When Technology Use Is Observable: Image Concerns in AI-Assisted Work*

David Almog†

This version: August 14, 2026

Abstract

Concerns about how workers are perceived can deter the use of productive workplace technologies. I study this friction in the context of artificial intelligence (AI). In a field experiment on an online labor market, I hired 450 workers to complete a task with AI recommendations and compete for contract extensions, randomly varying whether an HR evaluator observed their AI reliance. Visibility reduced reliance by 4.3 percentage points and accuracy by 2.7 percentage points, even though evaluators observed performance and were instructed to assess future accuracy on the same task. These effects persisted when workers were told evaluators had been reassured about their strong platform records. Leveraging the platform's public feedback feature, I employ a novel incentive-compatible elicitation showing that workers fear AI reliance will be interpreted as low confidence in their own initial judgment. When technology use is observable, anticipated evaluation penalties can lead workers to sacrifice performance.

*I am especially grateful for the guidance and support of Yuval Salant, Benjamin Golub, Alex Imas, Daniel Martin, Alvaro Sandroni, and Jörg Spenkuch. I also thank Meghan Busse, Joshua Dean, John Horton, Alexander Jakobsen, Rafael Jiménez-Durán, Diego Jiménez-Hernández, Andreas Kraft, Ro'ee Levy, Annie Liang, Ryan Oprea, Devin Pope, Mike Powell, Luis Rayo, Avner Strulov-Shlain, and audiences at Google, Harvard Business School, Hebrew University of Jerusalem, Tel Aviv University, Coller School of Management, Monash, University of Queensland, Rutgers-Newark, Advances with Field Experiments Conference, AI in the Workplace at UChicago, AI and the Future of Work at Wharton, Booth Behavioral Lab, and Kellogg Theory and Strategy Seminars for their valuable feedback. Refine.ink provided helpful comments during the writing stage. The experiment was pre-registered on aspredicted.org (239005 and 242197). The IRB ID at Northwestern University is STU00223689.

†Harvard Business School, dalmog@hbs.edu.

1 Introduction

Productive technologies change not only what workers can do, but also what organizations can observe about how work is done. Digital tools often leave detailed records of when workers consult them, follow them, or override them. When these choices are visible to managers, clients, or colleagues, a worker's production method can become a signal of ability, effort, or judgment. Workers may then choose how to produce output partly for what that choice communicates, rather than solely for its effect on performance. This paper asks whether greater observability of a productive input can reduce performance even when output itself is observed and rewarded.

AI is particularly well suited to studying this mechanism because it is both potentially productive and digitally traceable. In many workplaces, humans retain final authority while receiving AI recommendations, combining algorithmic predictions with human judgment. Because workers care not only about performing well but also about how they are viewed by colleagues, managers, and clients (Ellingsen and Johannesson 2008; Bandiera, Barankay, and Rasul 2005), visible AI reliance may itself become a signal. Heavy reliance might be interpreted as adaptability and technological sophistication, but it might instead be seen as reflecting low effort, limited competence, or weak independent judgment. Observability could therefore shift reliance in either direction. Yet evidence from hiring (Hoffman, Kahn, and D. Li 2018), radiology (Agarwal, Alex Moehring, et al. 2023), and pretrial courts (Angelova, Dobbie, and C. Yang 2026) suggests that professionals underuse AI recommendations, leaving substantial value unrealized. Existing explanations emphasize cognitive and information-aggregation biases; this paper asks whether evaluation concerns provide an additional channel.

Identifying this channel with observational data is difficult because workers facing greater scrutiny may differ in other ways, and changes in observability often coincide with changes in incentives or technology.[1] I address this problem with a field experiment on *Upwork*, one of the largest online labor-market platforms. I hired 450 experienced freelance workers to complete an AI-assisted categorization task and compete for higher-paid contract extensions. Workers were randomly assigned to an evaluation environment in which the HR specialist observed either final accuracy alone or both final accuracy and the worker's reliance on AI recommendations. In all conditions, workers knew what the evaluator would observe and were told that the evaluator's objective was to predict accuracy in another session of the same AI-assisted task.

Making AI reliance observable reduced reliance by 4.3 percentage points, or 14 percent relative to the control mean, and lowered final accuracy by 2.7 percentage points. The per-

[1] Angelova, Dobbie, and C. Yang (2024) show that when judges face greater public scrutiny, their use of algorithmic recommendations systematically changes.

formance loss was not driven by reduced effort and was not offset by better targeting of AI overrides. Workers whose reliance was visible spent somewhat longer on their initial choices, but their initial accuracy did not improve; nor did they become better at distinguishing correct from incorrect recommendations. Instead, observability led them to override additional useful recommendations and sacrifice measured performance.

To better understand the mechanism behind these results, I develop a novel incentive-compatible elicitation method using platform feedback to measure what workers fear their AI reliance signals. When the possibility that AI reliance would be disclosed in public feedback was salient, the share of workers choosing to emphasize confidence in their own judgment increased by 13.9 percentage points, more than doubling from 11.9 to 25.8 percent. Follow-up questionnaire evidence supports this interpretation, indicating that while confidence in judgment is generally viewed as the least important trait to signal, in AI-assisted work it emerges as the most important, surpassing both effort and skill. In a subsequent incentive-compatible evaluation exercise, returning workers also placed a substantial negative weight on other workers' AI reliance when predicting their future performance. Generic reassurance about candidates' strong platform records did not attenuate the behavioral response. Together, the results show that visible technology use can become a career-relevant signal and distort the way workers combine their own information with a productive external input.

