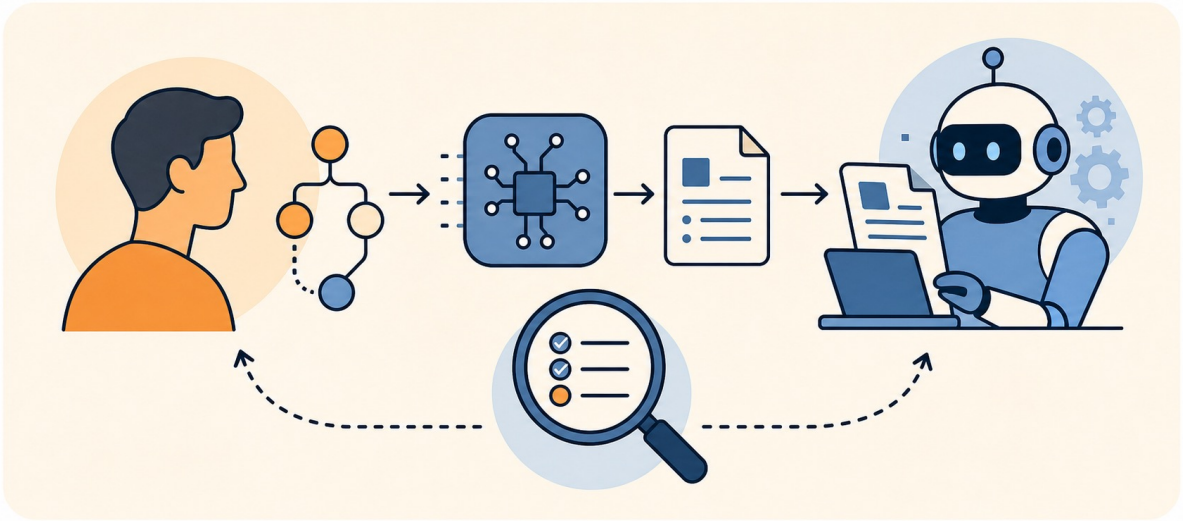
Delegation and Verification Under AI

Lingxiao Huang
Nanjing University

Wenyang Xiao
Nanjing University

Nisheeth K. Vishnoi
Yale University



How does AI reshape worker quality in the AI Age?

ICML 2026

Paper : https://arxiv.org/abs/2603.02961

Motivating Question: Can Cheaper Execution Lower Worker Quality?

Expected effect

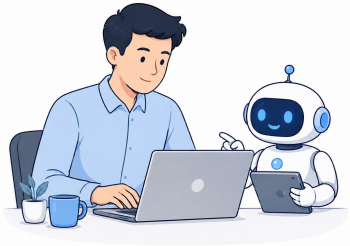
AI assistance

⇓

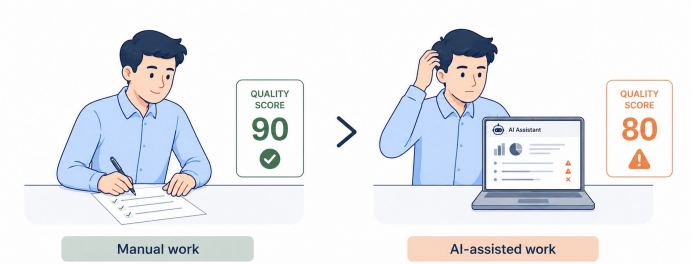
Lower execution cost
Higher execution efficiency

⇓

Always lead to better worker quality ?
+
Better AI, larger quality improvement ?



Observed puzzle



- **(Over-reliance)** Workers rely more on AI advice on harder tasks with lower AI accuracy [Zhang & Deniz, 2024]
- **(Oversight gap)** High-load tasks reduce oversight when AI errors need more scrutiny [Gaube et al., 2021]
- **(Mixed impact)** AI upgrades some workers but downgrades others, despite similar AI access [Brynjolfsson et al., 2025]

Existing explanations

- Attribute oversight decay to cognitive load, verification complexity, and time pressure. [Bergman, 2025]
- Treat uneven AI effects as task allocation or fixed human-skill differences. [Celis et al., 2025]

What is missing?

