b_w + \beta_F(1-b_w), \quad v_T^{AI} = (1-\beta_T)\, b_r, \quad v_F^{AI} = (1-\beta_F)\, b_w,$$

and $v_T^O - v_F^O = V^O$, $v_T^{AI} - v_F^{AI} = V^{AI}$: by Lemma 8, condition (ii) of Proposition 6 is exactly truthful reporting by both types.

**Lemma 9** (Technology)**.** *Given (ii), the high type has no profitable deviation to $AI$ iff $\bar\beta^H \geq \underline{\beta}^{sep}$, and the low type has no profitable deviation to effort iff $\bar\beta^L \leq \bar\beta^{sep}$.*

*Proof.* Given (ii), each type's best deviation reports truthfully. The high type has no profitable deviation to $AI$ iff

$$q_H v_T^O + (1-q_H) v_F^O - c \;\geq\; q_{AI} v_T^{AI} + (1-q_{AI}) v_F^{AI},$$

that is, if and only if

$$(q_H - q_{AI})\Delta b + \beta_T w_T^H + \beta_F w_F^H \;\geq\; c,$$

and dividing by $w_T^H + w_F^H = 1 - (q_H - q_{AI})\Delta b$ gives $\bar{\beta}^H \geq [c - (q_H - q_{AI})\Delta b]/[1 - (q_H - q_{AI})\Delta b] = \underline{\beta}^{sep}$ when the latter is positive (and holds trivially otherwise). The low type has no profitable deviation to $O$ iff

$$q_{AI} v_T^{AI} + (1 - q_{AI}) v_F^{AI} \;\geq\; q_L v_T^O + (1 - q_L) v_F^O - c,$$

that is, if and only if

$$c + (q_{AI} - q_L)\Delta b \;\geq\; \beta_T w_T^L + \beta_F w_F^L,$$

and dividing by $w_T^L + w_F^L = 1 + (q_{AI} - q_L)\Delta b$ gives $\bar{\beta}^L \leq \bar{\beta}^{sep}$. □

Lemmas 8 and 9 together prove Proposition 6. □

## Proof of Corollary 2

*Proof.* Consider first an audit that inspects failures alone, so $\beta_T = 0$. The high type's truthful-reporting condition, $V^O \geq 0$, then reads $\beta_F(1 - b_w) \leq \Delta b$, that is

$$\beta_F \;\leq\; \frac{\Delta b}{1 - b_w} \;<\; 1,$$

the inequality because $\Delta b = b_r - b_w < 1 - b_w$. In particular the deterministic rule $(\beta_T, \beta_F) = (0, 1)$ violates it, so auditing every failure and no success never sustains separation. Symmetrically, an audit that inspects successes alone has $\beta_F = 0$, and the low type's condition $V^{AI} \geq 0$ reads $\beta_T b_r \leq \Delta b$, that is $\beta_T \leq \Delta b / b_r < 1$.

For the last claim, suppose some audit that sustains separation sets one of the two probabilities to one. If $\beta_F = 1$, then $V^O \geq 0$ becomes $\Delta b + \beta_T(1 - b_r) \geq 1 - b_w$, which simplifies to $\beta_T(1 - b_r) \geq 1 - b_r$ and hence to $\beta_T = 1$. If instead $\beta_T = 1$, then $V^{AI} \geq 0$ becomes $\Delta b + \beta_F b_w \geq b_r$, which simplifies to $\beta_F b_w \geq b_w$ and hence to $\beta_F = 1$. Either way the audit is $(1, 1)$, which monitors after every outcome and therefore gives each type an effective audit probability of one. But separation requires the low type's effective audit probability to be at most $\bar{\beta}^{sep}$, and $\bar{\beta}^{sep} < 1$ whenever $c < 1$ by Proposition 2, so $(1, 1)$ violates Proposition 6(i). No audit that sustains separation audits either outcome with certainty. □

## Proof of Theorem 3 and Corollary 3

*Proof.* Write $\mathcal{S}(\alpha)$ for the set of policies that sustain $\sigma_\alpha$ with truthful reporting and $C(\alpha)$ for the minimum over it of the expected monitoring cost $P_r(\alpha)\beta_T + P_w(\alpha)\beta_F$, where $P_r(\alpha) = \mu q_H + (1 - \mu)[\alpha q_{AI} + (1 - \alpha) q_L]$ and $P_w(\alpha) = 1 - P_r(\alpha)$. At $\alpha = 1$ we drop the argument. By Proposition 6,

