DELEGATION AND VERIFICATION UNDER AI

By

Lingxiao Huang, Wenyang Xiao and Nisheeth K. Vishnoi

March 2026

COWLES FOUNDATION DISCUSSION PAPER NO. 2500

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Delegation and Verification Under AI

Lingxiao Huang
Nanjing University

Wenyang Xiao
Nanjing University

Nisheeth K. Vishnoi
Yale University

Abstract

As AI systems enter institutional workflows, workers must decide whether to delegate task execution to AI and how much effort to invest in verifying AI outputs, while institutions evaluate workers using outcome-based standards that may misalign with workers' private costs. We model delegation and verification as the solution to a rational worker's optimization problem, and define worker quality by evaluating an institution-centered utility (distinct from the worker's objective) at the resulting optimal action. We formally characterize optimal worker workflows and show that AI induces *phase transitions*, where arbitrarily small differences in verification ability lead to sharply different behaviors. As a result, AI can amplify workers with strong verification reliability while degrading institutional worker quality for others who rationally over-delegate and reduce oversight, even when baseline task success improves and no behavioral biases are present. These results identify a structural mechanism by which AI reshapes institutional worker quality and amplifies quality disparities between workers with different verification reliability.

Contents

1 Introduction

Across high-stakes domains such as medicine, finance, and law, task completion traditionally followed a simple structure: a human worker executed the task and bore responsibility for correctness. As artificial intelligence (AI) systems enter institutional workflows, execution can be delegated while responsibility remains human. Institutions increasingly expect workers to use AI for efficiency, standardization, or cost reduction, transforming task completion into a process of delegation and oversight [20, 30]. Work is therefore organized as a *delegation pipeline*: a worker may perform the task manually, delegate execution to AI, or delegate while verifying the AI's output.

In experimental and real-world settings, workers tend to rely more heavily on AI advice on harder and less predictable tasks, even when AI accuracy is lower in those regimes [3, 24, 29]. At the same time, verification and oversight often weaken under cognitive load, complexity, or time pressure, leading to reduced scrutiny of AI outputs precisely where failures are most likely [9, 2]. Together, these findings point to a structural tension between delegation and verification in AI-assisted work. Notably, recent empirical evidence suggests that AI-assisted work can unevenly affect worker outcomes, with some workers benefiting from AI access while others experience degraded performance despite similar exposure to AI tools [4].

These observations raise a basic question: *How does AI reshape the value that institutions derive from human workers?* We refer to this value as *worker quality*: the institutional utility generated by a worker's output, accounting not only for task success but also for verification costs and accountability borne by the institution. While workers choose among manual work, verified delegation, and pure delegation to optimize their own effort–utility trade-offs, institutional evaluation is typically outcome-based and agnostic to internal workflows. This misalignment implies that individually rational delegation strategies need not maximize institutional utility, motivating the framework we develop in this paper.

Related work. Our setting relates to prior work on human–AI delegation, reliance, and monitoring under costly verification, which differ in how delegation and verification decisions are modeled and in whether worker behavior and institutional outcomes are treated endogenously. A growing body of work studies how tasks should be allocated between humans and AI systems based on task attributes like criticality and verifiability, utilizing frameworks such as selective prediction, learning-to-defer, and complementary expertise frameworks [19, 6, 22, 14, 26]. These approaches design routing or deferral policies to optimize system-level accuracy or utility, typically assuming fixed human accuracy and exogenous human effort, and therefore do not model how worker-chosen delegation interacts with verification effort or translates into institutional worker quality.

Empirical work documents patterns of over-reliance and oversight decay in AI-assisted settings as functions of task difficulty, uncertainty, and cognitive load [23, 29, 18]. While this evidence shows that verification weakens in practice, it typically explains these effects in psychological or organizational terms rather than modeling the underlying decision problem linking effort costs, task difficulty, and outcomes.

Economic models of monitoring and verification study how costly inspection or auditing should be deployed to deter misreporting or errors [27, 8, 21, 10]. In these models, verification is controlled by an institutional principal, rather than chosen by workers on a per-task basis, and human and AI performance are not jointly modeled as functions of task hardness.

Recent work also develops mathematical frameworks for human–AI integration, characterizing how human and AI capabilities combine to affect job success and productivity [5], as well as benchmark- and value-of-information–based approaches to joint decision-making [15, 12, 11]. These frameworks characterize optimal decisions given available signals, but do not model how delegation and verification choices are shaped by effort costs or how they induce institutional outcomes. Taken together, existing approaches address task allocation, reliance behavior, auditing, or evaluation in isolation, but do not provide a unified account of how worker-chosen delegation and verification under AI shape institutional worker quality.

Our contributions. We introduce a mathematical framework that characterizes how AI reshapes institutional worker quality through delegation and verification. Our main contributions are as follows:

- *Utility-driven model of delegation and verification.* We model worker–AI interaction via a delegation–verification action pair $(d, s) \in [0, 1]^2$, where d is the probability of delegating execution to AI and s is verification effort. Worker ability is parameterized by verification reliability $\alpha \in \mathbb{R}_{\geq 0}$ and execution efficiency $\beta \in \mathbb{R}_{\geq 0}$. Worker behavior is given by the solution to a utility maximization problem (Eq. (2)), while *worker quality* is defined by evaluating an institution-centered utility at the resulting optimal action (Eq. (4)). This framework makes explicit the structural misalignment between worker incentives and institutional objectives.

- *Workflow phase transitions under AI assistance.* We characterize the worker's optimal workflow $(d^\star, s^\star)$ as a function of ability parameters (α, β) (Theorem 3.1). We show that AI induces a *nonlinear and discontinuous* mapping from ability to behavior, with sharp transitions between manual work, verified delegation, and pure delegation. These transitions arise from structural properties of the delegation pipeline, persist under general concave detection and convex cost functions, and explain why accountability failures can emerge abruptly rather than gradually in AI-assisted workflows.

- *Reshaping institutional worker quality.* We characterize when AI assistance improves or degrades institutional worker quality (Theorem 3.5) and identify which workers are upgraded or downgraded relative to the pre-AI baseline (Theorem 3.6). We identify a *compliance gain* regime, in which workers with high verification reliability become institutionally qualified due to AI access, and a *compliance loss* regime, in which workers with low verification reliability rationally over-delegate and experience quality degradation. These effects arise endogenously from rational optimization, without invoking behavioral bias or miscalibration. More broadly, our results suggest a *verification amplification* phenomenon: AI access can disproportionately benefit workers with strong evaluative capabilities, while disadvantaging others through rational over-delegation.

- *Closed-form characterization in a tractable instantiation.* In a tractable instantiation with inverse-linear verification reliability and linear execution cost, we obtain closed-form expressions for optimal actions and resulting worker quality (Figure 2). This highlights the central role of verification reliability and reveals countervailing effects: higher verification ability or higher AI quality can induce over-delegation and reduce quality for some workers.

- *Worker and institutional interventions.* We formulate a constrained optimization problem for worker up-skilling that trades off improvements in verification reliability α and execution efficiency β under a qualification constraint (Eq. (17)). We further analyze institutional interventions, including improvements in AI capability and changes to worker incentives, and show that both can have non-monotonic effects on institutional utility (Section 6.2).

We further extend the model to heterogeneous task difficulty, miscalibrated beliefs about AI capability, and partial re-execution (Section 7), showing that the core qualitative phenomena persist, and provide empirical grounding on real-world data to illustrate applicability (Section 4).

Overall, our framework shows that AI functions not only as a productivity enhancer but as a structural filter that induces discontinuous changes in institutional worker quality. By isolating rational over-delegation as the core mechanism, we show that verification ability, rather than execution efficiency, becomes the primary determinant of institutional viability in AI-assisted workflows.

2 Model

We study a static, per-task model of AI-assisted work in which a rational worker interacts with a fixed AI system, choosing delegation and verification effort, while the institution evaluates the worker based on task outcomes.

Action space and workflow. A worker W chooses an action $(d, s) \in [0, 1]^2$, where d denotes the probability of delegating task execution to AI and s denotes the verification effort exerted upon delegation. With probability $1 - d$, the worker executes the task independently. When the task is delegated, the worker verifies the AI's output with effort s. If an error is detected, the worker redoes the task independently of the AI's output. The pre-AI workflow corresponds to $(d, s) = (0, 0)$, in which the worker always executes the task directly.

Worker utility. The worker chooses an action to trade off task success against completion cost, which we model via a utility function depending on the probability of task success and the cost incurred. Let $p : [0, 1]^2 \to [0, 1]$ denote the *task success function* and $\mathrm{cost} : [0, 1]^2 \to \mathbb{R}_{\geq 0}$ the *cost function*. For an action (d, s), $p(d, s)$ is the probability of success and $\mathrm{cost}(d, s)$ the incurred cost. Both functions are increasing in verification effort s and depend on worker and AI abilities, as specified in Equations (5) and (6). Let $b_W \geq 0$ denote the worker's benefit from success and $\ell_W \geq 0$ the loss from failure. The worker utility is

$$U_W(d, s) := b_W \, p(d, s) - \ell_W \big(1 - p(d, s)\big) - \mathrm{cost}(d, s). \tag{1}$$

The worker selects an optimal action

$$(d^\star, s^\star) := \arg \max_{(d,s) \in [0,1]^2} U_W(d, s). \tag{2}$$

Since AI enlarges the action space, the worker's utility weakly improves relative to the no-AI baseline: $U_W(d^\star, s^\star) \geq U_W(0, 0)$.

Institutional utility. The institution derives benefits from task success and incurs costs associated with worker effort, for example through compensation required to cover increased workload [25]. Let $b_I, \ell_I \geq 0$ denote the institutional benefit from task success and loss from task failure, respectively, and let $\xi \geq 0$ scale the worker's cost from the institutional perspective. The institutional utility under action (d, s) is

$$U_I(d, s) := b_I\, p(d, s) - \ell_I\big(1 - p(d, s)\big) - \xi \operatorname{cost}(d, s). \tag{3}$$

Institutional utility depends on the worker's action, which is not directly controlled by the institution.

Worker quality. We define *worker quality* as the institutional utility induced by the worker's optimal action,

$$Q(W) := U_I(d^\star, s^\star). \tag{4}$$

If institutional and worker utilities are aligned (e.g., $b_I = b_W$, $\ell_I = \ell_W$, and $\xi = 1$), then $Q(W) = U_W(d^\star, s^\star)$ and worker quality coincides with the worker's realized utility, which weakly improves under AI assistance. In general, however, $U_I \neq U_W$. Institutions employ multiple workers with heterogeneous objectives, so institutional utility cannot coincide with each worker's individual utility. This misalignment allows worker quality to decrease under AI assistance even when workers act rationally, consistent with empirical observations [3, 24].

Reliability and task success. Task success under action (d, s) arises through three mutually exclusive pathways: (i) direct worker execution, (ii) direct AI execution, and (iii) worker success after detecting and correcting an AI error. To compute the probability for each pathway, we define the reliability of the worker and the AI. Let $p_w, p_a \in [0, 1]$ denote the task success probabilities of the worker and the AI, respectively. The probabilities of direct worker and AI success are $(1 - d)p_w$ and dp_a. The quantities p_w and p_a may depend on task difficulty. For simplicity, we assume uniform task difficulty in the main text so that p_w and p_a are fixed, and relax this assumption in Section 7.1. Worker correction after AI errors is governed by an *error detection function* $\phi : [0, 1] \to [0, 1]$, which maps verification effort s to the probability of detecting an AI error. Conditional on detecting an error, the worker's subsequent task success is assumed to be independent of the AI's original output. Thus, the probability of success via correction is $d(1 - p_a)\phi(s)p_w$. Combining these cases, the task success probability induced by action $(d, s) \in [0, 1]^2$ is

$$p(d, s) := (1 - d)p_w + dp_a + d(1 - p_a)\phi(s)p_w. \tag{5}$$

When $p_a < p_w$, delegation without verification (i.e., $d > 0$ and $s = 0$) can reduce the overall task success probability. We impose the following regularity conditions on the error detection function $\phi(\cdot)$: (i) $\phi(0) = 0$, so no errors are detected without verification effort. (ii) ϕ is increasing, continuously differentiable, and strictly concave in s, implying diminishing returns to verification effort. These conditions capture the well-established speed–accuracy tradeoff and insights from random search theory in human cognition [28, 13, 16].

A commonly used functional form is $\phi(s) := 1 - e^{-\alpha s}$, where $\alpha \geq 0$ represents the worker's verification reliability. Larger values of α yield higher detection probabilities for any fixed $s > 0$, i.e., $\phi(s)$ is strictly increasing in α. An alternative is the inverse-linear form $\phi(s) := 1 - \frac{1}{1+\alpha s}$, which is a first-order approximation to $1 - e^{-\alpha s}$ and is analytically convenient.

Execution efficiency and cost. We model the cost incurred under action (d, s) by aggregating execution, verification, and correction costs. Let $C_w, C_a \geq 0$ denote the execution costs of the worker and the AI, respectively. The expected execution cost is $(1-d)C_w + dC_a$. In practice, C_a is typically much smaller than C_w, and we often assume $C_a = 0$ for simplicity.