The rest of the paper is organized as follows. Section 1.1 discusses the related literature. Section 2 presents the experimental design, and Section 3 develops a framework of observable reliance. Section 4 reports the main results and mechanism evidence. Section 5 examines the reassurance intervention and workers' evaluations of their peers. Section 6 concludes.

1.1 Related Literature

This paper contributes to the literature on career concerns and organizational information structures (Holmström 1999; Dewatripont, Jewitt, and Tirole 1999). Most closely, Prat (2005) shows that observing an agent's action in addition to its consequence can distort decisions, while Dai and Singh (2020) show that reputational concerns can discourage experts from using informative diagnostic tools. I provide field-experimental evidence that observing how output is produced can similarly distort workers' use of a productive input, causing them to sacrifice performance even when it is directly observed and rewarded.

A large literature following Spence (1973) studies how workers use costly actions to convey information about their type to employers. Recent work shows that generative AI is rapidly reshaping this signaling environment by changing the cost, content, and informativeness of labor-market communication (Galdin and Silbert 2025; Cui, Dias, and Ye 2025; Wiles and Horton 2025; Cowgill, Hernández-Lagos, and Wright 2026; Gans 2024). This paper

identifies a complementary force: AI affects labor-market signaling not only by transforming established signals, but also by turning workers' reliance on AI into a signal in its own right.

The same mechanism also connects the paper to research on technology adoption and digitization. Technology adoption is often impeded by organizational frictions (Atkin et al. 2017) or limited familiarity with new tools (Almog and Bronsoler 2025). Digitization can create an additional barrier by generating new and durable records of how work is performed. As Goldfarb and Tucker (2012) emphasize, new technologies frequently enable firms to collect previously unavailable information. AI-assisted decision systems are particularly likely to generate such traces because workers' decisions to follow or override recommendations can be recorded and audited (Goldfarb and Tucker 2019). This differs from informal advice exchanged with a co-worker, for which attribution is diffuse and difficult to verify. Consistent with the broader importance of data observability, Houeix (2026) shows that observability slowed the adoption of digital payments among Senegalese taxi drivers. I document a related pattern in workplace production: observability reduces reliance on AI recommendations, with image concerns rather than contract enforcement as the operative mechanism.

The paper also contributes to the rapidly growing literature on AI-human collaboration. I focus on recommendation-based collaboration, in which AI provides advice while humans retain final authority, a structure common in settings such as radiology, lending, and pretrial adjudication. This format is useful for studying image concerns for two reasons. First, it predates newer forms of collaboration such as generative AI, helping reduce concerns that observed behavior is driven by novelty or limited familiarity. Second, reliance is relatively transparent and can be measured directly by observing whether individuals follow or override AI recommendations. [2]

Field studies document substantial departures from algorithmic recommendations in high-stakes professional settings, often suggesting unrealized gains (Agarwal, Alex Moehring, et al. 2023; Angelova, Dobbie, and C. Yang 2026; Hoffman, Kahn, and D. Li 2018; Stevenson and Doleac 2024), a pattern related to the broader literature on *algorithm aversion* (Dietvorst, Simmons, and Massey 2015). Prominent explanations emphasize overconfidence, miscalibration, and failures to aggregate human and algorithmic information.[3] I identify a distinct evaluative channel: by randomizing whether workers' reliance is observable while holding the task, recommendation process, and stated evaluation objective fixed, I show that observability reduces reliance even when doing so lowers accuracy.[4] Underuse can therefore

[2]Goldberg and Lam (2025) examine a relatively rare setting in which generative AI use can be directly tracked, studying its equilibrium effects in a creative-goods marketplace.

[3]Caplin et al. (2025) show that calibration and belief formation shape the gains from AI-assisted decisions, while Balakrishnan, Johnson Ferreira, and Tong (2025) study failures to recognize and combine private information with algorithmic advice.

[4]Related work examines how AI assistance can directly shape decision makers' preferences (McLaughlin and Spiess 2022; Albright 2024; Almog, Gauriot, et al. 2026).

arise not only from what decision makers believe about AI, but also from what they believe reliance on AI communicates about them.

Image concerns shape consequential behavior across domains, including voting, status consumption, educational investment, political contributions, preventive health take-up, and workplace productivity (DellaVigna et al. 2017; Bursztyn, Ferman, et al. 2018; Bursztyn, Egorov, and Jensen 2019; Perez-Truglia and Cruces 2017; Karing 2024; Jee, Karing, and Naguib 2025; Mas and Moretti 2009), and are closely related to the study of stigmatized behaviors (Chandrasekhar, Golub, and He Yang 2018; Celhay, Meyer, and Mittag 2025; Friedrichsen, König, and Schmacker 2018).[5] This paper contributes to these literatures in the increasingly important workplace domain of AI-human collaboration. Emerging evidence suggests that the social meaning of AI use affects both behavior and evaluation: Ash, Lou, and Song (2026) show that AI can alter social-image incentives in online interactions; Haiyang Yang et al. (2025) find in a vignette study that physicians evaluate peers who use generative AI less favorably; Reif, Larrick, and Soll (2025) document workplace evaluation penalties for AI users; and Ling, Kale, and Imas (2026) show that social desirability concerns lead individuals to underreport AI use even in low-stakes surveys. I complement this work with causal evidence from an incentive-compatible design implemented in a real labor-market setting, showing that making AI reliance visible changes workers' behavior and that workers view heavy reliance as carrying negative connotations. [6]

Finally, the paper contributes to the labor literature using online platforms such as Upwork to study market frictions (Pallais 2014; Stanton and Thomas 2016; Horton 2017; Barach and Horton 2021; He, Kang, and Lacetera 2021). Methodologically, I introduce a new incentive-compatible elicitation mechanism that allows workers to select features of the public feedback they receive after job completion, offering a new tool for studying signaling motives in platform environments.