$\mathcal{S}(1)$ is the intersection of four half-planes with the unit square, and its only lower boundary is the high type's effort constraint $\beta_T w_T^H + \beta_F w_F^H = c - c^*$, where

$$c^* = (q_H - q_{AI})\Delta b, \qquad c_S := c^* + \frac{w_T^H \Delta b}{b_r}, \qquad c_F := c^* + \frac{w_F^H \Delta b}{1 - b_w},$$

and $\hat{\mu}$ is the prior at which auditing successes and auditing failures are equally cost-effective, characterized in Step 2. We proceed in four steps.

*Step 1: the manager implements pooling on AI or some $\sigma_\alpha$ with $\alpha \in (0, 1]$.*

By Proposition 2, adapted to outcome-contingent policies exactly as Proposition 6 adapts the separating conditions, the equilibria under any policy are pooling on AI, the profiles $\sigma_\alpha$, and the profiles in which the high type mixes and the low type uses AI. The profile in which the high type uses AI and the low type exerts effort is never an equilibrium, as in the proof of Proposition 2.

Consider first a policy sustaining a high-type mixed-strategy equilibrium. Relative to separation at the same policy the high type's accuracy is lower, so $b_r$ is lower and $b_w$ higher, and the monitored posterior after AI is $b'_{AI} > 0$ rather than 0. Hence $B := (1 - \beta_T)b_r - (1 - \beta_F)b_w$ is larger under separation. The high type's indifference condition then implies her separation effort constraint,

$$\begin{aligned} c &= (q_H - q_{AI})B^{mix} + \beta_T q_H + \beta_F(1 - q_H) - \left[\beta_T q_{AI} + \beta_F(1 - q_{AI})\right] b'_{AI} \\ &\leq (q_H - q_{AI})B^{sep} + \beta_T q_H + \beta_F(1 - q_H), \end{aligned}$$

and $V^O = B + \beta_T - \beta_F$ and $V^{AI} = B$ are likewise larger under separation, so both truthful-reporting conditions hold there too. The policy therefore satisfies the high type's effort constraint and both conditions of Proposition 6(ii). On the effort constraint the low type's condition coincides with $V^O \geq 0$, since subtracting the two effort constraints and using $w_T^H - w_T^L = (q_H - q_L)(1 - b_r)$ and $w_F^H - w_F^L = -(q_H - q_L)(1 - b_w)$ turns it into $\beta_F(1 - b_w) - \beta_T(1 - b_r) \leq \Delta b$. Cost is increasing in both arguments, so its minimizer over that larger set lies on the effort constraint and hence in $\mathcal{S}(1)$: the policy costs at least $C(1)$, while delivering lower accuracy than separation. High-type mixing is therefore never implemented.

Pooling on effort is never implemented either. In that profile $b'_{O,r} = b_r$ and $b'_{O,w} = b_w$ evaluated at $(0, 0)$, and $b'_{AI} = 0$, so the low type's two deviations to AI are deterred if and only if

$$\begin{aligned} T &:= \beta_T q_{AI} b_r + \beta_F(1 - q_{AI})b_w - \left[c + (q_{AI} - q_L)\Delta b\right] \geq 0, \\ M &:= \beta_T(1 - q_{AI})b_r + \beta_F q_{AI} b_w - \left[c - (q_{AI} + q_L - 1)\Delta b\right] \geq 0, \end{aligned}$$

the high type's constraints being implied. Since $T - M = -(2q_{AI} - 1)V_0^{AI}$ with $V_0^{AI} := \Delta b - \beta_T b_r +$