We parameterize the worker's execution cost C_w by an efficiency parameter $\beta \geq 0$, where larger β corresponds to higher efficiency and lower execution cost. For example, C_w may scale linearly as $1 - \beta$ or inverse-linearly as $1/\beta$. Verification incurs an additional cost modeled by a *detection cost function* $C_v : [0, 1] \to \mathbb{R}_{\geq 0}$. We assume $C_v(0) = 0$ and that C_v is increasing, continuously differentiable, and strictly convex, implying increasing marginal verification costs [17]. Examples include a linear cost $C_v(s) \propto s$ or a convex cost such as $C_v(s) = s + \frac{s^2}{2}$. Since verification occurs only when the task is delegated to AI, the expected verification cost under action (d, s) is $d\,C_v(s)$. If an AI error is detected, the worker redoes the task at cost C_w, yielding an expected correction cost of $d(1-p_a)\phi(s)C_w$. The total cost under action (d, s) is

$$\mathrm{cost}(d, s) := (1-d)C_w + d\left(C_a + C_v(s) + (1-p_a)\phi(s)C_w\right). \tag{6}$$

Substituting (5) and (6) into (1) and (3) yields $U_W(d, s) = f_W(s)d + g_W$ and $U_I(d, s) = f_I(s)d + g_I$, both linear in the delegation level d. Here, g_W and g_I denote the worker's and institution's no-AI baseline utilities; in particular, g_I equals the pre-AI worker quality $Q_0(W)$. We restrict to the regime $g_W \geq 0$, ensuring nonnegative pre-AI worker utility. The terms $f_W(s)d$ and $f_I(s)d$ capture the incremental utilities from delegation with verification effort s. Expressions for $f_W(s)$, g_W, $f_I(s)$, and g_I are given in Section 5.1. Because worker and institutional objectives need not align, it is possible that $f_W(s) > 0$ while $f_I(s) < 0$, so delegation benefits the worker but harms the institution.

Figure 1 provides a flow chart for the workflow.

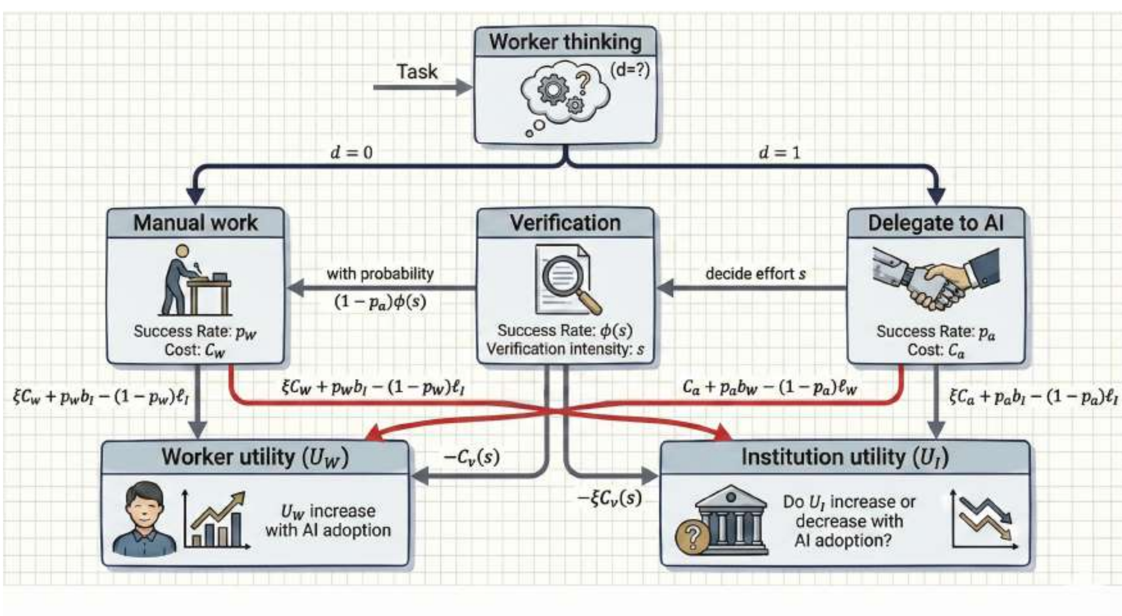
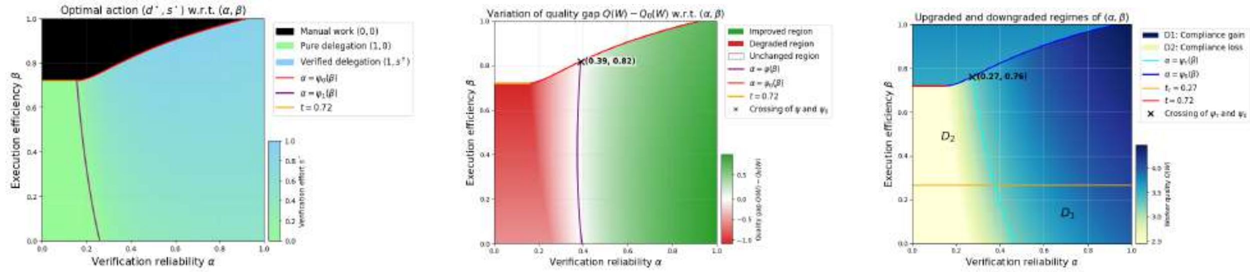


Figure 1: A flow chart for the workflow under AI assistance with worker action $(d, s) \in [0, 1]^2$.

Remark 2.1 (**Structural properties of delegation with verification**). The discontinuities identified in subsequent sections do not depend on specific functional forms, but arise from structural



(a) Worker action in Theorem 3.1 (b) Quality gap in Theorem 3.5 (c) Grading in Theorem 3.6

Figure 2: Plots illustrating the worker action and quality regimes characterized in Theorems 3.1, 3.5, and 3.6 as functions of abilities (α, β), under the default parameter setting $(b_W, \ell_W, b_I, \ell_I, \xi, \tau, p_a, C_a, p_w) = (8, 6, 14, 12, 0.3, 6.4, 0.65, 0, 0.75)$, and the functional choices $C_v(s) = s$, $C_w(\beta) = 5(1-\beta)$, and $\phi(s; \alpha) = 1 - \frac{1}{1+2\alpha s}$. Discussion on the choice of parameters and functions, and closed-form expressions for key quantities—including $(d^\star, s^\star)$ and the boundaries ψ_0, ψ_1, ψ, and ψ_τ—are provided in Section 5.10.

features of utility-driven delegation with costly verification. Suppose that (i) $\phi(s)$ is increasing and concave, (ii) $C_v(s)$ is increasing and convex, and (iii) utilities are affine in the delegation probability d for fixed s. Then $U_W(d, s) = f_W(s)d + g_W$, implying $d^\star(s) \in \{0, 1\}$ whenever $f_W(s) \neq 0$. As a result, small changes in worker or AI ability that flip the sign of $f_W(s)$ induce abrupt transitions between no delegation, verified delegation, and pure delegation. These phase transitions are therefore generic to delegation pipelines with costly verification.

In the remainder of the paper, we analyze how the optimal action $(d^\star, s^\star)$ and resulting quality $Q(W)$ vary with verification reliability α and execution efficiency β. By contrast, in the pre-AI baseline only execution efficiency β matters, as verification is unnecessary.

3 Theoretical results

This section presents our main theoretical results on how AI assistance reshapes worker behavior and institutional worker quality. Throughout, we fix the task profile (benefit parameters b_W, b_I, loss parameters ℓ_W, ℓ_I, and discount factor ξ), AI characteristics (task success probability p_a and execution cost C_a), and baseline worker characteristics (task success probability p_w and verification cost function $C_v(\cdot)$. Under this setup, the worker's optimal action $(d^\star, s^\star)$ and resulting worker quality $Q(W)$ depend only on verification reliability α and execution efficiency β, which parameterize the detection function $\phi(\cdot)$ and execution cost C_w.

We first show that the optimal worker action is nonlinear and discontinuous in (α, β) (Theorem 3.1). We then analyze how AI assistance alters worker quality relative to the pre-AI baseline: Theorem 3.5 characterizes when AI improves worker quality, while Theorem 3.6 identifies which workers are upgraded or downgraded, highlighting the central role of verification ability.

Characterization of worker action. To compute the verification decision, we define the *verification surplus* function

$$\Phi(s; \alpha, \beta) := (1 - p_a)\big((b_W + \ell_W)p_w - C_w\big)\,\phi(s) - C_v(s),$$

which captures the marginal benefit of detecting AI errors net of verification cost. This function corresponds to the s-dependent component of $f_W(s)$. Let

$$s^\dagger(\alpha,\beta) \in \arg\max_{s\in[0,1]} \Phi(s;\alpha,\beta)$$

denote the optimal verification effort, conditional on verification being employed. Since $f_W(s) = \Phi(s;\alpha,\beta) + \text{const}$, where the constant term does not depend on s, this is equivalent to $s^\dagger(\alpha,\beta) \in \arg\max_{s\in[0,1]} f_W(s)$. By the definition of Φ, this is equivalent to $s^\dagger(\alpha,\beta) \in \arg\max_{s\in[0,1]} f_W(s)$. We also define the *manual-to-delegation gain*

$$\Delta(\beta) := U_W(1,0) - U_W(0,0) = f_W(0), \tag{7}$$

which represents the worker's net utility change when switching from manual work to pure delegation (i.e., delegating execution to AI without verification). The last equality follows from the decomposition $U_W(d,s) = f_W(s)\,d + g_W$. Since no verification effort is exerted, we have $\phi(0) = 0$ and $C_v(0) = 0$ for any α, and hence Δ depends only on the execution efficiency β. Below we characterize the utility-driven worker action.

Theorem 3.1 (Worker optimal action). *Fix the task profile, AI characteristics, and baseline worker characteristics. Let $t \geq 0$ be a threshold such that $\Delta(t) = 0$. There exist continuous functions $\psi_0 : [t,\infty) \to \mathbb{R}_{\geq 0}$, monotonically increasing, and $\psi_1 : [0,t] \to \mathbb{R}_{\geq 0}$, monotonically decreasing, such that, for all β in their respective domains,*

$$f_W\big(s^\dagger(\psi_0(\beta),\beta)\big) = 0 \quad \text{and} \quad \partial_s\Phi\big(0;\psi_1(\beta),\beta\big) = 0.$$

Then the optimal action $(d^\star, s^\star)$ takes the following form:

- **(Manual work)** $(0,0)$ *if* $(\beta \geq t)$ *and* $(\alpha < \psi_0(\beta))$*;*
- **(Pure delegation)** $(1,0)$ *if* $(\beta < t)$ *and* $(\alpha < \psi_1(\beta))$*;*
- **(Verified delegation)** $(1, s^\dagger(\alpha,\beta))$ *otherwise.*

This result shows that the worker's optimal action lies in one of three regimes: manual work $(0,0)$, pure delegation $(1,0)$, and verified delegation $(1, s^\dagger(\alpha,\beta))$. Consequently, AI induces a nonlinear and discontinuous mapping from ability to action. In particular, the optimal delegation decision $d^\star$ is binary, leading to sharp regime shifts. This discreteness follows from the fact that worker utility $U_W(d,s)$ is affine in the delegation variable d, highlighting the structural source of abrupt action transitions under AI assistance (Figure 2a). For example, as β decreases past a threshold t, the worker may abruptly switch from manual work to pure delegation. Similarly, as verification reliability α crosses a threshold $\psi_0(\beta)$, the optimal verification effort $s^\star$ can jump discontinuously from 0 to $s^\dagger(\alpha,\beta)$. These phase transitions provide an explanation for empirical findings that accountability failures can emerge abruptly rather than gradually in AI-assisted workflows [29].

The threshold condition $\Delta(t) = 0$ characterizes indifference between manual work and pure delegation. The condition $f_W\big(s^\dagger(\psi_0(\beta),\beta)\big) = 0$ implies that, at $(\alpha,\beta) = (\psi_0(\beta),\beta)$, the worker is indifferent between manual work and verified delegation, so ψ_0 separates these regimes. Finally, the condition $\partial_s\Phi\big(0;\psi_1(\beta),\beta\big) = 0$ implies that verification yields zero marginal surplus at $s = 0$, so ψ_1 separates pure and verified delegation. When α is sufficiently large, the worker enters the verified-delegation regime, underscoring the role of verification ability in preventing over-delegation. Moreover, the monotonicity of the separatrices ψ_0 and ψ_1 implies that the optimal delegation decision $d^\star$ is non-increasing in execution efficiency β: higher execution efficiency makes manual work more attractive, reducing reliance on AI.

Worker quality under AI assistance. Given the worker's optimal action $(d^\star, s^\star)$, we can compute the resulting worker quality $Q(W)$. A central question is whether $Q(W)$ exceeds the no-AI baseline quality $Q_0(W)$. This question is practically relevant for workers, who seek to remain qualified and avoid job loss. It is equally important for institutions, for whom worker quality directly determines productivity, liability, and risk. From the institutional perspective, a key concern is whether workers are upgraded or downgraded under AI assistance, as this directly affects productivity. Let $\tau \geq 0$ denote a qualification threshold for worker quality. Formally, we ask which workers are *upgraded*—moving from $Q_0(W) < \tau$ to $Q(W) \geq \tau$—and which are *downgraded*—moving from $Q_0(W) \geq \tau$ to $Q(W) < \tau$.

To address these questions, we introduce technical assumptions. The first ensures that the institution places higher marginal value on task success than the worker does.

Assumption 3.2 (Institutional dominance). Assume that $(b_I + \ell_I) > \xi\,(b_W + \ell_W)$.

This assumption is typically reasonable in practice: institutions often have larger benefits from success and larger losses from failure than individual workers (i.e., $b_I > b_W$ and $\ell_I > \ell_W$), while the institution may bear only a discounted fraction of the worker's private cost (i.e., $\xi < 1$).

A second assumption ensures that, under the worker's optimal verification choice, higher verification reliability α leads to higher detection probability.