Underlying mechanism on how AI reshape institutional worker quality is still unclear

Our Contributions

- **(Rational action)** Model worker action (delegation & verification) as optimality for individual utility (task success v.s. cost)
- **(Worker quality)** Define institutional worker quality as institutional gains induced by worker's optimal action
- **(Compliance regime)** *Characterize when AI assistance improves or degrades institutional worker quality; identify which workers are upgraded or downgraded*
- **(Verification amplifier)** *Obtain closed-form expressions for optimal actions and resulting worker quality, which highlights the importance of verification reliability*

Cheaper execution helps only when workers can verify

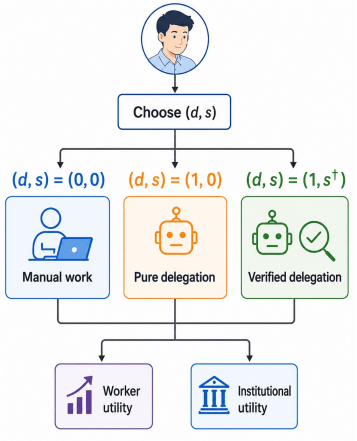
Model: Delegation, Verification, and Worker Quality

Action space & workflow

We study a static, per-task model of AI-assisted work

Worker action space:
- $d \in [0,1]$: delegation prob.
- $s \in [0,1]$: verification effort after delegation

Workflow logic:
- **Manual work**: worker performs the task manually
- **Pure Delegation**: worker accepts AI output directly
- **Verified delegation**: worker checks AI output before accepting or corrcting it



Worker utility

- $p(d,s)$: the success probability
- $cost(d,s)$: the incurred cost
- b_W/ℓ_W: worker's success gain/failure loss

Worker utility: $U_W(d,s) \coloneqq b_W p(d,s) - \ell_W(1-p(d,s)) - cost(d,s)$

Worker action -- Optimizing worker utility $(d^\star, s^\star) \in \arg\ max_{(d,s)\in[0,1]^2}\ U_W(d,s)$.

Worker quality

- b_I/ℓ_I: institution's success gain/failure loss
- ξ: institution-paid worker cost share
- Assume $b_I + \ell_I > \xi(b_W + \ell_W)$, e.g., $b_I = \ell_I, b_W = \ell_W, \xi = 0.3$

Institutional utility: $U_I(d,s) \coloneqq b_I p(d,s) - \ell_I(1-p(d,s)) - \xi cost(d,s)$

Institutional worker quality is induced by worker's optimal action: $Q(W) \coloneqq U_I(d^\star, s^\star)$

- Misalignment: Since $U_I \neq U_W$, rational worker choice always improve worker utility, while can harm institutional worker quality!

Task success probability

Three task-success paths:
- **Manual work**: worker performs task, succeed with $(1-d)p_w$, p_w: worker success rate
- **Pure AI success**: task delegated to AI, succeeds with dp_a, p_a: AI success rate
- **Detect & correct**: if AI fails, worker detects error with $\phi(s)$, redoes task, succeeds with p_w, $\phi(s)$: error-detection probability, e.g., $\phi(s) = 1 - e^{-\alpha s}$ with verification reliability α