$\beta_F b_w$, exactly one of the two binds at the cheapest such policy $P_0$, according to the sign of $V_0^{AI}$. If $V_0^{AI}(P_0) \geq 0$ then $T(P_0) = 0$, and as $\alpha \to 0$ the conditions defining $\mathcal{S}(\alpha)$ converge to those defining the pooling set, so $C(\alpha) \to C(0)$ while $W(\alpha) - W(0) = (1-\mu)(q_{AI} - q_L)\alpha > 0$: a little AI use by the low type is sustained at the same cost with higher accuracy. If $V_0^{AI}(P_0) < 0$ then $M(P_0) = 0$ and $P_0$ audits successes alone, with $\beta_T = [c - (q_{AI} + q_L - 1)\Delta b]/((1-q_{AI})b_r)$ and $c$ in the interval $\big(q_L\Delta b,\ q_L\Delta b + (1-q_{AI})b_w\big]$. Both $C(0)$ and the symmetric separating cost $\underline{\beta}^{sep}(c) \geq C(1)$ are affine in $c$ there, so it suffices to compare them at the endpoints: at the left endpoint $C(0) = P_r\Delta b/b_r = P_r^2\Delta b/(\mu q_H) \geq q_L\Delta b \geq \underline{\beta}^{sep}$, and at the right endpoint $\beta_T = 1$, $C(0) = P_r$ and $c \leq q_L b_r \leq P_r$, both using $P_r^2 \geq \mu q_H q_L$. Hence $C(1) \leq C(0)$ while $W(1) > W(0)$, and separation is cheaper and more accurate. This proves the last claim of Corollary 3.

*Step 2: the cheapest separating audit.*

Since cost is increasing in both arguments and the effort constraint is the only lower boundary of $\mathcal{S}(1)$, the minimizer lies on that constraint, at $(0,0)$ if $c \leq c^*$. Along it $d\beta_F/d\beta_T = -w_T^H/w_F^H$, so cost falls as $\beta_T$ rises if and only if $P_r/w_T^H < P_w/w_F^H$, that is, if and only if the cost of a unit of incentive is lower on the success side. Substituting $w_T^H = q_H q_{AI}/P_r$ and $w_F^H = (1-q_H)(1-q_{AI})/P_w$, the two unit costs are

$$\frac{P_r}{w_T^H} = \frac{P_r^2}{q_H\, q_{AI}}, \qquad \frac{P_w}{w_F^H} = \frac{P_w^2}{(1-q_H)(1-q_{AI})},$$

so the comparison is between $(P_r/P_w)^2$ and $\big[q_H/(1-q_H)\big]\big[q_{AI}/(1-q_{AI})\big]$. Taking square roots, successes are the cheaper side if and only if the odds of a correct outcome fall short of the geometric mean of the two types' odds of being right, $P_r/P_w < \sqrt{q_H q_{AI}/\big((1-q_H)(1-q_{AI})\big)}$. As $\mu$ rises from 0 to 1 the odds $P_r/P_w$ increase continuously from $q_{AI}/(1-q_{AI})$ to $q_H/(1-q_H)$, and the geometric mean of these two lies strictly between them, so $\hat{\mu}$ exists and is unique, and the audit targets successes if $\mu \leq \hat{\mu}$ and failures if $\mu > \hat{\mu}$.

At the minimizer the high type's effort constraint holds with equality, and the policies satisfying it with equality together with the remaining conditions form a segment. The minimum is at the end of that segment in the cost-reducing direction, and it remains to determine which condition binds there. By the identity of Step 1, on the effort constraint the low type's condition is $V^O \geq 0$. Moving toward successes relaxes it and $V^O$ alike, since $w_T^L/w_F^L < w_T^H/w_F^H$, so the segment ends at $\beta_F = 0$ or where $V^{AI} = 0$, and moving toward failures it ends at $\beta_T = 0$ or where $V^O = 0$. The success-only point $\beta_T = (c - c^*)/w_T^H$ satisfies $V^{AI} \geq 0$, that is $\beta_T b_r \leq \Delta b$, if and only if $c \leq c_S$, and the failure-only point $\beta_F = (c - c^*)/w_F^H$ satisfies $V^O \geq 0$ if and only if $c \leq c_F$. This gives the