Assumption 3.3 (Monotonicity of detection). Assume that for any $\alpha, \beta \geq 0$, $\partial_\alpha\, \phi\big(s^\dagger(\alpha, \beta)\big) \geq 0$.

This assumption is intuitive: more reliable verifiers should detect more under optimal behavior. We provide concrete examples validating this condition in Section 5.2. We first study whether AI assistance improves worker quality relative to the no-AI baseline, i.e., whether $Q(W) - Q_0(W) > 0$ or not. Recall that $Q_0(W) = g_I$, which immediately implies the following claim.

Proposition 3.4 (Worker quality difference). $Q(W) - Q_0(W) = f_I(s^\star)\, d^\star$.

Consequently, the question reduces to determining the sign of $f_I(s^\star)\, d^\star$, which leads to the following theorem.

Theorem 3.5 (Worker quality improvement v.s. no AI). *Fix the task profile, AI characteristics, and baseline worker characteristics. Let t and $\psi_0(\cdot)$ be as given in Theorem 3.1. Under Assumptions 3.2 and 3.3, there exists a continuous function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that*

$$(f_I(s^\star) = 0) \wedge (d^\star = 1) \text{ for any pair } (\alpha, \beta) = (\psi(\beta), \beta).$$

Moreover,

- **(Improved)** $Q(W) > Q_0(W)$ *holds iff* $\alpha > \psi(\beta)$*;*
- **(Unchanged)** $Q(W) = Q_0(W)$ *holds iff* $\big((\beta \geq t) \wedge (\alpha < \psi_0(\beta))\big)$ *or* $\alpha = \psi(\beta)$*;*
- **(Degraded)** $Q(W) < Q_0(W)$ *holds otherwise.*

The result indicates that AI assistance reshapes worker quality in a heterogeneous manner. The regime in which quality remains unchanged largely coincides with the manual-work regime ($(d^\star, s^\star) = (0, 0)$). Beyond this region, worker quality improves under high verification reliability α, but can degrade when α is low, as over-delegation reduces task success. Moreover, higher execution efficiency

β does not necessarily translate into improved worker quality under AI assistance since delegation behavior can weaken the induced utility gain.

Figure 2b visualizes the three regimes governing how worker quality varies under AI assistance in Theorem 3.5. Notably, there is a specific region in which quality is degraded, located below the intersection point $(\alpha, \beta) = (0.39, 0.82)$ of the separatrices ψ and ψ_0. This region highlights a countervailing effect: increasing verification reliability can induce workers to switch from manual work to delegation even when reliability remains insufficient to make delegation quality-improving, thereby triggering over-delegation and reducing quality.

Moreover, we have $t = 1 - \frac{C_a + (b_W + \ell_W)(p_w - p_a)}{5}$ (see (16)), which is monotonically increasing in AI ability (i.e., higher p_a or lower C_a). Consequently, more capable AI enlarges the quality-degradation regime when α is close to 0. This reveals a second countervailing effect: improved AI capability can induce workers with low verification reliability to over-delegate, thereby reducing quality; see also Figure 3b.

Next, we characterize the conditions under which workers are upgraded or downgraded under AI assistance.

Theorem 3.6 (Worker upgraded v.s. downgraded). *Fix the task profile, AI characteristics, and baseline worker characteristics. Fix a grading threshold $\tau \geq 0$. Let $t_\tau \geq 0$ be the solution to $Q_0(W; \beta) = \tau$. Under Assumptions 3.2 and 3.3, there exists a monotonically decreasing function $\psi_\tau : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $Q(W) = \tau$ for the ability pair $(\alpha, \beta) = (\psi_\tau(\beta), \beta)$. Then*

- **(Compliance gain)** $(Q(W) > \tau) \wedge (Q_0(W) < \tau)$ *holds iff* $\big(\beta < \min\{t, t_\tau\}\big) \wedge \big(\alpha > \psi_\tau(\beta)\big)$ *or* $\big(t \leq \beta < t_\tau\big) \wedge \big(\alpha > \max\{\psi_0(\beta), \psi_\tau(\beta)\}\big)$.
- **(Compliance loss)** $(Q(W) < \tau) \wedge (Q_0(W) > \tau)$ *holds iff* $\big(t_\tau < \beta \leq t\big) \wedge \big(\alpha < \psi_\tau(\beta)\big)$ *or* $\big(\beta > \max\{t, t_\tau\}\big) \wedge \big(\psi_0(\beta) < \alpha < \psi_\tau(\beta)\big)$.

This result identifies both a *compliance gain* regime and a *compliance loss* regime, separated by a boundary that depends on verification reliability α. In particular, high α enables workers with low execution efficiency ($\beta < t_\tau$) to upgrade, while low α can cause workers with high execution efficiency ($\beta > t_\tau$) to be downgraded. The monotonicity of the separatrix ψ_τ indicates that workers with lower execution efficiency require higher verification reliability to remain qualified.

Together with Theorem 3.5, this yields a *verification amplification* phenomenon: AI access can disproportionately benefit workers with strong evaluative capabilities, while disadvantaging others through rational over-delegation. Notably, these effects arise endogenously from rational optimization, without invoking behavioral bias or miscalibration.

Figure 2c visualizes the upgraded and downgraded regimes characterized in Theorem 3.6. When $\beta < t$, worker quality $Q(W)$ increases more rapidly with α at low values of α, revealing diminishing returns to verification upskilling. This highlights that verification upskilling is most valuable for workers with low initial reliability.

Interventions and extensions. Section 6 discusses potential interventions to improve worker quality under AI assistance. From the worker side, we formulate a constrained optimization problem that trades off investments in verification reliability α and execution efficiency β under a qualification constraint (Eq. (17)). Figure 3a illustrates that improving α can be more cost-effective, reinforcing the central role of verification reliability.

From the institutional side, we analyze two levers: deploying more capable AI systems and reshaping worker incentives. We show that both can have non-monotonic effects on institutional utility, underscoring the complexity of designing effective organizational policies in the AI age.

Section 7 extends the model to heterogeneous task difficulty, miscalibrated beliefs about AI capability, and partial re-execution. We show that the core qualitative phenomena—phase transitions, compliance gain/loss regimes, and rational over-delegation—persist under these extensions, validating the robustness of our results.

3.1 Overview of the proofs

Table 1: Summary of monotonicity properties of key quantities with respect to the ability parameters α and β. Here $\uparrow$ denotes increasing, $\downarrow$ denotes decreasing, / denotes no clear monotonic pattern, and $\times$ denotes no dependence.

	$s^\dagger$	$d^\star$	$s^\star$	Q_0	$Q\mid_{d^\star=1}$	$Q\mid_{d^\star=1} - Q_0$	$\psi_0(\beta)$	$\psi_1(\beta)$	$\psi(\beta)$	$\psi_\tau(\beta)$	$f_W(s^\dagger)$	$\partial_s\Phi(0)$	$f_I(s^\dagger)$
α	/	$\uparrow$	/	$\times$	$\uparrow$	$\uparrow$	$\times$	$\times$	$\times$	$\times$	$\uparrow$	$\uparrow$	$\uparrow$
β	$\uparrow$	$\downarrow$	$\uparrow$	$\uparrow$	$\uparrow$	/	$\uparrow$	$\downarrow$	/	$\downarrow$	$\downarrow$	$\uparrow$	/

We summarize the main proof ideas; full proofs appear in Section 5. All arguments rely on monotonicity properties of a small number of quantities with respect to verification reliability α and execution efficiency β, despite the fact that optimal verification effort depends implicitly on both parameters. These relationships are summarized in Table 1.

To Theorem 3.1. We first characterize the structure of the worker's optimal workflow. Recall that $U_W(d,s) = f_W(s)\,d + g_W$ is affine in the delegation variable d, with all nontrivial dependence on verification captured by $f_W(s)$. Under strict concavity of ϕ and strict convexity of C_v, $f_W(s)$ is strictly concave, implying a unique optimal verification effort $s^\dagger(\alpha,\beta) \in \arg\max_{s\in[0,1]} f_W(s)$. The optimal delegation decision is therefore binary:

$$d^\star = \mathbb{I}\Big[f_W\big(s^\dagger\big) \geq 0\Big],$$

with the boundary between delegation and manual work given by $f_W\big(s^\dagger\big) = 0$.

To locate this boundary in ability space, we study how $f_W\big(s^\dagger(\alpha,\beta)\big)$ varies with α and β.

Lemma 3.7 (Effects of α and β on f_W). *$f_W\big(s^\dagger(\alpha,\beta)\big)$ is non-decreasing in α and non-increasing in β.*

When $s^\dagger = 0$, the condition $f_W(s^\dagger) = 0$ coincides with the manual–delegation indifference condition $\Delta(\beta) = 0$, yielding the threshold t separating manual work from pure delegation. When $s^\dagger > 0$, the same condition defines the separatrix ψ_0 between manual work and verified delegation. The monotonicity of ψ_0 follows directly from Lemma 3.7.

The remaining boundary ψ_1, separating pure from verified delegation, is determined by whether verification yields positive marginal surplus at $s = 0$.

Lemma 3.8 (Effects of α and β on Φ). *$\partial_s\Phi(0;\alpha,\beta)$ is non-decreasing in α and non-decreasing in β.*

Since ψ_1 is defined by $\partial_s \Phi(0; \alpha, \beta) = 0$, this lemma determines its geometry. Together, Lemmas 3.7 and 3.8 characterize the three workflow regimes and explain the origin of discontinuous transitions.

To Theorem 3.5. Theorem 3.5 compares worker quality with and without AI assistance. By Proposition 3.4, the quality gap reduces to the sign of $f_I(s^\star)d^\star$, so the key question is how this quantity varies with verification reliability.

Lemma 3.9 (**Effect of** α **on** $f_I(s^\star)d^\star$). *$d^\star$ is non-decreasing in α; and for ability pairs (α, β) with $d^\star = 1$, $f_I(s^\star)$ is non-decreasing in α.*

Higher α weakly increases the likelihood of delegation and, conditional on delegation, increases the institutional value of verification. Assumptions 3.2 and 3.3 ensure that detection improvements translate into institutional value at least as fast as into worker utility, yielding a separatrix ψ such that worker quality improves exactly when $\alpha > \psi(\beta)$.

To Theorem 3.6. Finally, we characterize which workers are upgraded or downgraded relative to a qualification threshold τ. This requires understanding how worker quality varies jointly with α and β in the delegation regime.

Lemma 3.10 (**Effects of** α, β **on** Q). *For ability pairs with $d^\star = 1$, worker quality $Q(W)$ is non-decreasing in α and non-decreasing in β.*

This lemma determines the geometry of the level sets $Q(W) = \tau$, yielding the separatrix ψ_τ that divides compliance gain from compliance loss. Combined with Theorem 3.1, this shows how AI assistance can upgrade workers with strong verification ability while downgrading others through rational over-delegation.

4 From data to model parameters

While our analysis is structural rather than empirical, the model is designed to interface with observational data from AI-assisted workflows. This section outlines how commonly available institutional data can be mapped to the model's parameters; full implementation details appear in Section A.

Observational inputs. We use the Collab-CXR dataset [1], which contains diagnostic data for 324 patient cases from 33 clinicians. For each clinician–case pair, the dataset records predictions over 14 pathologies in three settings: clinician alone, AI-alone, and clinician with access to the AI's output. The dataset also reports time spent per case in the clinician-alone and clinician-with-AI conditions. Expert diagnoses from 10 radiologists provide a reference for evaluating correctness, and a washout period of at least two weeks is enforced between clinician-alone and clinician-with-AI sessions.

Mapping observables to model quantities. From these data, we map each clinician to the model's core quantities: task success probabilities p_w and p_a, success probability under AI assistance $p(1, s^\dagger)$, execution costs C_w and C_a, and the assisted cost $\mathrm{cost}(1, s^\dagger)$. We illustrate this mapping for clinician `vn_exp_65341958`.

Deriving p_w, p_a, and $p(1,s^\dagger)$. Task success is defined as matching expert labels on all 14 pathologies. Under this criterion, we obtain success indicators for $n = 38$ cases in the clinician-alone, AI-alone, and clinician-with-AI conditions. Averaging yields $p_w = 0.447$, $p_a = 0.368$, and $p(1,s^\dagger) = 0.395$. In the model, the clinician-with-AI condition corresponds to $d = 1$, with the clinician exerting optimal verification effort $s^\dagger$.
Deriving C_w, C_a, and $\text{cost}(1,s^\dagger)$. We assume worker cost is proportional to task completion time. Since AI execution time is negligible, we set $C_a = 0$. Average completion times over the n cases yield $C_w = 172.5$ and $\text{cost}(1,s^\dagger) = 116.8$.

We next infer the clinician's ability parameters (α,β).
Deriving $\phi(s^\dagger)$. By (5), $\phi(s^\dagger) = \frac{p(1,s^\dagger)-p_a}{(1-p_a)p_w} = 0.093$.
Deriving $C_v(s^\dagger)$. By (6) with $C_a = 0$, $C_v(s^\dagger) = \text{cost}(1,s^\dagger) - (1-p_a)\phi(s^\dagger)C_w = 106.7$.
Choice of ability functions. We model cost through time expenditure, taking linear verification cost and normalizing verification effort and execution efficiency via $C_v(s) = T_v^{\max}s$ and $C_w(\beta) = T_w^{\max}(1-\beta)$, where $T_v^{\max} = 118.1$ and $T_w^{\max} = 262.3$ are the maximum observed verification and execution times. These normalizations ensure $s^\dagger, \beta \in [0,1]$. We use $\phi(s;\alpha) = 1 - e^{-\alpha s}$ [13].
Deriving β. From C_w, we obtain $\beta = 1 - \frac{C_w}{T_w^{\max}} = 0.342$.
Deriving α. From $C_v(s) = T_v^{\max}s$, we obtain $s^\dagger = 0.903$. Using $\phi(s^\dagger;\alpha)$ and the estimated $\phi(s^\dagger)$ yields $\alpha = -\frac{\ln(1-\phi(s^\dagger))}{s^\dagger} = 0.108$.