2 Field Experiment Design

I hired 450 workers[7] over a two-week period in July 2025 by posting a 30-minute image categorization job on Upwork under the registered employer name *BuildingAI*, a company that advertises AI annotation solutions on its webpage.[8] Upwork is a leading online marketplace

[5] See Bursztyn and Jensen (2017) for a review of the recent literature on social image concerns.

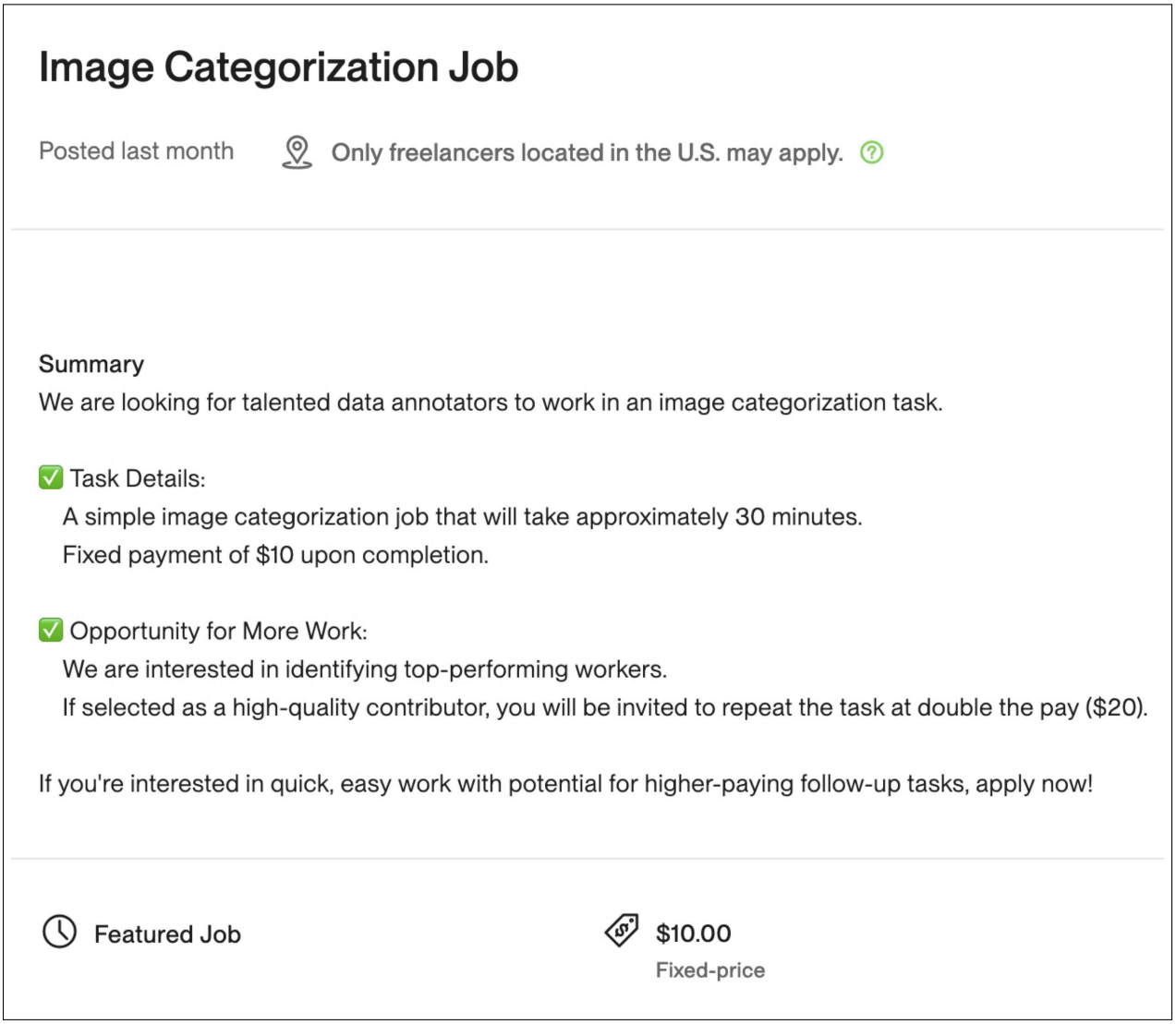
[6] In a companion paper (Almog 2025), I use the same task with a Prolific sample to isolate non-instrumental image concerns, showing that observability can reduce AI reliance even when it has no monetary consequences.

[7] After debriefing, one worker elected to withdraw their data, leaving 449 workers for research purposes, without affecting any of the paper's results.