audit of Corollary 3(i), and substituting into the cost gives

$$C(1) = \begin{cases} 0, & c \le c^*, \\ \dfrac{P_r^2\,(c-c^*)}{q_H\,q_{AI}}, & \mu \le \hat{\mu} \text{ and } c^* < c \le c_S, \\ \dfrac{P_w^2\,(c-c^*)}{(1-q_H)(1-q_{AI})}, & \mu > \hat{\mu} \text{ and } c^* < c \le c_F, \\ c-(1-c)\,\dfrac{(1-\mu)^2(q_H-q_{AI})^2}{q_H(1-q_H)}, & \mu \le \hat{\mu} \text{ and } c > c_S, \\ c-(1-c)\,\dfrac{\mu^2(q_H-q_{AI})^2}{q_{AI}(1-q_{AI})}, & \mu > \hat{\mu} \text{ and } c > c_F, \end{cases}$$

the last two obtained by solving the effort constraint together with $V^{AI} = 0$ and with $V^O = 0$ respectively. The two one-sided costs are mirror images, as are the two in which the truthful-reporting constraint binds: the success audit is priced by the accuracy of the two technologies and the failure audit by their inaccuracy, and the prior enters through $1-\mu$ in the first and through $\mu$ in the second.

For the remaining claims of Corollary 3(ii), no separating audit sets either probability to one by Corollary 2, so both components are strictly below one. On $V^{AI} = 0$ we have $\beta_T - \beta_F = \Delta b(1-\beta_F)/b_r > 0$ and on $V^O = 0$ we have $\beta_F - \beta_T = \Delta b(1-\beta_T)/(1-b_w) > 0$, so the targeted component strictly exceeds the other. Raising $c$ shifts the effort constraint outward along a line of positive slope, so both components are nondecreasing in $c$.

*Step 3: the audit is cheaper than symmetric monitoring.*

Let $c > c^*$ and $\mu \neq \hat{\mu}$. The symmetric policy $(\underline{\beta}^{sep}, \underline{\beta}^{sep})$ lies on the effort constraint and satisfies the other conditions strictly: the low type's because $\underline{\beta}^{sep} < \bar{\beta}^{sep}$ by Proposition 2, and both values of being right because at a symmetric $\beta < 1$ each equals $\Delta b(1-\beta) > 0$. By Step 2 a small move along the constraint in the cost-reducing direction remains in $\mathcal{S}(1)$ and strictly lowers cost, so $C(1) < \underline{\beta}^{sep}$. Under an outcome-contingent policy the manager can sustain separation at cost $C(1)$, so she monitors whenever $k\,C(1) \le \mu(q_H - q_{AI})$, while under symmetric monitoring she does so only if $k\,\underline{\beta}^{sep} \le \mu(q_H - q_{AI})$ by Theorem 1. Since $C(1) < \underline{\beta}^{sep}$, the set of parameters on which she monitors rather than accepting pooling on AI is strictly larger, which proves Corollary 3(iii).

*Step 4: separation against partial separation.*

For $\alpha \in (0,1)$ the low type must be exactly indifferent, so $\mathcal{S}(\alpha)$ lies on a line $L_\alpha$, which converges as $\alpha \to 1$ to the line $L_1$ on which the low type's separating condition holds with equality, the other conditions converging to those of $\mathcal{S}(1)$. Hence $\lim_{\alpha\to 1} C(\alpha)$ is the cost of the cheapest

separating audit on $L_1$, which is at least $C(1)$ and strictly more whenever the minimizer of Step 2 is not on $L_1$. By Step 2 that minimizer is the success-only point, the $V^{AI} = 0$ point, or the failure-only point in all cases except $\mu > \hat{\mu}$ with $c > c_F$, and at each of those three the low type's condition is slack. Partial separation therefore has strictly higher cost than separation, and is never chosen, unless $\mu > \hat{\mu}$ and $c > c_F$.

In that remaining case the minimizer is the vertex at which the effort constraint and $V^O = 0$ bind, and $V^O = 0$ is the low type's condition with equality, so the minimizer lies on $L_1$ and $C$ is continuous at 1. For $\alpha$ near one the cheapest audit sustaining $\sigma_\alpha$ solves the low type's indifference condition together with the high type's truthful-reporting condition at the beliefs of $\sigma_\alpha$,

$$\begin{aligned} c = (q_L - q_{AI})\big[(1-\beta_T)\, b_r(\alpha) - (1-\beta_F)\, b_w(\alpha)\big] \\ + \beta_T q_L\, b'_{O,r}(\alpha) + \beta_F(1-q_L)\, b'_{O,w}(\alpha), \\ \beta_T\, b'_{O,r}(\alpha) + (1-\beta_T)\, b_r(\alpha) = \beta_F\, b'_{O,w}(\alpha) + (1-\beta_F)\, b_w(\alpha), \end{aligned}$$