Identifying regime membership. Given (α,β), the model predicts the clinician's optimal action and resulting quality.
Identifying the action regime. By rationality, $s^\dagger$ satisfies $\partial_s\Phi(s^\dagger;\alpha,\beta) = 0$, yielding

$$b_W + \ell_W = \frac{T_v^{\max} + (1-p_a)\alpha e^{-\alpha s^\dagger}C_w}{(1-p_a)\alpha e^{-\alpha s^\dagger}p_w} = 4646.0.$$

This implies $f_W(s^\dagger) = -189.26 < 0$, so $(d^\star, s^\star) = (0,0)$. The clinician therefore lies in the manual-work regime of Theorem 3.1.
Identifying the quality regime. Setting $b_I = 2787.6$, $\ell_I = 1858.4$, and $\xi = 0.5$ (Assumption 3.2) yields $Q(W) = Q_0(W) = 133.77$. Thus, the clinician lies in the quality-unchanged regime of Theorem 3.5, and is neither upgraded nor downgraded in Theorem 3.6.

Interpreting interventions. Suppose the institution raises the qualification threshold to $\tau = 150$. As discussed in Section 3, the model can evaluate interventions such as upskilling worker abilities (increasing α or β) or deploying more capable AI (increasing p_a). To achieve $Q(W) \geq \tau$, the institution may consider: (i) increasing α from 0.108 to 0.335; (ii) increasing β from 0.342 to 0.479; or (iii) increasing p_a from 0.368 to 0.432. This numerical analysis can serve as a practical guide for institutions when designing intervention policies.

Scope and limitations. This calibration illustrates applicability rather than identification. The model does not estimate causal effects or capture dynamic learning or organizational adaptation. Parameters such as α and β may be interpreted at the role or population level, depending on the setting.

5 Omitted proofs and details from Section 3

This section provides proofs of the results in Section 3 and gives detailed derivations for Figure 2.

5.1 Explicit forms of utility functions

Recall that

$$U_W(d,s) = f_W(s)d + g_W, \text{ and } U_I(d,s) = f_I(s)d + g_I.$$

We provide the explicit form of the coefficients below:

$$\begin{aligned} f_W(s) &= (1-p_a)\left((b_W+\ell_W)p_w - C_w\right)\phi(s) - C_v(s) - (b_W+\ell_W)(p_w-p_a) + C_w - C_a \\ g_W &= (b_W+\ell_W)p_w - \ell_W - C_w \\ f_I(s) &= (1-p_a)\left((b_I+\ell_I)p_w - \xi C_w\right)\phi(s) - \xi C_v(s) - (b_I+\ell_I)(p_w-p_a) + \xi(C_w - C_a) \\ g_I &= (b_I+\ell_I)p_w - \ell_I - \xi C_w. \end{aligned}$$

Furthermore, note that $f_W(s) = \Phi(s;\alpha,\beta) + \Delta(\beta)$.

5.2 An illustrative example for Assumption 3.3: monotonicity of detection

Assumption 5.1 (Restatement of Assumption 3.3). Assume that for any $\alpha, \beta \geq 0$,

$$\partial_\alpha\, \phi\big(s^\dagger(\alpha,\beta)\big) \geq 0.$$

Consider the specific functional forms:

$$\phi(s;\alpha) := 1 - \frac{1}{1+10\alpha s}, \quad \text{and} \quad C_v(s) := 0.5s. \tag{8}$$

The constants 10 and 0.5 are chosen for convenience and can be replaced by other values, illustrating the generality of the assumption. Let $K_W(\beta) := (1-p_a)\left[(b_W+\ell_W)p_w - C_w\right]$. The worker chooses effort $s \in [0,1]$ to maximize the net benefit:

$$f_W(s) = K_W(\beta)\cdot\phi(s;\alpha) - C_v(s) + C_w - (b_W+\ell_W)(p_w-p_a) - C_a. \tag{9}$$

The first order condition with respect to s is:

$$\frac{\partial f_W(s)}{\partial s} = K_W(\beta)\cdot\frac{10\alpha}{(1+10\alpha s)^2} - 0.5 = 0. \tag{10}$$

Solving for the interior root s_0, we obtain:

$$s_0(\alpha) = \sqrt{\frac{K_W(\beta)}{5\alpha}} - \frac{1}{10\alpha}. \tag{11}$$

It is straightforward to verify that $s_0(\alpha)$ is unimodal in α (increasing then decreasing), with a maximum value of $(s_0)_{\max} = K_W(\beta)/2$. The constrained optimal effort is $s^\dagger = \min\left(1,\ \max(0, s_0)\right)$.

We now verify that the total derivative $\frac{d}{d\alpha}\phi(s^\dagger;\alpha)$ is strictly positive. Note that ϕ depends on α solely through the product $x := \alpha s$. Thus, $\phi(x) = 1 - (1+10x)^{-1}$, and $\frac{d\phi}{dx} = \frac{10}{(1+10x)^2} > 0$. It suffices to show that $\frac{d}{d\alpha}(\alpha s^\dagger) \geq 0$.

1. **Corner Solution ($s^\dagger = 0$):**
   In this regime, $s^\dagger = 0$, implying $\phi(0; \alpha) = 0$. The derivative is trivially non-negative.

2. **Corner Solution ($s^\dagger = 1$):**
   In this regime, $s^\dagger = 1$. The product becomes $x = \alpha$. The derivative is:

$$\frac{d}{d\alpha}\phi(1; \alpha) = \frac{d\phi}{dx} \cdot \frac{dx}{d\alpha} = \frac{10}{(1 + 10\alpha)^2} \cdot 1 > 0. \tag{12}$$

3. **Interior Solution ($s^\dagger \in (0, 1)$):**
   In this regime, $s^\dagger = s_0$. The product $x = \alpha s_0$ simplifies to:

$$x(\alpha) = \alpha \left(\sqrt{\frac{K_W(\beta)}{5\alpha}} - \frac{1}{10\alpha} \right) = \sqrt{\frac{\alpha K_W(\beta)}{5}} - 0.1. \tag{13}$$

   Differentiating x with respect to α:

$$\frac{dx}{d\alpha} = \frac{1}{2}\sqrt{\frac{K_W(\beta)}{5\alpha}} > 0. \tag{14}$$

   Consequently, by the Chain Rule:

$$\partial_\alpha \phi(s^\dagger) = \frac{10}{(1 + 10x)^2} \cdot \frac{1}{2}\sqrt{\frac{K_W(\beta)}{5\alpha}} > 0. \tag{15}$$

Combining all cases, we conclude that $\frac{d}{d\alpha}\phi(s^\dagger) \geq 0$ for all $\alpha > 0$.

5.3 Proof of Theorem 3.1: Worker optimal action

Theorem 5.2 (Restatement of Theorem 3.1). *Fix the task profile, AI characteristics, and baseline worker characteristics. Let $t \geq 0$ be a threshold such that $\Delta(t) = 0$. There exist continuous functions $\psi_0 : [t, \infty) \to \mathbb{R}_{\geq 0}$, monotonically increasing, and $\psi_1 : [0, t] \to \mathbb{R}_{\geq 0}$, monotonically decreasing, such that, for all β in their respective domains,*

$$f_W\big(s^\dagger(\psi_0(\beta), \beta)\big) = 0 \qquad \text{and} \qquad \partial_s \Phi\big(0; \psi_1(\beta), \beta\big) = 0.$$

Then the optimal action $(d^\star, s^\star)$ takes the following form:

- **(Manual work)** $(0, 0)$ *if* $(\beta \geq t)$ *and* $(\alpha < \psi_0(\beta))$*;*
- **(Pure delegation)** $(1, 0)$ *if* $(\beta < t)$ *and* $(\alpha < \psi_1(\beta))$*;*
- **(Verified delegation)** $(1, s^\dagger(\alpha, \beta))$ *otherwise.*

Proof. We proceed by analyzing the properties of the threshold functions and then deriving the optimal actions for each regime.

Monotonicity of $\psi_0(\beta)$. Recall that $\psi_0(\beta)$ is defined implicitly by the condition

$$f_W(s^\dagger(\psi_0(\beta), \beta)) = 0.$$

Consider two values β_1, β_2 with $\beta_1 > \beta_2$. By Lemma 3.7, the function $f_W(s^\dagger(\alpha, \beta))$ is non-decreasing in α and non-increasing in β. Comparing the values at $(\psi_0(\beta_1), \beta_1)$ and $(\psi_0(\beta_1), \beta_2)$, we have:

$$0 = f_W\big(s^\dagger(\psi_0(\beta_1), \beta_1)\big) \le f_W\big(s^\dagger(\psi_0(\beta_1), \beta_2)\big).$$

Since $f_W\big(s^\dagger(\psi_0(\beta_2), \beta_2)\big) = 0$, it follows that $f_W\big(s^\dagger(\psi_0(\beta_1), \beta_2)\big) \ge f_W\big(s^\dagger(\psi_0(\beta_2), \beta_2)\big)$. Because f_W is non-decreasing in α, this inequality implies $\psi_0(\beta_1) \ge \psi_0(\beta_2)$. Thus, $\psi_0(\beta)$ is monotonically increasing in β.

Monotonicity of $\psi_1(\beta)$. Recall that $\psi_1(\beta)$ is defined implicitly by $\partial_s\Phi(0; \psi_1(\beta), \beta) = 0$. Consider $\beta_1 > \beta_2$. By Lemma 3.8, the marginal benefit $\partial_s\Phi(0; \alpha, \beta)$ is non-decreasing in α and non-decreasing in β. Comparing the values at $(\psi_1(\beta_1), \beta_1)$ and $(\psi_1(\beta_1), \beta_2)$, we have:

$$0 = \partial_s\Phi(0; \psi_1(\beta_1), \beta_1) \ge \partial_s\Phi(0; \psi_1(\beta_1), \beta_2).$$

Since $\partial_s\Phi(0; \psi_1(\beta_2), \beta_2) = 0$, it follows that $\partial_s\Phi(0; \psi_1(\beta_1), \beta_2) \le \partial_s\Phi(0; \psi_1(\beta_2), \beta_2)$. Because $\partial_s\Phi$ is non-decreasing in α, this implies $\psi_1(\beta_1) \le \psi_1(\beta_2)$. Thus, $\psi_1(\beta)$ is monotonically decreasing in β.

Relationship between $f_W(0)$ and t. The term $f_W(0)$ represents the net benefit of pure delegation relative to manual work. Note that $f_W(0) = C_w(\beta) - C_a - (b_W + \ell_W)(p_w(\beta) - p_a)$. Since the manual cost C_w is decreasing in β, we have $\partial_\beta f_W(0) < 0$. The threshold t is defined such that $\Delta(t) = f_W(0)|_{\beta=t} = 0$. Therefore:

- If $\beta \ge t$, then $f_W(0) \le 0$.
- If $\beta < t$, then $f_W(0) > 0$.

Optimal action regimes. We now characterize the optimal strategy $(d^\star, s^\star)$ based on the sign of the optimal verification intensity $s^\dagger$ and the maximal value function $f_W(s^\dagger)$.

1. **Manual work** $(0, 0)$**.** The worker chooses manual work if and only if the utility of delegation (even with optimal verification) does not exceed manual utility, i.e., $f_W(s^\dagger(\alpha, \beta)) \le 0$. Since $f_W(s^\dagger)$ is maximized at $s^\dagger$ and bounded below by pure delegation $f_W(0)$, a necessary condition is $f_W(0) \le 0$, which implies $\beta \ge t$. Given $\beta \ge t$, the condition $f_W(s^\dagger(\alpha, \beta)) \le 0$ holds if and only if $\alpha \le \psi_0(\beta)$ since f_W is increasing in α and zero at ψ_0. Thus, the condition is: $\beta \ge t$ and $\alpha < \psi_0(\beta)$.

2. **Pure delegation** $(1, 0)$**.** The worker chooses pure delegation if and only if verification is not worthwhile ($s^\dagger = 0$) and delegation is better than manual work ($f_W(0) > 0$). First, $f_W(0) > 0$ implies $\beta < t$. Second, $s^\dagger = 0$ corresponds to the corner solution where the marginal benefit of verification at zero is non-positive: $\partial_s\Phi(0; \alpha, \beta) \le 0$. Since $\partial_s\Phi$ is increasing in α and zero at ψ_1, this requires $\alpha \le \psi_1(\beta)$. Thus, the condition is: $\beta < t$ and $\alpha < \psi_1(\beta)$.

3. **Verified delegation** $(1, s^\dagger)$**.** The worker chooses verified delegation in all other cases, specifically when $s^\dagger > 0$ and $f_W(s^\dagger) > 0$. This occurs if:

- $\beta < t$ and $\alpha > \psi_1(\beta)$ (Delegation is profitable, and verification is valuable).
- $\beta \geq t$ and $\alpha > \psi_0(\beta)$ (Delegation is only profitable because verification is sufficiently valuable to overcome the manual baseline). Note that when $\beta \geq t$, we must have $\psi_0(\beta) \geq \psi_1(\beta)$ for a solution to exist, ensuring consistency.