[8] Screenshots of the *BuildingAI* webpage are available in Appendix B.3.

that connects employers with independent contractors for a wide range of remote tasks. The job offered a fixed payment of $10 upon completion and highlighted the firm's interest in identifying top-performing workers, who would have the opportunity to be invited back to repeat the task at double the pay rate. Figure 1 shows how the job posting appeared on Upwork.

Figure 1: Upwork Job Posting



During the instructions, workers were informed that the firm would extend second job offers to 30% of participants, with decisions guided by feedback from an HR specialist reviewing their anonymized participation profiles. Workers were randomly assigned to one of three experimental conditions, which varied the information available to the HR specialist at the time of evaluation. Importantly, each worker was also made aware of the specific

information that the HR specialist would have in their respective condition.

There is a *Private* condition, which serves as the study's control. In this condition, AI reliance, defined as the frequency with which workers changed their answers to match the AI's recommendation, remains private, and HR specialists observe only task accuracy. Throughout the paper, accuracy will serve as the main performance benchmark and is defined as the percentage of correct answers after considering the AI recommendation. Workers in the *Private* condition were informed of the evaluation process as follows:

> *"The HR specialist will only see each worker's percentage of final correct answers (after considering the AI recommendation)."*

The first treatment group is the *Public* condition. In this condition, HR specialists observe not only workers' accuracy but also their AI reliance. Workers in the *Public* condition were informed of the evaluation process as follows:

> *"The HR specialist will only see each worker's percentage of final correct answers (after considering the AI recommendation) and how often each worker changed their answer after seeing the AI recommendation."*

The second treatment group is the *Public with Evaluator Reassurance* condition. As in the *Public* condition, HR specialists observe both workers' accuracy and their AI reliance. In addition, only workers in this condition are informed that the HR specialist has been assured of the rigorous pre-screening process and the strong track record of all candidates on the platform. This is not new information to the workers, as they were already told on the initial screen that they had been selected based on their strong Upwork history. The intervention therefore manipulates only workers' beliefs about what the HR specialist knows. The exact message provided exclusively to workers in the *Public with Evaluator Reassurance* condition was:

> *"We also assured the HR specialist that all candidates had been carefully pre-screened and had a solid track record of positive experiences on Upwork, minimizing the risk of unnecessary concerns about worker quality."*

Ex ante, I expected reassurance about candidates' screening and platform records to reduce any anticipated penalty arising from evaluator uncertainty about general worker quality. The preregistered prediction was therefore that AI reliance and accuracy would lie between the *Private* and *Public* conditions.

In all experimental conditions, workers were told that HR specialists were explicitly instructed to evaluate candidates based on who was expected to perform most accurately if

invited back for another session of the same AI-assisted categorization task. The instructions highlighted accuracy as the decisive factor in achieving a higher score.

At the end of the categorization task, without prior knowledge of this component, workers participated in a novel, incentive-compatible exercise designed to reveal what they fear high AI reliance may signal to employers. All workers were told that our firm typically leaves public feedback for employees, intended to be informative for future employers on the platform. They were informed that the feedback would include a brief description of the completed task, in this case, image categorization with AI assistance.

Workers were then given the opportunity to select one of three statements they wished to emphasize at the end of their feedback:[9]

(i) You are a hard worker and put in strong effort.

(ii) You are highly skilled and capable in these tasks.

(iii) You have confidence in your own judgment.

In the treatment groups, a few modifications were introduced. Workers were told that their public feedback would also indicate whether they relied more or less on the AI recommendations than the average worker for the same job. They were further reminded that heavy reliance on AI might raise concerns for future employers, and that their selected sentence could help address those concerns if they were identified as high AI users.[10]

For the control group, this exercise captures the baseline distribution of preferences over signaling effort, skill, or confidence in judgment. For the treatment groups, it captures preferences over these same signals under the potential threat of being exposed as a high AI user. Taken together, this allows us to illustrate, at the aggregate level, what workers fear their AI use may signal to employers.

After all workers completed the first job, an experienced HR specialist hired on Upwork was familiarized with the task and proceeded to assign each worker a score from 0 to 100, adhering strictly to the advertised information in their condition (observing only accuracy for the control group, and both accuracy and AI reliance for the treatment groups). Based

[9]These three options reflected the main mechanisms identified in a pilot survey as reasons why workers may fear using AI when it is observable.

[10]Northwestern's human subjects committee was concerned that feedback evaluations disclosing AI reliance could harm workers. In response, treatment group workers were asked a follow-up question on whether they wanted the AI reliance information included in their feedback. They were informed that if at least one worker objected, the information would be removed for everyone for fairness considerations. Because about 30% of the workers requested its removal, the information was ultimately excluded.

on a rehiring rule that depended on both score and chance (described in the next section), we offered contract extensions to 140 workers, of which 135 accepted within one week.[11]

For rehired workers originally assigned to a treatment group, the second job began with an evaluation exercise. They assessed 20 profiles of other returning workers based solely on accuracy (correct answers) and AI reliance rates, replicating the HR specialist's evaluation in the first job. This component was designed to identify whether changes in AI reliance induced by the treatment were driven solely by anticipated external evaluation or also reflected self-projection into the evaluator's role. After this, all rehired workers completed a short questionnaire followed by a 10-round version of the categorization task. Control-group workers did not perform the evaluation task, since AI reliance had not been visible in their prior environment and introducing it at this stage would have been unnatural.