where $b_r(\alpha)$ and $b_w(\alpha)$ are the unmonitored posteriors when the low type's accuracy is $\alpha q_{AI} + (1-\alpha)q_L$, and $b'_{O,r}(\alpha) = \mu q_H/(\mu q_H + (1-\mu)(1-\alpha)q_L)$ and $b'_{O,w}(\alpha) = \mu(1-q_H)/(\mu(1-q_H) + (1-\mu)(1-\alpha)(1-q_L))$. The system is linear in $(\beta_T, \beta_F)$, and $c$ enters it in one place only, as a constant in the low type's indifference condition. By Cramer's rule each of $\beta_T(\alpha)$ and $\beta_F(\alpha)$ is therefore affine in $c$, with coefficients rational in $\alpha$ and independent of $c$, and agreeing at $\alpha = 1$ with the audit of Step 2. Since $P_r(\alpha)$ and $P_w(\alpha)$ do not depend on $c$, the cost $C(\alpha) = P_r(\alpha)\beta_T(\alpha) + P_w(\alpha)\beta_F(\alpha)$ is affine in $c$ for every $\alpha$, and so is its derivative. Write

$$C'(1^-) \;=\; A\,c + B,$$

where $A$ and $B$ depend on $(\mu, q_H, q_{AI}, q_L)$ alone. If $A > 0$, then $C'(1^-) > 0$ if and only if $c > \hat{c} \coloneqq -B/A$. If $A < 0$ the inequality is reversed, and if $A = 0$ the sign of $C'(1^-)$ is that of $B$, independently of $c$.

*Conclusion.*

A profile $\sigma_\alpha$ beats separation if and only if $k[C(1) - C(\alpha)] > W(1) - W(\alpha)$, so the manager prefers some partial-separation profile if and only if $k > \underline{k}$, where

$$\underline{k} \;\coloneqq\; \inf\left\{\tfrac{W(1)-W(\alpha)}{C(1)-C(\alpha)} \;:\; \alpha < 1 \text{ with } C(\alpha) < C(1)\right\},$$

with $\underline{k} \coloneqq +\infty$ when $C(\alpha) \geq C(1)$ for every $\alpha$. Step 4 gives a sufficient condition for $\underline{k}$ to be finite, together with an upper bound on it. If $\mu \leq \hat{\mu}$ or $c \leq c_F$, then $C(\alpha) > C(1)$ for every $\alpha$ near one,

and $\underline{k} = +\infty$ unless some profile far from separation is cheaper. If $\mu > \hat{\mu}$ and $c > c_F$, then $C$ is continuous at 1 and $C'(1^-) > 0$ exactly when $c > \hat{c}$. In that case profiles just below separation are cheaper, so $\underline{k}$ is finite and, taking $\alpha \uparrow 1$ in the definition,

$$\underline{k} \;\leq\; \frac{(1-\mu)(q_{AI} - q_L)}{C'(1^-)}.$$

It remains to compare the best of these profiles with pooling on AI, whose accuracy is $q_{AI}$ and which costs nothing. The objective $W(\alpha) - kC(\alpha)$ is continuous on $[0, 1)$, and at $\alpha = 1$ its limit from the left is no greater than its value, since $C(1) \leq \lim_{\alpha\uparrow 1} C(\alpha)$ by Step 4. It is therefore upper semicontinuous on the compact interval $[0, 1]$ and attains a maximum there. Write $\alpha^{opt}$ for a maximizer, which is interior or equal to one, since $\alpha = 0$ is pooling on effort and is dominated by Step 1. The manager monitors if and only if $k\,C(\alpha^{opt}) \leq W(\alpha^{opt}) - q_{AI}$, and by the previous paragraph $\alpha^{opt} = 1$ exactly when $k \leq \underline{k}$. Note that the comparison must be made at $\alpha^{opt}$ rather than at separation: since $C(\alpha) < C(1)$ is possible, $\sigma_\alpha$ can be worth its cost when separation is not. Together with Step 1 this gives the three cases of the theorem, and Steps 2 and 3 give Corollary 3. □