This concludes the proof of the characterization. □

5.4 Proof of Theorem 3.5: Worker quality improvement v.s. no AI

Theorem 5.3 (Restatement of Theorem 3.5). *Fix the task profile, AI characteristics, and baseline worker characteristics. Let t and $\psi_0(\cdot)$ be as given in Theorem 3.1. Under Assumptions 3.2 and 3.3, there exists a continuous function $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that*

$$(f_I(s^\star) = 0) \wedge (d^\star = 1) \text{ for any pair } (\alpha, \beta) = (\psi(\beta), \beta).$$

Moreover,

- **(Improved)** *$Q(W) > Q_0(W)$ holds iff $\alpha > \psi(\beta)$;*
- **(Unchanged)** *$Q(W) = Q_0(W)$ holds iff $\big((\beta \geq t) \wedge (\alpha < \psi_0(\beta))\big)$ or $\alpha = \psi(\beta)$;*
- **(Degraded)** *$Q(W) < Q_0(W)$ holds otherwise.*

Proof. We analyze the relationship between the equilibrium quality $Q(W)$ and the baseline manual quality $Q_0(W)$. Recall that the difference is given by $Q(W) - Q_0(W) = f_I(s^\star) \cdot d^\star$. Thus, the deviation depends on the delegation decision $d^\star$ and the sign of the institution's net benefit $f_I(s^\star)$.

Construction of $\psi(\beta)$. We define the threshold function $\psi(\beta)$ implicitly by the condition that the institution breaks even under delegation. Specifically, for an ability pair $(\alpha, \beta) = (\psi(\beta), \beta)$, we satisfy $d^\star = 1$ and $f_I(s^\dagger(\psi(\beta), \beta)) = 0$. By Lemma 3.9, $f_I(s^\dagger(\alpha, \beta))$ is non-decreasing in α, and the optimal delegation $d^\star$ is non-decreasing in α. Given that $d^\star = 1$ at the threshold, it follows that for any $\alpha > \psi(\beta)$, we have $d^\star = 1$ and $f_I(s^\dagger) > 0$. Conversely, if $\alpha \leq \psi(\beta)$ and delegation occurs, $f_I(s^\dagger) \leq 0$.

Improved ($Q(W) > Q_0(W)$). This regime holds if and only if the worker delegates ($d^\star = 1$) and the institutional benefit is strictly positive ($f_I(s^\dagger) > 0$). Based on the monotonicity established above, $f_I(s^\dagger) > 0$ holds if and only if $\alpha > \psi(\beta)$. Since $\alpha > \psi(\beta)$ implies $d^\star = 1$ (delegation is optimal for the worker), the condition simplifies to:

$$\alpha > \psi(\beta).$$

Unchanged ($Q(W) = Q_0(W)$). This regime holds if either the worker chooses manual work ($d^\star = 0$) or delegates with zero net institutional gain ($d^\star = 1 \wedge f_I(s^\dagger) = 0$).

- From Theorem 3.1, manual work ($d^\star = 0$) is chosen if and only if $\beta \geq t$ and $\alpha < \psi_0(\beta)$.
- Verified delegation with zero gain occurs exactly at the threshold $\alpha = \psi(\beta)$.

Combining these disjoint cases yields the condition:

$$\big((\beta \geq t) \wedge (\alpha < \psi_0(\beta))\big) \vee (\alpha = \psi(\beta)).$$

Degraded ($Q(W) < Q_0(W)$). This regime holds if and only if the worker delegates ($d^\star = 1$) despite the institutional benefit being negative ($f_I(s^\dagger) < 0$). First, $f_I(s^\dagger) < 0$ implies $\alpha < \psi(\beta)$. Second, for the worker to delegate in this region, they must satisfy the delegation conditions from Theorem 3.1: either ($\beta < t$) or ($\beta \geq t \wedge \alpha > \psi_0(\beta)$). Thus, degradation occurs if:

$$(\alpha < \psi(\beta)) \wedge \big[(\beta < t) \vee (\beta \geq t \wedge \alpha > \psi_0(\beta))\big].$$

□

5.5 Proof of Theorem 3.6: Worker upgraded v.s. downgraded

Theorem 5.4 (Restatement of Theorem 3.6). *Fix the task profile, AI characteristics, and baseline worker characteristics. Fix a grading threshold $\tau \geq 0$. Let $t_\tau \geq 0$ be the solution to $Q_0(W;\beta) = \tau$. Under Assumptions 3.2 and 3.3, there exists a monotonically decreasing function $\psi_\tau : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $Q(W) = \tau$ for the ability pair $(\alpha, \beta) = (\psi_\tau(\beta), \beta)$. Then*

- **(Compliance gain)** $(Q(W) > \tau) \wedge (Q_0(W) < \tau)$ *holds iff* $\big(\beta < \min\{t, t_\tau\}\big) \wedge \big(\alpha > \psi_\tau(\beta)\big)$ *or* $\big(t \leq \beta < t_\tau\big) \wedge \big(\alpha > \max\{\psi_0(\beta), \psi_\tau(\beta)\}\big)$.
- **(Compliance loss)** $(Q(W) < \tau) \wedge (Q_0(W) > \tau)$ *holds iff* $\big(t_\tau < \beta \leq t\big) \wedge \big(\alpha < \psi_\tau(\beta)\big)$ *or* $\big(\beta > \max\{t, t_\tau\}\big) \wedge \big(\psi_0(\beta) < \alpha < \psi_\tau(\beta)\big)$.

Proof. We characterize the regions where the worker's performance relative to the grading threshold τ changes due to AI availability.

Threshold properties. Let t_τ be the solution to $Q_0(W;\beta) = \tau$. Recall that $Q_0(W) = (b_I + \ell_I)p_W(\beta) - \ell_I - \xi C_w(\beta)$. Since the manual cost C_w is decreasing in β, $Q_0(W)$ is strictly increasing in β. Therefore:

$$Q_0(W) > \tau \iff \beta > t_\tau \quad \text{and} \quad Q_0(W) < \tau \iff \beta < t_\tau.$$

Next, let $\psi_\tau(\beta)$ be the ability level α such that the delegated quality equals τ, i.e., $Q(W)|_{d^\star=1} = \tau$. By Lemma 3.10, $Q(W)$ under delegation is non-decreasing in α. Thus:

$$Q(W)|_{d^\star=1} > \tau \iff \alpha > \psi_\tau(\beta).$$

Compliance Gain ($(Q(W) > \tau) \wedge (Q_0(W) < \tau)$). This scenario implies the worker fails manually but succeeds with AI. First, $Q_0(W) < \tau$ necessitates $\beta < t_\tau$. Second, $Q(W) > \tau$ implies the worker must delegate ($d^\star = 1$), because manual work would yield $Q_0 < \tau$. Thus, we require $d^\star = 1$ and $\alpha > \psi_\tau(\beta)$. We analyze delegation in two sub-regions of β:

- If $\beta < t$: Delegation is always preferred. Combining with $\beta < t_\tau$, we have $\beta < \min\{t, t_\tau\}$. The only remaining condition is $\alpha > \psi_\tau(\beta)$.
- If $\beta \geq t$: Delegation occurs only if $\alpha > \psi_0(\beta)$. Combining with $\beta < t_\tau$, we have $t \leq \beta < t_\tau$. The condition becomes $\alpha > \max\{\psi_0(\beta), \psi_\tau(\beta)\}$.

Combining these yields:

$$\big(\beta < \min\{t, t_\tau\} \wedge \alpha > \psi_\tau(\beta)\big) \vee \big(t \leq \beta < t_\tau \wedge \alpha > \max\{\psi_0(\beta), \psi_\tau(\beta)\}\big).$$

Compliance Loss ($(Q(W) < \tau) \wedge (Q_0(W) > \tau)$). This scenario implies the worker succeeds manually but fails with AI. First, $Q_0(W) > \tau$ necessitates $\beta > t_\tau$. Second, $Q(W) < \tau$ implies the worker must delegate ($d^\star = 1$), because manual work would yield $Q_0 > \tau$. Thus, we require $d^\star = 1$ and $\alpha < \psi_\tau(\beta)$. We analyze delegation in two sub-regions of β:

- If $\beta \le t$: Delegation is always preferred. Combining with $\beta > t_\tau$, we have $t_\tau < \beta \le t$. The condition is simply $\alpha < \psi_\tau(\beta)$.
- If $\beta > t$: Delegation occurs only if $\alpha > \psi_0(\beta)$. Combining with $\beta > t_\tau$, we have $\beta > \max\{t, t_\tau\}$. The condition becomes $\psi_0(\beta) < \alpha < \psi_\tau(\beta)$.

Combining these yields:

$$\big(t_\tau < \beta \le t \wedge \alpha < \psi_\tau(\beta)\big) \vee \big(\beta > \max\{t, t_\tau\} \wedge \psi_0(\beta) < \alpha < \psi_\tau(\beta)\big).$$

□

5.6 Proof of Lemma 3.7: effects of α, β on f_W

Lemma 5.5 (Restatement of Lemma 3.7). *$f_W\big(s^\dagger(\alpha, \beta)\big)$ is non-decreasing in α and non-increasing in β.*

Proof. Recall the definition of the worker's net benefit from delegation:

$$f_W(s) = K_w(\beta)\,\phi(s;\alpha) - C_v(s) - C_a + C_w(\beta) - (b_W + \ell_W)(p_w - p_a),$$

where

$$K_w(\beta) := (1 - p_a)\big((b_W + \ell_W)p_w - C_w(\beta)\big).$$

Let $F_W(\alpha, \beta) := \max_{s\in[0,1]} f_W(s)$. Note that $K_w(\beta) \ge 0$ holds under the viability assumption.

Monotonicity in α. Fix β. For any fixed $s \in [0, 1]$, differentiating with respect to α yields:

$$\frac{\partial}{\partial\alpha} f_W(s) = K_w(\beta)\,\frac{\partial}{\partial\alpha}\phi(s;\alpha).$$

Since $K_w(\beta) \ge 0$ and the detection probability $\phi(s;\alpha)$ is non-decreasing in verification ability α, we have $\frac{\partial}{\partial\alpha} f_W(s) \ge 0$. By the Envelope Theorem, since the objective function shifts upward pointwise, the maximum value $F_W(\alpha, \beta)$ must be non-decreasing in α. Specifically, for $\alpha_2 > \alpha_1$:

$$F_W(\alpha_2, \beta) = f_W(s^\dagger_{\alpha_2} \mid \alpha_2) \ge f_W(s^\dagger_{\alpha_1} \mid \alpha_2) \ge f_W(s^\dagger_{\alpha_1} \mid \alpha_1) = F_W(\alpha_1, \beta).$$

Monotonicity in β. Fix α and $s \in [0, 1]$. We differentiate $f_W(s)$ with respect to β. Note that β affects f_W primarily through the manual cost $C_w(\beta)$. Expanding the term involving C_w:

$$\text{Terms with } C_w(\beta) = -(1 - p_a)C_w(\beta)\phi(\alpha, s) + C_w(\beta) = C_w(\beta)\left[1 - (1 - p_a)\phi(s;\alpha)\right].$$

Differentiating with respect to β:

$$\frac{\partial}{\partial\beta} f_W(s) = \left[1 - (1 - p_a)\phi(s;\alpha)\right]\frac{\partial C_w(\beta)}{\partial\beta}.$$

First, observe the bracketed term. Since $\phi \in [0,1]$ and $p_a \in [0,1]$, we have $(1-p_a)\phi \leq 1$, which implies $[1-(1-p_a)\phi] \geq 0$. Second, by assumption, manual cost decreases with skill β, so $\frac{\partial C_w}{\partial \beta} < 0$. Combining these, we obtain:

$$\frac{\partial}{\partial \beta} f_W(s) \leq 0.$$

Thus, $f_W(s)$ is pointwise non-increasing in β. It follows that the maximum value $F_W(\alpha, \beta)$ is non-increasing in β. □

5.7 Proof of Lemma 3.8: effects of α, β on Φ

Lemma 5.6 (**Restatement of Lemma 3.8**)**.** $\partial_s \Phi(0; \alpha, \beta)$ *is non-decreasing in* α *and non-decreasing in* β*.*

Proof. Recall the definition of the net verification benefit function:

$$\Phi(s; \alpha, \beta) := K_w(\beta)\phi(s; \alpha) - C_v(s),$$

where the coefficient $K_w(\beta)$ is defined as:

$$K_w(\beta) := (1-p_a)\left((b_W + \ell_W)p_w - C_w(\beta)\right).$$

Recall that $K_w(\beta) \geq 0$ for all $\beta \geq 0$. Let $g(\alpha, \beta) := \partial_s \Phi(0; \alpha, \beta)$. Differentiating Φ with respect to s and evaluating at $s = 0$, we obtain:

$$g(\alpha, \beta) = K_w(\beta) \cdot \partial_s \phi(0; \alpha) - C_v'(0).$$

Monotonicity in α. Differentiating $g(\alpha, \beta)$ with respect to α:

$$\frac{\partial}{\partial \alpha} g(\alpha, \beta) = K_w(\beta) \cdot \frac{\partial}{\partial \alpha}\left(\partial_s \phi(0; \alpha)\right).$$

Since $K_w(\beta) \geq 0$, the sign of the derivative depends on the term $\frac{\partial}{\partial \alpha}\partial_s \phi(0; \alpha)$. By definition of the derivative at zero and the property that $\phi(0) = 0$:

$$\partial_s \phi(0; \alpha) = \lim_{t \to 0^+} \frac{\phi(t; \alpha) - \phi(0; \alpha)}{t} = \lim_{t \to 0^+} \frac{\phi(t; \alpha)}{t}.$$

We know $\phi(t; \alpha)$ is non-decreasing in α for any fixed $t > 0$. Therefore, for $\alpha_2 > \alpha_1$, we have $\phi(t; \alpha_2) \geq \phi(t; \alpha_1)$, which implies:

$$\frac{\phi(t; \alpha_2)}{t} \geq \frac{\phi(t; \alpha_1)}{t}.$$

Since the inequality holds for all $t > 0$, it is preserved in the limit as $t \to 0$. Thus, $\partial_s \phi(0; \alpha)$ is non-decreasing in α, implying $\frac{\partial}{\partial \alpha} g(\alpha, \beta) \geq 0$.