The evaluation exercise was incentive compatible: workers were told that they would be matched with one of the evaluated workers, and that the higher a worker ranked in their ordered score ranking, the greater the chance of being paired with them. Each worker could earn up to an additional \$5, receiving \$0.25 for every correct answer of their matched partner and \$0.25 for each of their own correct answers. This design provided incentives both to evaluate carefully and to perform as well as possible themselves.

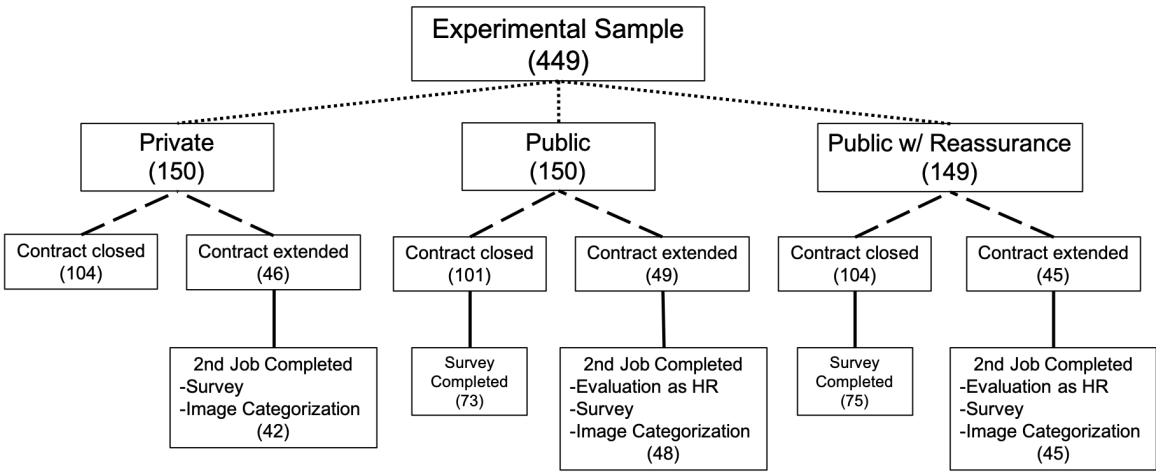
Finally, workers from the treatment groups who were not selected for a contract extension were given an opportunity to earn \$5 by completing the short questionnaire component of the second job, of which 148 out of 205 agreed. Because most of the questions were relevant only for treated workers, they were the ones invited to complete the questionnaire. Figure 2 presents a diagram of the experimental design.

2.1 Upwork and Sample Selection

Upwork (formerly oDesk) has been widely used in field experiments because it combines real-market conditions with strong experimental control (Horton, Rand, and Zeckhauser (2011) discuss in detail the virtues of using online labor markets for conducting experiments). For example, Pallais (2014) examined hiring inefficiencies caused by information frictions, and Coffman, Collis, and Kulkarni (2024) registered as an employer to study gender differences in job applications. A key advantage of Upwork for this study is its ability to foster authentic employer–employee relationships: the platform enables the use of actual contract extensions as incentives. Recruited workers perform data annotation professionally or as a side job and maintain reputations for doing that through ratings and public reviews. This provides a unique opportunity to study how workers in their natural workplace environment respond

[11] The 96.4% acceptance rate reinforces that the rehiring incentives were strong and that workers on the platform are committed to longer-horizon jobs.

Figure 2: Experiment Design



Notes: Short dashed lines denote random assignment, while long dashed lines denote quasi-random assignment based on the rehiring probability rule. The number of workers in each node is reported in parentheses.

to an exogenous shock that makes their AI use observable to someone with influence over their career prospects.

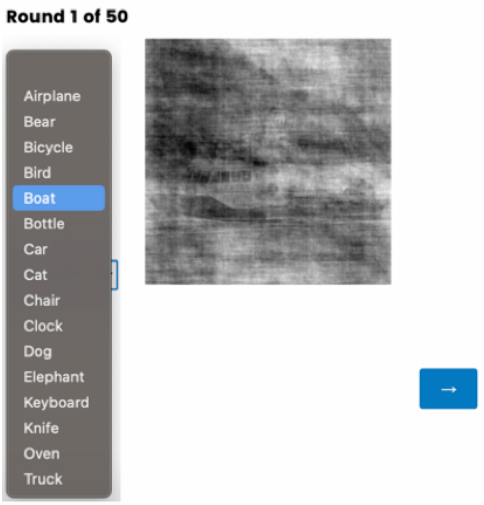
To participate, workers had to satisfy three criteria. First, they had to be located in the United States, situating the findings in a region with substantial AI exposure while reducing concerns about language barriers or misinterpreted instructions. Second, workers needed prior platform experience in related fields such as data annotation or data entry. Because prior work has shown that new workers face substantial barriers (Pallais 2014; Stanton and Thomas 2016), I restricted the sample to workers with previous experience; 81% had already earned over $100 on the platform. Finally, workers had to be registered as freelancers rather than part of an agency, ensuring direct business relationships without intermediaries.