Thus, $p(d,s) = (1-d)p_w + dp_a + d(1-p_a)\phi(s)p_w$

Cost function

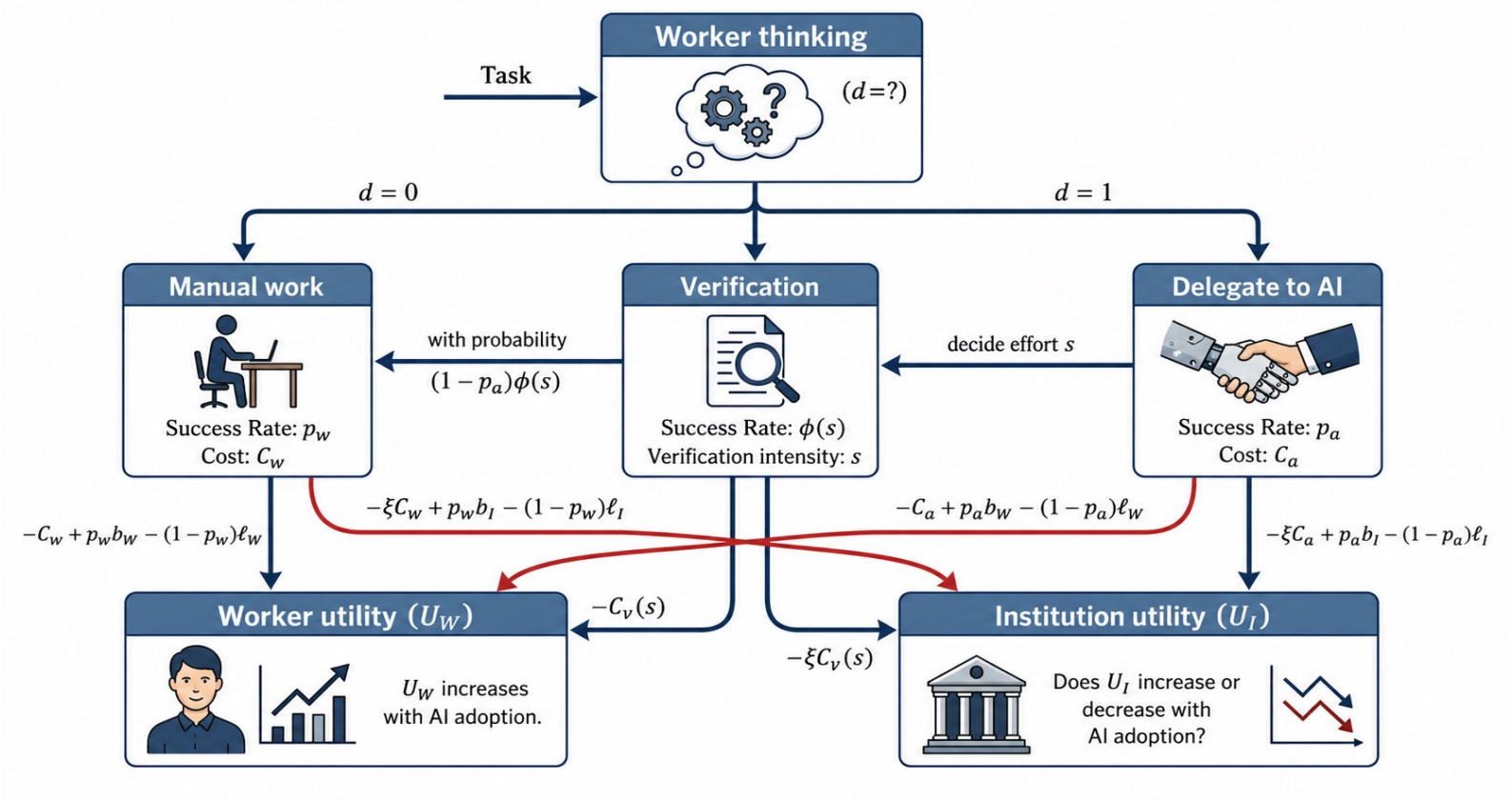
Three cost components:
- **Manual work:** worker execution cost $(1-d)C_w$, e.g., $C_w(\beta) = 1-\beta$ with execution efficiency β
- **AI execution:** AI execution cost dC_a
- **Verify & correct:** worker verifies AI output at cost $C_v(s)$; if AI error detected, worker redoes task with cost C_w

$C_v(s)$: verification cost, e.g., $C_v(s) = s + \frac{s^2}{2}$

Thus, $cost(d,s) = (1-d)C_w + dC_a + dC_{v(s)} + d(1-p_a)\phi(s)C_w$

We analyze how optimal action $(d^\star, s^\star)$ and quality $Q(W)$ vary with verification reliability α and execution efficiency β

Mechanism: How (d, s) Determines Worker Quality



A flow chart for the workflow under AI assistance

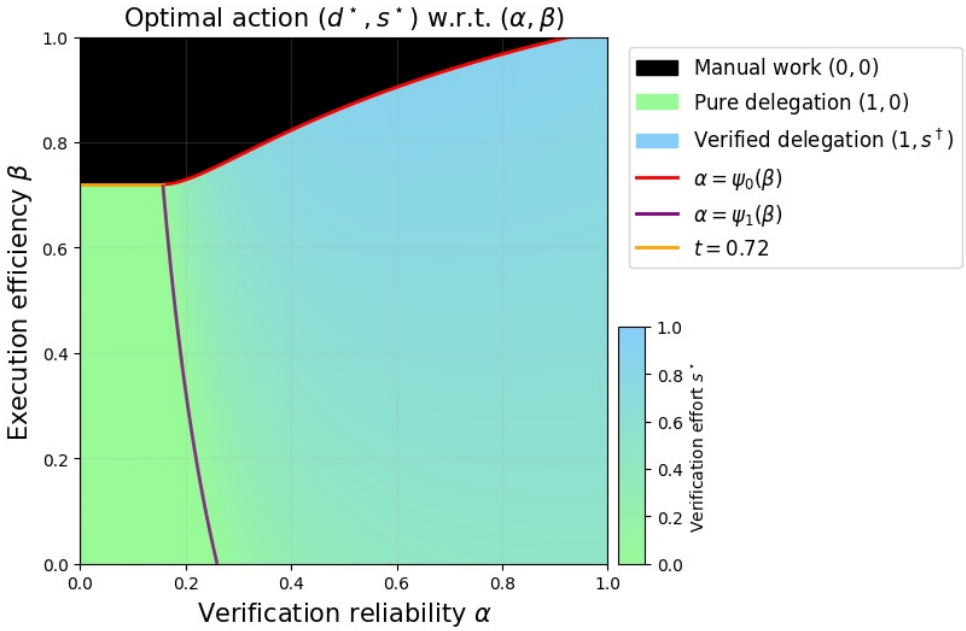
Result 1 — Workers sort into three workflow regimes