Monotonicity in β. Differentiating $g(\alpha, \beta)$ with respect to β:

$$\frac{\partial}{\partial \beta} g(\alpha, \beta) = \frac{\partial K_w(\beta)}{\partial \beta} \cdot \partial_s \phi(0; \alpha).$$

First, consider the derivative of $K_w(\beta)$. Assuming p_w is constant or absorbs into the cost scaling, the dependence on β comes from the manual cost $C_w(\beta)$:

$$\frac{\partial K_w(\beta)}{\partial \beta} = -(1 - p_a) \frac{\partial C_w(\beta)}{\partial \beta}.$$

Since the manual cost decreases with skill ($\frac{\partial C_w}{\partial \beta} < 0$) and $p_a \leq 1$, we have $\frac{\partial K_w}{\partial \beta} > 0$. Second, consider $\partial_s \phi(0)$. Since $\phi(t)$ is a probability increasing from $\phi(0) = 0$, we have $\phi(t) \geq 0$ for $t > 0$, implying:

$$\partial_s \phi(0) = \lim_{t \to 0} \frac{\phi(t)}{t} \geq 0.$$

Thus, $\frac{\partial}{\partial \beta} g(\alpha, \beta) \geq 0$. We conclude that $\partial_s \Phi(0; \alpha, \beta)$ is non-decreasing in β. □

5.8 Proof of Lemma 3.9: effect of α on $f_I(s^\star)$

Lemma 5.7 (**Restatement of Lemma 3.8**). *$d^\star$ is non-decreasing in α; and for ability pairs with $d^\star = 1$, $f_I(s^\star)$ is non-decreasing in α.*

Proof. We address the two assertions sequentially.

Monotonicity of $d^\star$. Recall the optimal policy from Theorem 3.1: the worker delegates ($d^\star = 1$) if and only if the net benefit of delegation exceeds the manual baseline, i.e., $f_W(s^\dagger) > \max(0, f_W(0))$. From Lemma 3.7, the value function $f_W(s^\dagger(\alpha, \beta))$ is non-decreasing in α. Consequently, if delegation is optimal for a given α_1 (implying $f_W(s^\dagger(\alpha_1)) >$ threshold), it must also be optimal for any $\alpha_2 > \alpha_1$, as $f_W(s^\dagger(\alpha_2)) \geq f_W(s^\dagger(\alpha_1))$. Thus, $d^\star$ is non-decreasing in α.

Monotonicity of $f_I(s^\star)$ under delegation. Assume $d^\star = 1$. We analyze the institutional utility $f_I(s^\star)$ by comparing it to the worker's utility $f_W(s^\star)$. Recall that the effective benefit coefficients for the institution and the worker are:

$$K_I(\beta) := (1 - p_a)\big((b_I + \ell_I)p_w - \xi C_w(\beta)\big), \quad K_W(\beta) := (1 - p_a)\big((b_W + \ell_W)p_w - C_w(\beta)\big).$$

Using these coefficients, the utility functions under delegation ($d = 1$) can be written as:

$$\begin{aligned} f_W(s) &= K_W(\beta)\phi(s; \alpha) - C_v(s) + \Delta_W, \\ f_I(s) &= K_I(\beta)\phi(s; \alpha) - \xi C_v(s) + \Delta_I, \end{aligned}$$

where Δ_W and Δ_I are terms independent of s and α. Observe that we can express $f_I(s)$ as a linear combination of $f_W(s)$ and the detection probability $\phi(s)$:

$$f_I(s) = \xi f_W(s) + (K_I(\beta) - \xi K_W(\beta))\phi(s; \alpha) + \text{Constant}.$$

We now differentiate the equilibrium value $f_I(s^\star)$ with respect to α.

1. **Case 1: Pure delegation ($s^\star = 0$).** In this case, $s^\star$ is constant locally. Since $\phi(0;\alpha) = 0$, $f_I(0)$ is constant with respect to α. Thus, $\frac{d}{d\alpha} f_I(s^\star) = 0$.

2. **Case 2: Verified delegation ($s^\star > 0$).** In this case, $s^\star = s^\dagger(\alpha, \beta)$. Differentiating the linked expression with respect to α:

$$\frac{d}{d\alpha} f_I(s^\dagger) = \xi \frac{d}{d\alpha} f_W(s^\dagger) + (K_I(\beta) - \xi K_W(\beta)) \frac{d}{d\alpha} \phi(s^\dagger; \alpha).$$

We analyze the signs of these terms:

- $\frac{d}{d\alpha} f_W(s^\dagger) \geq 0$ by Lemma 3.7.
- By Assumption 3.2, $(b_I + \ell_I) \geq \xi(b_W + \ell_W)$, which implies $K_I(\beta) \geq \xi K_W(\beta)$. Since $K_W > 0$, the coefficient $(K_I - \xi K_W)$ is non-negative.
- The term $\frac{d}{d\alpha} \phi(s^\dagger; \alpha)$ represents the total change in equilibrium detection probability. By Assumption 3.3, the equilibrium detection probability is non-decreasing in α.

Combining these non-negative terms, we conclude that $\frac{d}{d\alpha} f_I(s^\star) \geq 0$.

□

5.9 Proof of Lemma 3.10: effects of α, β on $Q(W)$

Lemma 5.8 (Restatement of Lemma 3.10). *For ability pairs with $d^\star = 1$, worker quality $Q(W)$ is non-decreasing in α and non-decreasing in β.*

Proof. Let $G(\alpha, \beta) := Q(W)|_{d^\star = 1}$. Using the effective benefit coefficients defined in the proof of Lemma 3.9, we write the institution's quality and the worker's objective function under delegation as:

$$G(\alpha, \beta) = K_I(\beta)\phi(s^\dagger; \alpha) - \xi C_v(s^\dagger) + \Gamma_I,$$
$$f_W(s^\dagger) = K_W(\beta)\phi(s^\dagger; \alpha) - C_v(s^\dagger) + \Gamma_W,$$

where Γ_I, Γ_W are constants with respect to s and α. Recall that $K_I(\beta) \geq \xi K_W(\beta) > 0$, and $\frac{dK_I}{d\beta} > 0$, $\frac{dK_W}{d\beta} > 0$ (due to decreasing manual costs).

Monotonicity of $s^\dagger$ in β. We first establish that the optimal effort $s^\dagger$ is non-decreasing in β. Consider the interior first order condition for the worker:

$$g(s, \beta) := \frac{\partial f_W}{\partial s} = K_W(\beta)\phi_s(s; \alpha) - C_v'(s) = 0.$$

By the implicit function theorem, the sensitivity of the interior solution s_0 is:

$$\frac{\partial s_0}{\partial \beta} = -\frac{\partial g / \partial \beta}{\partial g / \partial s}.$$

The denominator $\frac{\partial g}{\partial s} = \frac{\partial^2 f_W}{\partial s^2} < 0$ by the second-order condition for a maximum. The numerator is $\frac{\partial g}{\partial \beta} = K_W'(\beta)\partial_s \phi(s)$. Since $K_W'(\beta) > 0$ and $\partial_s \phi(s) > 0$, the numerator is positive. Therefore, $\frac{\partial s_0}{\partial \beta} > 0$. Since $s^\dagger = \min(1, \max(0, s_0))$, the monotonicity holds for the constrained solution as well.

Monotonicity of $Q(W)$ in β. Differentiating $G(\alpha, \beta)$ with respect to β:

$$\frac{\partial G}{\partial \beta} = K_I'(\beta)\phi(s^\dagger) + \frac{\partial s^\dagger}{\partial \beta}\left[K_I(\beta)\partial_s\phi(s^\dagger) - \xi C_v'(s^\dagger)\right].$$

From the worker's FOC, we know that for an interior solution, $C_v'(s^\dagger) = K_W(\beta)\partial_s\phi(s^\dagger)$. If $s^\dagger = 0$, the derivative term vanishes or is positive; if $s^\dagger = 1$, $\frac{\partial s^\dagger}{\partial \beta} = 0$. Focusing on the interior case, we substitute C_v':

$$\frac{\partial G}{\partial \beta} = K_I'(\beta)\phi(s^\dagger) + \frac{\partial s^\dagger}{\partial \beta}\phi_s(s^\dagger)\left[K_I(\beta) - \xi K_W(\beta)\right].$$

All terms are non-negative:

- $K_I'(\beta) \geq 0$ (decreasing manual cost increases net institutional benefit).
- $\frac{\partial s^\dagger}{\partial \beta} \geq 0$ (from above).
- $K_I(\beta) \geq \xi K_W(\beta)$ (Assumption 3.2).

Thus, $\frac{\partial G}{\partial \beta} \geq 0$.

Monotonicity of $Q(W)$ in α. Differentiating $G(\alpha, \beta)$ with respect to α:

$$\frac{\partial G}{\partial \alpha} = K_I(\beta)\left[\partial_\alpha\phi(s^\dagger) + \partial_s\phi(s^\dagger)\frac{\partial s^\dagger}{\partial \alpha}\right] - \xi C_v'(s^\dagger)\frac{\partial s^\dagger}{\partial \alpha}.$$

Rearranging terms:

$$\frac{\partial G}{\partial \alpha} = K_I(\beta)\partial_\alpha\phi(s^\dagger) + \frac{\partial s^\dagger}{\partial \alpha}\left[K_I(\beta)\partial_s\phi(s^\dagger) - \xi C_v'(s^\dagger)\right].$$

Again, using the worker's first-order condition $C_v' \leq K_W\partial_s\phi$, which implies that $K_I\partial_s\phi - \xi C_v' \geq K_I\phi_s - \xi K_W\phi_s$:

$$\frac{\partial G}{\partial \alpha} \geq K_I(\beta)\partial_\alpha\phi(s^\dagger) + \frac{\partial s^\dagger}{\partial \alpha}\partial_s\phi(s^\dagger)\left[K_I(\beta) - \xi K_W(\beta)\right].$$

Since $\partial_\alpha\phi \geq 0$, $\frac{\partial s^\dagger}{\partial \alpha} \geq 0$ (Lemma 3.9), and $K_I \geq \xi K_W$, the total derivative is non-negative. □

5.10 Details of the illustrative example in Figure 2

We first illustrate the choice of parameters and functions in Figure 2. Consider a worker with a linear verification cost $C_v(s) = s$, an execution cost $C_w(\beta) = 5(1-\beta)$ for efficiency $\beta \in [0, 1]$ (where the factor 5 reflects that execution is substantially more costly than verification), and an inverse-linear error-detection function $\phi(s; \alpha) = 1 - \frac{1}{1+2\alpha s}$. To match the scale of β, we restrict to $\alpha \in [0, 1]$, where the constant 50 serves as a scaling parameter for this normalization. Below, we select a representative parameter setting for analysis.
Task profile. We set $b_W = 8$, $\ell_W = 6$, $b_I = 14$, $\ell_I = 12$, $\xi = 0.3$, and $\tau = 6.4$, which satisfies Assumption 3.2.
AI characteristics. We set $p_a = 0.65$ (moderate reliability) and $C_a = 0$ (negligible execution cost).
Worker characteristics. We set $p_w = 0.75$, so that the worker outperforms the AI in task success probability.