2.2 Image Categorization Task

The first job consisted of 50 rounds of an image categorization task with AI recommendation assistance. In each round, the workers observed a blurred image with the objective of selecting the correct category from a list of 16 options: airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven, and truck. After making their initial choice, workers were shown an AI recommendation. If their choice matched the AI recommendation, they were notified of the agreement and proceeded to the next image. If the recommendation differed, they could revise their answer and switch to the AI's recommendation. Workers were incentivized to put effort into their initial choice because, once the AI recommendation was revealed, they could only choose between their original answer

and the AI's recommendation; all other options were no longer available. The dynamics of this two-stage decision process are illustrated in Figure 3. The images and AI model were obtained from Steyvers et al. (2022), and Appendix B.4 provides additional technical details on the task construction.

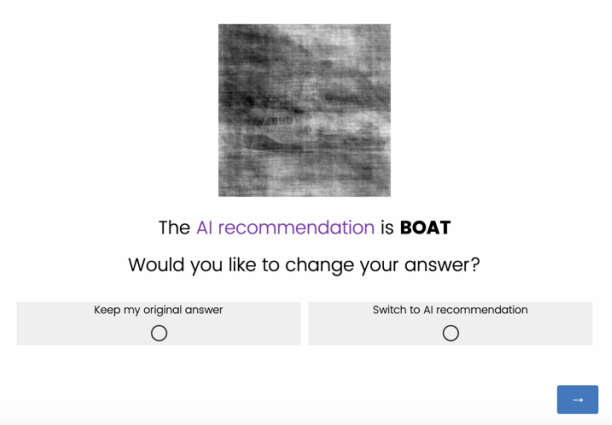
Figure 3: Task interface showing the initial choice and AI recommendation stages.



Initial Choice



When the recommendation agrees



When the recommendation differs

Although AI has made significant progress in image categorization, human input remains essential in supervised learning processes such as data annotation. This task is particularly well-suited for studying AI-human collaboration, as it encompasses: (i) images that appear simple to humans but are misclassified by the AI; (ii) images that the AI classifies correctly but are extremely difficult for humans, making blind delegation to the AI the most effective strategy; and (iii) cases where humans often classify correctly, but those who do not may be guided to the correct answer after reconsidering the image from the perspective suggested

by the AI.[12]

Additional advantages of this task are that it is a well-established and popular category on online labor platforms such as Upwork,[13] it is easy to explain and requires little to no training (especially for platform users with prior annotation experience), and it offers excellent experimental control. The dataset further provides ground-truth labels, presents each image at varying noise levels to manipulate difficulty, has been validated as an efficient tool for studying AI-human collaboration, and Steyvers et al. (2022) also made available high-quality AI recommendations.

2.3 Design Choices

In this section, I discuss several of the main choices I faced while designing the experiment.

Between-subjects. Workers are assigned to a single experimental condition to avoid experimenter demand or priming effects. While a within-subjects design could provide insight into the distribution of treatment effects, which is particularly relevant here since effects may vary in direction, the risk of eliciting mechanical responses to treatment is substantial.

Recommendation timing. Providing the recommendation after an initial choice allows us to clearly identify AI's impact on the final decision. This approach avoids ambiguity in cases where the choice and recommendation align, but it is unclear whether the choice was influenced by the AI or made independently. This design has two key advantages: first, it enables precise measurement of how often recommendations are adopted; second, it conveys to workers that their choices to adopt recommendations are being reliably tracked. In some work environments, identifying which employees rely more heavily on recommendations can take considerable time (e.g., days or even months), but this study lacks the timeframe for extended observation, which would likely be necessary if recommendations were provided before making an initial choice.

Information provided on AI recommendations. Workers were informed that the recommendations came from an AI model trained on the same task and that it was accurate 85% of the time, a statement that holds for both the full dataset and the subset of images used in this experiment. They were also told that AI might perform well on images that humans find difficult but could occasionally miss images that appear easy to the human eye. This statement reflects a genuine feature of the original study by Steyvers et al. (2022). Prior research (Dietvorst, Simmons, and Massey 2015; Dreyfuss and Raux 2024) shows that people may overreact to a single poor algorithmic recommendation, so this disclaimer was intended to reduce the likelihood that participants would lose trust in AI based on an isolated er-

[12] See Figure A.1 for examples.

[13] Upwork has a job category named "AI Data Annotation and Labeling."

ror. Such reactions represent an interesting phenomenon that is orthogonal to my research question, and given the short horizon of this study, I found it preferable to avoid them.

Incentives. Workers receive a fixed $10 payment for completing the job. However, the most important aspect of the incentive scheme is the opportunity to be rehired for a similar task at double the pay rate ($20). The rehiring opportunity allows workers to derive instrumental value from third-party beliefs, here represented by the evaluation of an HR specialist. This approach mirrors a realistic aspect of the labor market, where employees care about how others perceive them, as these perceptions can ultimately affect hiring, promotion, or layoff decisions.

Rehiring decisions. As advertised to workers, second job offers were made using feedback from a real HR specialist hired on the same platform. Rehiring decisions depended on both score and luck, with the probability of rehiring for each worker determined by the following formula:

$$P(\text{Rehire}) = 0.2 + (\text{Decile} \times 0.02), \quad \text{Decile} = 1, \ldots, 10$$

Deciles were determined by workers' scores within their experimental condition. The rehiring rule served two purposes. First, it rewarded higher scores. Although workers were never told explicitly that higher scores would lead to better outcomes, it was natural to interpret the scores in this way, and the design relies on this perception.[14] Second, it ensured that the pool for the second job was more representative, rather than consisting only of top performers. This design makes the second-job results more generalizable. Although the rule does not guarantee a realized rehiring rate of exactly 30%, it implies an expected rate of approximately 31%, close to the advertised share.