Intuition: depending on execution efficiency and verification reliability, the worker either works manually, delegates blindly, or delegates with verification

Theorem (Worker optimal action)

There exists threshold t, monotonically increasing function $\psi_0: [0,1] \to \mathbb{R}_{\geq 0}$, monotonically decreasing function $\psi_1: [0,1] \to \mathbb{R}_{\geq 0}$, function $s^{\dagger}: [0,1]^2 \to [0,1]$, such that the optimal action $(d^{\star}, s^{\star})$ takes the following form:

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



Workflow regimes: workers switch sharply between manual work, pure delegation, and verified delegation

Analysis

- **Three workflow regimes**

Ability space (α, β) is partitioned into manual work, pure delegation, and verified delegation

- **Sharp workflow switches**

Small changes in (α, β) can abruptly switch workflow choices, consistent with observations in [Zhang & Deniz, 2024]

- **Verification reliability drives oversight**

When α is sufficiently high, verification becomes worthwhile, moving the worker from pure delegation or manual work to verified delegation

- **Execution efficiency reduces AI reliance**

Higher β lowers manual execution cost, making manual work more attractive and reducing delegation incentives

Verification reliability separates verified delegation from unchecked reliance.

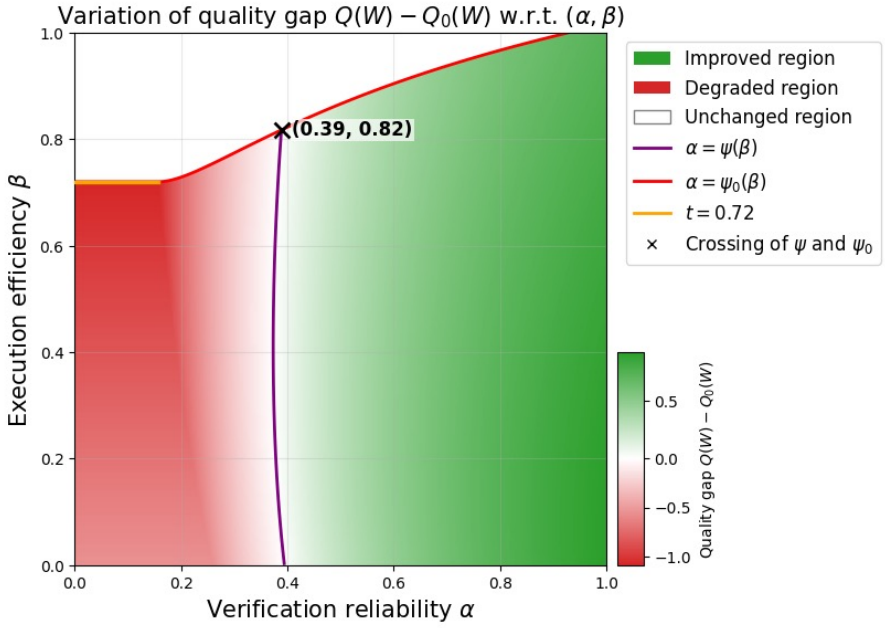
Result 2 — AI can improve or degrade institutional worker quality

Intuition: AI improves quality only when the induced workflow is institutionally valuable. Otherwise, rational over-delegation can reduce quality.