Characterization of key quantities. We next compute the key quantities that govern the regime characterization in our theoretical results.
To Theorem 3.1. To compute the threshold t, we first obtain

$$\Delta(\beta) = 5(1-\beta) - C_a - (b_W + \ell_W)(p_w - p_a).$$

Under the above parameter choices, this yields

$$t = 1 - \frac{C_a + (b_W + \ell_W)(p_w - p_a)}{5} = 0.72. \tag{16}$$

We also derive closed-form expressions for the two regime boundaries (on their respective domains):

$$\psi_0(\beta) = \frac{20}{\left(\sqrt{77+70\beta} - \sqrt{200\beta - 144}\right)^2}, \qquad \psi_1(\beta) = \frac{20}{77+70\beta}.$$

Finally, within the verified-delegation regime, the optimal action takes the form $(d^\star, s^\star) = (1, s^\dagger(\alpha,\beta))$ with

$$s^\dagger(\alpha,\beta) = \sqrt{\frac{1.925 + 1.75\beta}{\alpha}} - \frac{1}{\alpha}.$$

To Theorem 3.5. The key is to characterize the separatrix ψ. To this end, we first define a function $\psi' : [0,1] \to [0,1]$ (monotonically decreasing in β) such that

$$f_I\Big(s^\dagger(\alpha,\beta)\Big) = 0 \text{ for any ability pair } (\alpha,\beta) = (\psi'(\beta), \beta).$$

Extending $\psi_0(\beta) = 0$ for $\beta < t$, Theorem 3.5 implies that $\psi(\beta) = \max\{\psi_0(\beta), \psi'(\beta)\}$. Finally, the closed-form condition of $f_I\left(s^\dagger(\alpha,\beta)\right) = 0$ is:

$$5.2 - 0.975\beta = \frac{6.3 + 0.525\beta}{10\sqrt{\alpha(0.077 + 0.07\beta)} - 1} + \frac{15\sqrt{\alpha(0.077+0.07\beta)} - 3}{10\alpha}$$

To Theorem 3.6. To compute the threshold t_τ, we first obtain

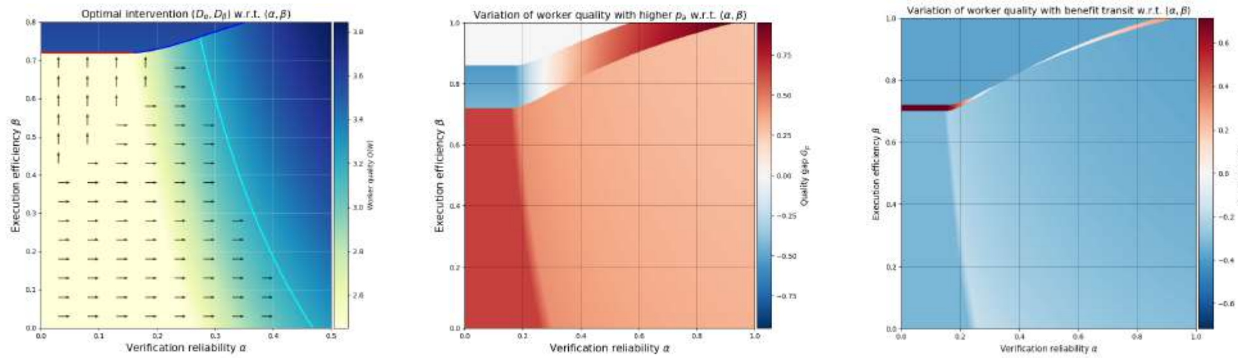
$$Q_0(W) = g_I = (b_I + \ell_I)p_w - \ell_I - 5\xi(1-\beta).$$

Solving $Q_0(W) = \tau$ yields the threshold $t_\tau = \frac{4}{15}$. under our parameter choices. We also derive a closed-form condition for $Q(W) = \tau$, which defines the separatrix ψ_τ:

$$\frac{(252 + 21\beta)(2 - \sqrt{7\alpha(1.1+\beta)})}{40(\sqrt{7\alpha(1.1+\beta)} - 1)} - \frac{3\sqrt{1.75(1.1+\beta)}}{10\sqrt{\alpha}} + \frac{0.3}{\alpha} + 4.9 = \tau.$$

6 Interventions to improve worker quality

This section studies interventions for workers who are unqualified under AI assistance (i.e., $Q(W) < \tau$), with the goal of improving their quality to at least τ.



(a) Optimal intervention (b) Heatmap of G_p with $D_p = 0.05$ (c) Heatmap of G_b with $D_b = 1$

Figure 3: Plots illustrating the effects of interventions from the worker side and the institutional side as functions of abilities (α, β), under the default parameter setting $(b_W, \ell_W, b_I, \ell_I, \xi, \tau, p_a, C_a, p_w) = (8, 6, 14, 12, 0.3, 6.4, 0.65, 0, 0.75)$, and the functional choices $C_v(s) = s$, $C_w(\beta) = 5(1-\beta)$, $\phi(s;\alpha) = 1 - \frac{1}{1+2\alpha s}$, and $h_1(x) = h_2(x) = x$. In Figure 3a, we restrict the domain to $[0, 0.5] \times [0, 0.8]$ for clarity. In Figures 3b and 3c, we analyze how worker quality varies under a more capable AI and under a larger marginal gain, respectively.

6.1 Worker-side interventions

It follows from Lemma 3.10 that increasing either verification reliability α or execution efficiency β can improve worker quality $Q(W)$, and thus both can serve as worker-side interventions. To determine which intervention is more cost-effective, we formulate a constrained optimization problem.

We consider two interventions: increasing α by $D_\alpha \geq 0$ and increasing β by $D_\beta \geq 0$. Let $Q(D_\alpha, D_\beta)$ denote the resulting worker quality under abilities $(\alpha + D_\alpha, \beta + D_\beta)$. The goal is to ensure $Q(D_\alpha, D_\beta) \geq \tau$ while minimizing intervention cost.

Let $h_1, h_2 : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be monotone functions that model the costs of increasing α and β, respectively. A simple specification is linear, $h_i(x) = c_i x$ for $i \in \{1, 2\}$, which is justified as a first-order approximation when x is small. More generally, one may take $h_i(x) = c_i x^\rho$ for $\rho > 1$, capturing increasing marginal costs of continued upskilling.

Optimization problem. We formalize the intervention design as:

$$\min_{D_\alpha, D_\beta \geq 0} h_1(D_\alpha) + h_2(D_\beta) \quad \text{s.t.} \quad Q(D_\alpha, D_\beta) \geq \tau. \tag{17}$$

Analysis. We use the same setting as in Figure 2 and set $h_1(x) = h_2(x) = x$. Figure 3a plots the optimal intervention ways for workers with various (α, β). Observe that improving α is usually more cost-efficient, validating the central role of verification reliability in the AI age.

6.2 Institutional-side interventions

From the institutional-side, we propose two potential interventions: deploying more capable AI systems and reshaping worker incentives. Below we illustrate these two interventions.

Deploying more capable AI systems. This intervention can be implemented by training a task-specific agent and corresponds to increasing the AI success probability p_a. Under this intervention, both the worker's optimal action and the resulting quality may change. Let $D_p \in [0, 1 - p_a]$ denote the increase in p_a, and let $Q(D_p)$ denote the worker quality under AI capability $p_a + D_p$. We define the *quality variation value* as

$$G_p(D_p) := Q(D_p) - Q(0).$$

Figure 3b plots a heatmap of G_p with $D_p = 0.05$. We observe that G_p can be positive in some regions of (α, β) but negative in others, indicating that improving AI capability may either increase or decrease worker quality depending on worker abilities. Thus, deploying more capable AI systems can have a non-monotonic effect on worker quality. Specifically, workers with high execution efficiency β but low verification reliability α may switch from manual work to pure delegation, which can reduce quality.

Reshaping worker incentives. Another strategy is to increase the worker's success benefit b_W, which strengthens incentives to verify AI outputs. Such an increase can be implemented by transferring institutional benefit to the worker, thereby reducing the institution's marginal gain and shifting the institutional utility. Let $D_b \geq 0$ denote the increase in b_W, resulting in benefits $b_W + D_b$ for the worker and $b_I - D_b$ for the institution. Let $Q(D_b)$ denote the worker quality under the reshaped incentive $b_W + D_b$, and define the *quality variation value* as

$$G_b(D_b) := Q(D_b) - Q(0).$$

Figure 3c plots a heatmap of G_p with $D_b = 1$. We observe that G_p can be positive in some regions of (α, β) but negative in others, indicating that incentive reshaping can either improve or degrade quality depending on worker abilities. Thus, reshaping worker incentives can also have a non-monotonic effect on worker quality. Specifically, workers with high execution efficiency β but low verification reliability α may switch from pure delegation to manual work when incentives increase, which can improve quality. We also note that, compared to increasing p_a, the quality-improvement region induced by benefit transfer is much narrower, suggesting that this policy is less effective.

Overall, these results underscore the complexity of designing effective organizational policies in the AI age.

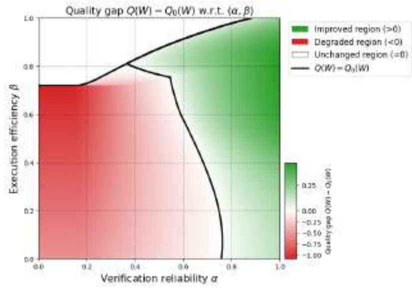
7 Model extension

In this section, we study three extensions to validate the robustness of our analysis.

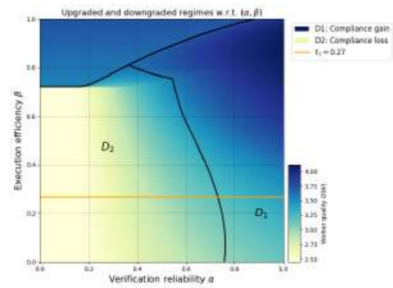
7.1 Heterogeneous task difficulty

Our model implicitly assumes that tasks have uniform difficulty, leading to constant success probabilities and costs. In practice, workers may face tasks of varying difficulty—for example, clinical cases can differ substantially in diagnostic hardness. To study the effect of heterogeneous task difficulty, we extend the model as follows.

Let $h \in [0, 1]$ denote task difficulty, where $h = 0$ is easiest and $h = 1$ is hardest. We model success probabilities as $p_w(h) = 1 - 0.5h$ and $p_a(h) = 1 - 0.7h$ for the worker and the AI, respectively. Both $p_w(h)$ and $p_a(h)$ are monotonically decreasing in h. We model the execution cost and verification



(a) Quality gap with heterogeneous task difficulty



(b) Quality grading with heterogeneous task difficulty

Figure 4: Plots illustrating the worker quality regimes with heterogeneous task difficulty as functions of abilities (α, β), under the default parameter setting $(b_W, \ell_W, b_I, \ell_I, \xi, \tau, C_a) = (8, 6, 14, 12, 0.3, 6.4, 0)$, and the functional choices $p_w(h) = 1 - 0.5h$, $p_a(h) = 1 - 0.7h$, $C_v(s; h) = (0.5 + h)s$, $C_w(\beta; h) = 10(1 - \beta)h$, and $\phi(s; \alpha) = 1 - \frac{1}{1+2\alpha s}$.

cost as $C_w(\beta; h) = 10(1 - \beta)h$ and $C_v(s; h) = (0.5 + h)s$, respectively, both are monotonically increasing in h for any fixed s. All other parameters and functional forms are set as in Figure 2. In particular, when $h \equiv 0.5$, this specification reduces to the setting of Figure 2.

For each difficulty level h, we can compute the worker's optimal action $(d^\star(h), s^\star(h))$ and the resulting quality $Q(W; h)$. Let the task-difficulty distribution be $\mu = \mathrm{Unif}[0, 1]$. The expected worker quality under heterogeneous task difficulty is then

$$Q(W) = \int_0^1 Q(W; h)\, dh.$$

Analogous to Section 3, we can study quality variation and grading outcomes by comparing $Q(W)$ with the no-AI baseline $Q_0(W)$.

Analysis. Figure 4 visualizes how worker quality varies under AI assistance when task difficulty is heterogeneous. The core qualitative phenomena from Figure 2 persist—including phase transitions, compliance gain/loss regimes, and rational over-delegation—supporting the robustness of our results.

7.2 Miscalibrated beliefs about AI capability

Our model implicitly assumes that the AI ability p_a is known, so that the worker can optimize its action given p_a. However, AI performance is often jagged across instances in practice [7]. Let $\widehat{p}_a \in [0, 1]$ denote the success probability that the worker believes the AI to have. This miscalibrated belief modifies the worker utility to

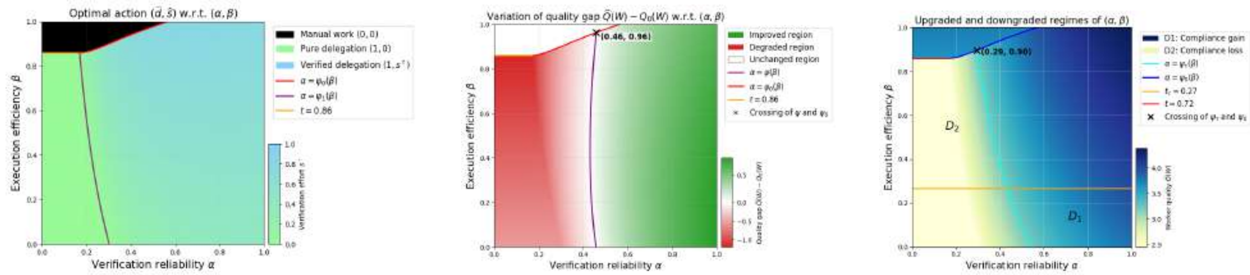
$$\widehat{U}_W(d, s) = \big[(1 - \widehat{p}_a)\big((b_W + \ell_W)p_w - C_w\big)\phi(s) - C_v(s) - (b_W + \ell_W)(p_w - \widehat{p}_a) + C_w - C_a\big]\, d + g_W.$$

Let

$$(\widehat{d}, \widehat{s}) \in \arg\max_{(d,s)\in[0,1]^2} \widehat{U}_W(d, s)$$

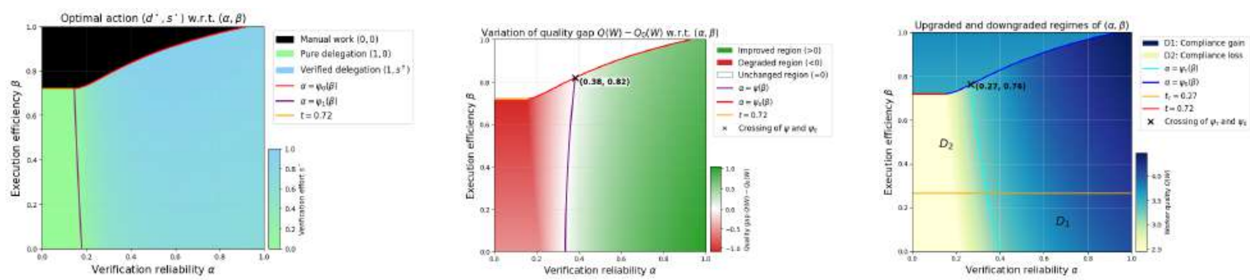
denote the worker's optimal action under the miscalibrated utility. The resulting worker quality is

$$\widehat{Q}(W) := U_I(\widehat{d}, \widehat{s}).$$



(a) Worker action with $\widehat{p}_a = 0.7$ (b) Quality gap with $\widehat{p}_a = 0.7$ (c) Quality grading with $\widehat{p}_a = 0.7$

Figure 5: Plots illustrating the worker action and quality regimes under miscalibrated beliefs about AI capability as functions of abilities (α, β), under the default parameter setting $(b_W, \ell_W, b_I, \ell_I, \xi, \tau, p_a, C_a, p_w, \widehat{p}_a) = (8, 6, 14, 12, 0.3, 6.4, 0.65, 0, 0.75, 0.7)$, and the functional choices $C_v(s) = s$, $C_w(\beta) = 5(1 - \beta)$, and $\phi(s; \alpha) = 1 - \frac{1}{1+2\alpha s}$.