Fixing the supply of second job offers. Fear of AI-driven replacement could provide an alternative explanation for changes in AI adoption if workers believe that relying on AI may either reveal their relative value compared with the algorithm or help train and improve the very system that could eventually replace them, as in Cullen, D. Li, and S. Li (2026). Advertising that second-job offers would be extended to 30% of participants helps mitigate workers' concerns that their performance in the current task could affect the future availability of work. This design choice helps isolate the signaling channel by focusing workers' attention on how AI reliance shapes perceptions of their current value relative to other workers, rather than on fears that AI adoption will alter the future supply of work.

Treatment design. The variation across experimental conditions was designed to differ

[14] Figure A.2 suggests that workers were not reporting a sophisticated strategy aimed at maximizing the probability of clearing the rehiring threshold. Instead, their responses point to a simpler objective of trying to do as well as possible overall, consistent with the model presented in the next section.

only in the information available to the HR specialist at the moment of evaluation, and workers were informed openly and transparently about what information the HR specialist would have at their disposal. Across all conditions, participants were also made aware that the HR specialist was instructed to evaluate solely on which workers were expected to perform most accurately if brought back for another session of the same AI-assisted categorization task.

3 A Framework of Observable Reliance

I introduce a framework of observable reliance on external advice and then map it to the AI-assisted categorization task used in the experiment. In each round, the worker first makes an unaided choice and holds a private confidence p in the selected category. The worker then receives a recommendation that she evaluates using a fixed accuracy assessment κ. When the recommendation differs from the initial choice, the worker decides whether to retain that choice or adopt the recommendation. An evaluator always observes final performance a and may additionally observe reliance r, defined as the share of decisions in which the worker switches to the recommendation. The worker chooses when to rely on the recommendation in order to maximize the score she expects the evaluator to assign. When reliance is unobserved, the worker's decision depends only on its contribution to performance. When reliance is observable, however, the production method itself may affect evaluation, shifting the worker's decision rule away from the performance-maximizing benchmark. Taking the worker's anticipated evaluation rule as given, the framework formalizes this distortion and yields testable predictions for how observability affects both reliance and performance.

3.1 A Model of Worker Behavior under Observable Reliance

In the experiment, the external recommendation is generated by an AI system, reliance is measured by switching to its recommendation, and performance is final categorization accuracy. Consider a categorization task that consists of n distinct categories (in the experiment, $n = 16$). The true category, or *state*, is denoted by $\omega \in \{\omega_1, \ldots, \omega_n\}$, and the worker chooses a category $y \in \{y_1, \ldots, y_n\}$. An action is correct when $y = y_i$ for the corresponding state $\omega = \omega_i$; that is, when the true category is chosen.

A worker has baseline type $\theta \in [1/n, 1]$ summarizing her average *unaided* accuracy (bounded below by $1/n$, which corresponds to random choice). Before seeing the AI recommendation, she selects an initial category y^0 from a private signal. The AI then recommends a category $\hat{y}$, which the worker evaluates using a fixed assessment $\kappa \in (1/n, 1)$ of the recommendation's accuracy. Because the AI is a black-box information source, the worker does not

know how its recommendation is statistically related to her own private signal. I therefore take a reduced-form no-updating approach: the worker does not revise either her confidence in her initial answer or her assessment of the AI's accuracy based on whether the two answers agree or disagree.[15]

Let y denote the worker's final label after considering the AI recommendation. We track two reduced-form objects across tasks:

$$a \equiv \Pr(y = \omega) \quad \text{(expected accuracy)}, \qquad r \equiv \Pr\left(y = \hat{y} \neq y^0\right) \quad \text{(AI reliance)}.$$

Thus a is average accuracy after considering the AI recommendation, and r is the share of tasks on which the worker *changes* her initial answer to the AI's recommendation (only relevant when $y^0 \neq \hat{y}$).

The worker decides whether to switch to the AI's recommendation in order to maximize the score assigned by an evaluator who observes certain variables. In the control regime the worker's subjective expected evaluation scoring rule depends only on accuracy,

$$S_c = S_c(a),$$

whereas in the treatment regime it also depends on AI reliance,

$$S_t = S_t(a, r).$$

The framework places no restriction on whether the worker's perceived evaluation rule is correct. The anticipated penalty may reflect accurate beliefs about evaluators, mistaken beliefs, or a social norm against heavy AI reliance.

Assumption (Accuracy is rewarded). The expected scoring rule is strictly increasing in accuracy:

$$\frac{dS_c(a)}{da} > 0 \quad \text{and} \quad \frac{\partial S_t(a, r)}{\partial a} > 0 \qquad \forall\, a \in (0, 1),\ r \in [0, 1].$$

3.2 Decision Rule and Threshold Optimality

I deliberately take a reduced-form approach to belief formation: I do not model priors, signals, or an updating rule. All we require is that, after choosing a category y^0, the worker can summarize her information by a scalar

$$p \equiv \Pr(y^0 = \omega \mid \text{worker's private information}) \in \left[1/n,\ 1\right],$$

[15] This abstraction is consistent with the experimental instructions, which prompted workers to treat the AI's recommendation as independent of their own signal.

interpreted as the probability that her initial choice of category is correct.[16]

When $y^0 \neq \hat{y}$, the worker chooses one source to follow. Under the reduced-form no-updating approach, she evaluates the accuracy gain from switching to the AI as $\kappa - p$. To isolate the evaluation channel, I take this perceived gain to coincide with the expected change in correctness.