Theorem (Worker quality improvement v.s. no AI)

Let $Q_0(W) \coloneqq U_I(0,0)$ denote no-AI baseline. There exists a continuous function $\psi\colon [0,1] \to \mathbb{R}_{\geq 0}$ such that:

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



Quality regimes: low verification reliability can create quality loss even when AI lowers execution cost

Analysis

- **Three quality regimes**
  Ability space (α, β) is partitioned into quality gain, unchanged, and loss regimes

- **Execution efficiency alone is insufficient**
  Higher β improves no-AI baseline $Q_0(W)$, but AI's effect depends on induced workflow

- **Quality loss stems from rational over-delegation**
  Worker-optimal delegation can be institutionally harmful, without bias or miscalibration

- **Verification reliability separates gain from loss**
  High α enables error detection and correction. Low α leaves delegation weakly supervised and can reduce quality even under rational worker choice

AI does not uniformly raise worker quality: verification reliability α determines whether delegation becomes a gain or a loss

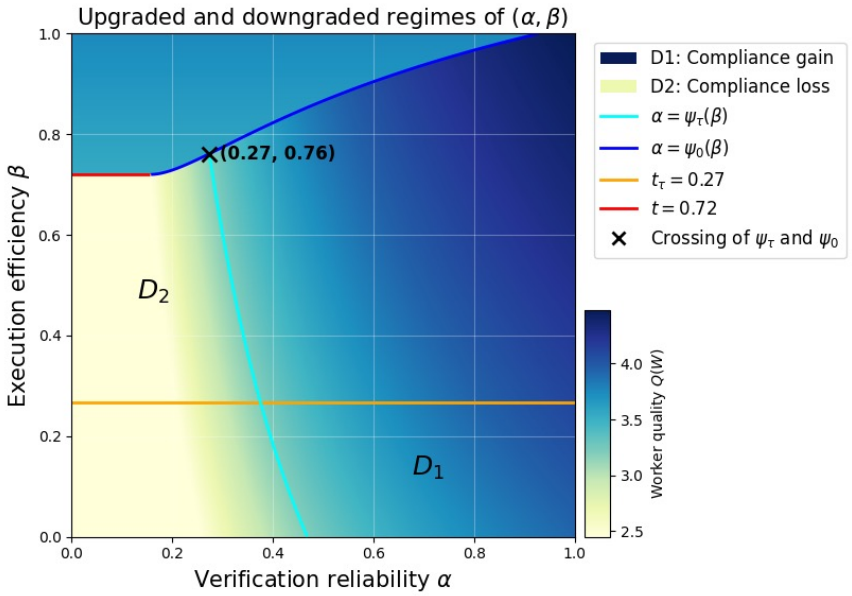
Result 3 — AI can move workers above or below the institutional threshold

Intuition: reliable verifiers can be upgraded by AI, while weak verifiers may be downgraded through over-delegation.

Theorem (Worker upgraded v.s. downgraded)

Let τ be the institutional qualification threshold. There exists threshold t_τ, a decreasing boundary $\psi_\tau: [0,1] \to \mathbb{R}_{\geq 0}$ such that:

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



Compliance effects: AI can move workers above or below the institutional quality threshold

Analysis

- **High α enables compliance gain**
  Workers with low baseline quality can become qualified when reliable verification makes AI delegation institutionally valuable

- **Low α causes compliance loss**
  Workers who were qualified without AI may become unqualified through rational over-delegation and weak oversight

- **Verification amplification**
  AI disproportionately benefits strong verifiers while disadvantaging weak verifiers, widening quality differences

AI amplifies verification ability: high-α workers gain compliance, low-α workers risk compliance loss.

Takeaways, Summary, and Future Work

Worker side

- **Work becomes a delegation–verification choice**

  Workers choose not only whether to delegate, but how much to verify

- **Verification becomes the key sorting dimension**

  High verification reliability enable upgrade; weak verification causes loss with rationally over-delegate

Institution side

- **AI is a structural filter, not just a productivity tool**

  AI amplifies differences in verification ability: strong verifiers benefit, weak verifiers may be downgraded

- **Better AI dosen not automatically improve worker quality**

  Higher AI capability may trigger rational over-delegation among low-α workers

- **Oversight must be designed, not assumed**

  Use verification protocols, correction rules, and rewards for substantive checking

Summary

- Built delegation–verification model for AI-assisted institutional work
- Defined worker quality as institutional utility under the worker-optimal workflow
- Characterized workflow regimes and phase transitions: manual, pure delegation, verified delegation
- Identified verification amplification: compliance gain/loss and rational over-delegation
- Analyzed type-specific interventions with heterogeneous and non-monotonic effects
- Tested robustness across action, belief, task, workflow, output, and regularity variants

Limitations and future work

- Extend beyond the static single-worker setting: dynamic learning, deskilling, and multi-worker collaboration
- Move from structural analysis to empirical estimation: quantify compliance gain/loss in real domains
- From oversight to governance: workflow observability, strategic behavior, responsibility, and fairness

Thank you!

Visit us at Poster · Session 8 · Jul 9, 17:00 · Hall A