(a) Worker action with $\kappa = 0.8$ (b) Quality gap with $\kappa = 0.8$ (c) Quality grading with $\kappa = 0.8$

Figure 6: Plots illustrating the worker action and quality regimes with partial re-execution cost as functions of abilities (α, β), under the default parameter setting $(b_W, \ell_W, b_I, \ell_I, \xi, \tau, p_a, C_a, p_w, \kappa) = (8, 6, 14, 12, 0.3, 6.4, 0.65, 0, 0.75, 0.8)$, and the functional choices $C_v(s) = s$, $C_w(\beta) = 5(1 - \beta)$, and $\phi(s; \alpha) = 1 - \frac{1}{1+2\alpha s}$.

In particular, $\widehat{Q}(W) = Q(W)$ when $\widehat{p}_a = p_a$.

Analysis. Figure 5 visualizes how worker action and worker quality varies under AI assistance when beliefs about AI capability are miscalibrated, with $\widehat{p}_a = 0.7 > p_a = 0.65$. The core qualitative phenomena from Figure 2 persist—including phase transitions, compliance gain/loss regimes, and rational over-delegation—supporting the robustness of our results. Moreover, over-delegation becomes more pronounced, expanding both the quality-degradation region and the compliance-loss region. This highlights the risks of overestimating AI capability and the importance of calibrating users' beliefs about AI performance.

7.3 Partial re-execution

Our model assumes that after detecting an AI error, the worker fully re-executes the task and incurs the full cost C_w. In practice, re-execution may be only partial, since the worker can often leverage the AI output (e.g., in essay writing) to reduce effort.

Let $\kappa \geq 0$ denote a re-execution discount factor. Under this extension, the worker utility becomes

$$U_W(d,s) = \big[(1-p_a)\big((b_W+\ell_W)p_w - \kappa C_w\big)\phi(s) - C_v(s) - (b_W+\ell_W)(p_w-p_a) + C_w - C_a\big]\, d + g_W,$$

which introduces the additional parameter κ. We study how this extension reshapes worker action and quality.

Analysis. Figure 6 visualizes how the worker's action and quality vary under partial re-execution cost with $\kappa = 0.8$. The core qualitative phenomena from Figure 2 persist—including phase transitions, compliance gain/loss regimes, and rational over-delegation—supporting the robustness of our results. Notably, Figure 6 is nearly identical to Figure 2, suggesting that partial re-execution has only a mild effect on worker quality in this regime.

8 Conclusions, limitations, and future work

This paper develops a formal framework for understanding how AI assistance reshapes institutional worker quality through endogenous delegation and verification decisions. By modeling worker behavior as the solution to a rational optimization problem and evaluating outcomes through an institution-centered notion of quality, we identify a structural mechanism by which AI can induce sharp, nonlinear changes in behavior and outcomes.

We show that AI assistance transforms the mapping from worker ability to workflow choice. Small differences in verification reliability can trigger abrupt regime shifts between manual work, verified delegation, and pure delegation. These phase transitions arise from structural properties of delegation with costly verification and persist under general assumptions on detection and cost functions.

As a result, AI assistance reshapes worker quality heterogeneously. Workers with strong verification ability may experience *compliance gains*, becoming institutionally qualified despite lower execution efficiency, while workers with weaker verification ability may rationally over-delegate and incur *compliance loss*, even when baseline task success improves. Notably, these effects arise endogenously from rational optimization, without invoking behavioral bias or miscalibration.

Taken together, our results suggest that in AI-assisted workflows, the ability to evaluate and verify AI outputs becomes a central determinant of institutional outcomes. AI thus functions not only as a productivity-enhancing tool, but as a structural filter that amplifies differences in verification ability.

Our framework abstracts away from several real-world considerations to isolate this core mechanism. We study a single worker performing a single task, treat verification reliability as exogenous, and do not model dynamic effects such as learning, deskilling, or reputation. Institutional evaluation is outcome-based and does not condition on workflow observability. Moreover, our analysis is structural rather than empirical and does not estimate effect sizes in specific domains.

These limitations suggest several directions for future work. Promising extensions include dynamic models in which execution efficiency and verification reliability evolve over time, analyses of institutional interventions such as verification requirements or incentive schemes, and multi-worker or team-based settings with distributed oversight. Empirical studies mapping real-world tasks and verification practices to model parameters could further enable quantitative assessment of compliance gain, compliance loss, and verification amplification in practice.

Acknowledgments

We thank Arnab Bhattacharyya and Savik Kinger for useful discussions. This work was funded by NSF Award CCF-2112665.

References

[1] Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. *SSRN Electronic Journal*, 2023.

[2] Robert Bergman. The meaning-and illusion-of human oversight of AI, 2025.

[3] Eric Bogert, Aaron Schecter, and Richard T Watson. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, 11(1):8028, 2021.

[4] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025.

[5] L. Elisa Celis, Lingxiao Huang, and Nisheeth K. Vishnoi. A mathematical framework for ai-human integration in work. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

[6] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International conference on algorithmic learning theory*, pages 67–82. Springer, 2016.

[7] Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013), 2023.

[8] Douglas Gale and Martin Hellwig. Incentive-compatible debt contracts: The one-period problem. *The Review of Economic Studies*, 52(4):647–663, 1985.

[9] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):31, 2021.

[10] George Georgiadis and Balazs Szentes. Optimal monitoring design. *Econometrica*, 88(5):2075–2107, 2020.

[11] Wilson Guo, Jason D. Hartline, and Jessica Hullman. The value of information in human–AI decision-making. *arXiv preprint*, arXiv:2502.06152, 2025.

[12] Wilson Guo, Tongshuang Wu, Jason Hartline, and Jessica Hullman. A decision-theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM, 2024.

[13] Richard P Heitz. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, 8:150, 2014.

[14] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. Human-AI collaboration: The effect of AI delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 453–463. ACM, March 2023.

[15] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2018.

[16] Bernard O Koopman. The theory of search. II. target detection. *Operations research*, 4(5):503–531, 1956.

[17] Jean-Jacques Laffont and David Martimort. *The theory of incentives: the principal-agent model.* Princeton university press, 2002.

[18] David Lyell and Enrico Coiera. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431, 2017.

[19] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

[20] Hannah Mayer, Lareina Yee, Michael Chui, and Roger Roberts. Superagency in the workplace: Empowering people to unlock AI's full potential. Research Report January 28, 2025, McKinsey & Company, 2025.

[21] Dilip Mookherjee and Ivan Png. Optimal auditing, insurance, and redistribution. *The Quarterly Journal of Economics*, 104(2):399–415, 1989.

[22] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR, 2020.

[23] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, 2000.

[24] Samir Passi and Mihaela Vorvoreanu. Overreliance on AI literature review. *Microsoft Research*, 339:340, 2022.

[25] Sherwin Rosen. The theory of equalizing differences. *Handbook of labor economics*, 1:641–692, 1986.

[26] Nenad Tomavsev, Matija Franklin, and Simon Osindero. Intelligent AI delegation. 2026.

[27] Robert M Townsend. Optimal contracts and competitive markets with costly state verification. *Journal of Economic theory*, 21(2):265–293, 1979.

[28] Wayne A Wickelgren. Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, 41(1):67–85, 1977.

[29] Tony Zhang and Sencer Melih Deniz. The effect of perceived task difficulty on human overreliance on AI. Humane AI Network, 2024.

[30] Araz Zirar, Syed Imran Ali, and Nazrul Islam. Worker and workplace artificial intelligence (AI) coexistence: Emerging themes and research agenda. *Technovation*, 124:102747, 2023.

Table 2: Notations used in this paper

Symbol	**Domain**	**Description**
d	$[0,1]$	Delegation rate (fraction of task delegated to AI)
s	$[0,1]$	Verification (scrutiny) effort
$p(d,s)$	$[0,1]^2 \to [0,1]$	Task success probability induced by (d,s)
$\mathrm{cost}(d,s)$	$[0,1]^2 \to \mathbb{R}_{\geq 0}$	Total cost induced by (d,s)
$d^\star$	$[0,1]$	Worker-optimal delegation choice
$s^\star$	$[0,1]$	Worker-optimal verification choice
b_W	$\mathbb{R}_{\geq 0}$	Worker valuation per unit success probability
b_I	$\mathbb{R}_{\geq 0}$	Institutional valuation per unit success probability
ℓ_W	$\mathbb{R}_{\geq 0}$	Worker loss per unit failure probability
ℓ_I	$\mathbb{R}_{\geq 0}$	Institutional loss per unit failure probability
ξ	$\mathbb{R}_{\geq 0}$	Institutional cost discount factor
p_w	$[0,1]$	Baseline task success probability of worker
p_a	$[0,1]$	Baseline task success probability of AI
$\phi(s)$	$[0,1] \to [0,1]$	Detection probability as a function of verification effort s
C_w	$\mathbb{R}_{\geq 0}$	Worker task execution cost
C_a	$\mathbb{R}_{\geq 0}$	AI task execution cost
$C_v(s)$	$[0,1] \to \mathbb{R}_{\geq 0}$	Verification cost as a function of scrutiny s
α	$\mathbb{R}_{\geq 0}$	Detection sensitivity parameter in ϕ
β	$\mathbb{R}_{\geq 0}$	Worker task efficiency parameter
$U_W(d,s)$	$[0,1]^2 \to \mathbb{R}$	Worker utility
$U_I(d,s)$	$[0,1]^2 \to \mathbb{R}$	Institutional utility
$Q(\alpha,\beta)$	$\mathbb{R}^2_{\geq 0} \to \mathbb{R}$	Worker quality with AI
$Q_0(\alpha,\beta)$	$\mathbb{R}^2_{\geq 0} \to \mathbb{R}$	Worker quality without AI
$\Delta_Q(\alpha,\beta)$	$\mathbb{R}^2_{\geq 0} \to \mathbb{R}$	Quality gap induced by AI assistance
τ	$\mathbb{R}_{\geq 0}$	Qualification threshold

A Omitted details in Section 4: from data to model parameters

We provide additional details for the example stated in Section 4.

Collab-CXR dataset. The Collab-CXR dataset is organized into multiple experimental designs. We focus on diagnoses collected under **Design 2**, in which each clinician is randomly assigned 60 cases out of the 324 total cases. For each case, the clinician makes predictions under one of four information environments: (1) X-ray only; (2) X-ray + AI-predicted probabilities; (3) X-ray

+ medical history; and (4) X-ray + medical history + AI-predicted probabilities. The study is conducted over four sessions separated by at least two weeks. In each session, the clinician diagnoses cases under exactly one of the four information environments, and across the four sessions each of the 60 cases is eventually observed under all four environments. This design enables direct comparisons of clinician performance with and without AI assistance.

Since the AI diagnoses cases using only the X-ray, we focus on information environments (1) and (2) to enable a fair comparison between the AI and clinicians. Accordingly, we treat environment (1) as the clinician-alone condition and environment (2) as the clinician-with-AI condition.

Details of case success. Given a case j, let $X_a^{(j)}, X_w^{(j)} \in \{0,1\}^{14}$ denote the AI's and the clinician's diagnosis-indicator vectors, respectively. The i-th coordinate equals 1 if the prediction for the i-th pathology matches the expert label. Let $I_a^{(j)}, I_w^{(j)} \in \{0,1\}$ denote the corresponding *case-success* indicators for the AI and the clinician. We define

$$I_a^{(j)} = \prod_{i\in[14]} (X_a^{(j)})_i, \qquad I_w^{(j)} = \prod_{i\in[14]} (X_w^{(j)})_i.$$

Case clean. Case difficulty may vary, which can induce substantial variation in the clinician-alone time cost C_w. To restrict attention to cases of comparable hardness, we retain only those cases whose clinician-alone completion time lies within [0.5×, 2×] the clinician's average completion time.

In addition, there are a small number of cases j with $I_a^{(j)} = 1$ but $I_w^{(j)} = 0$, i.e., cases where the clinician appears to override an AI output that matches the expert labels. This behavior is not captured by our baseline model, which assumes clinicians only detect AI errors. Since such cases account for only 4.9% of all cases, we treat them as erroneous-detection cases and exclude them from the analysis.

After this data cleaning, we retain 38 cases for clinician `vn_exp_65341958`.

Details of deriving $\mathrm{cost}(1, s^\dagger)$. We observe that the average time spent in the clinician-with-AI condition, denoted C_{wa}, is larger than the clinician-alone average time C_w. This is because recorded time includes not only diagnostic reasoning but also time spent interacting with the interface (e.g., reviewing the AI output and entering information). In particular, even when the clinician does not modify the AI output (captured by $X_w = X_a$), there is still nontrivial interface/entry time.

We approximate this fixed entry-time component by C_w and use the decomposition

$$C_{wa} = \mathrm{cost}(1, s^\dagger) + \Pr[X_w = X_a] \cdot C_w.$$

Consequently, we estimate

$$\mathrm{cost}(1, s^\dagger) = C_{wa} - \Pr[X_w = X_a] \cdot C_w.$$