A simple and natural decision rule is a cutoff policy: the worker adopts the AI's recommendation whenever her own confidence p is below a threshold $\tau \in [1/n, 1]$, and otherwise keeps her initial choice y^0.

For the switching problem, which is only relevant on disagreement tasks, assume that $p \mid (y^0 \neq \hat{y})$ admits a cdf F with density f satisfying

$$f(p) > 0 \quad \text{for all } p \in (1/n, 1).$$

Assumption (Single-peaked objective). The worker's expected evaluation score induced by a cutoff policy is strictly quasiconcave in the cutoff τ.

Proposition 1 (Threshold optimality) (i) Scoring when AI reliance is not observed. *If the worker maximizes $S_c(a)$ with $S_c'(a) > 0$, the accuracy-maximizing policy is the cutoff rule with*

$$\tau^* = \kappa, \qquad \textit{i.e., switch to AI iff } p \leq \kappa.$$

(ii) Scoring with observable AI reliance. *Suppose the worker maximizes $S_t(a, r)$, where S_t is continuously differentiable, $\partial S_t/\partial a > 0$, the single-peaked-objective assumption holds, and the optimal cutoff is interior. Then an optimal policy is again a cutoff rule:*

$$\tau^* \;=\; \kappa - \lambda^*, \qquad \lambda^* \;\equiv\; -\frac{\partial S_t/\partial r}{\partial S_t/\partial a}\bigg|_{(a^*, r^*)}$$

Here, λ^* is the worker's perceived marginal evaluation penalty on reliance, measured in accuracy units. To simplify notation, I suppress the star below. If workers anticipate that reliance is penalized ($\partial S_t/\partial r < 0$) then $\tau^* < \kappa$; if workers anticipate that reliance is rewarded ($\partial S_t/\partial r > 0$) then $\tau^* > \kappa$.[17]

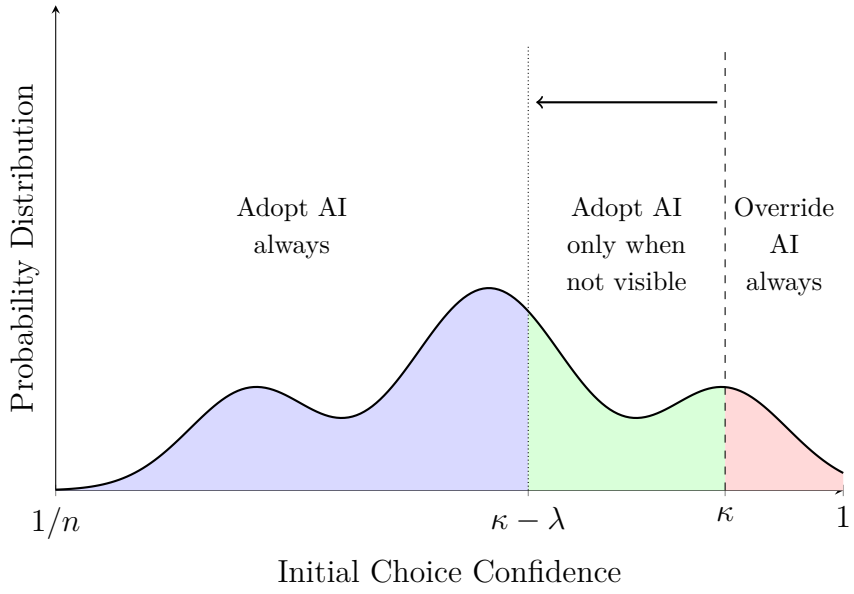
When workers anticipate AI reliance to be penalized, they are willing to ignore AI recommendations for any private confidence level $p \in (\kappa - \lambda, \kappa)$, a range in which accuracy would

[16] A simple microfoundation is Bayesian: the worker begins with a prior over categories, observes a private signal s, forms posteriors $q_j(s) = \Pr(\omega = \omega_j \mid s)$, chooses $y^0 = \arg\max_j q_j(s)$, and sets $p = \max_j q_j(s)$. The analysis relies only on the scalar p and abstracts from the underlying structure that generates it.

[17] For simplicity, I focus on the interior case $1/n < \tau^* < 1$.

improve by following the AI recommendation, but not enough to offset the expected penalty for relying on AI. Conversely, when AI reliance is expected to be rewarded, the worker is willing to adopt AI recommendations when her private confidence level $p \in (\kappa, \kappa-\lambda)$, a range in which accuracy would decrease by following the AI recommendation, but the reward for adopting AI more than compensates for the loss in accuracy. Figure 4 illustrates how the threshold rule dictates whether to adopt or override the AI recommendation as a function of initial choice confidence.

Figure 4: Confidence-Based Threshold Decision Rule



Notes: The distribution of private confidence p is partitioned by two different thresholds that dictate whether to adopt AI. When p is below the threshold, the worker follows the AI recommendation; when p is above it, she keeps her original choice. The higher threshold κ represents the optimal rule when AI reliance is not observable. When reliance becomes observable and evaluators penalize it, the threshold shifts left to $\kappa - \lambda$. The shaded green region between $\kappa - \lambda$ and κ highlights initial confidence levels for which adopting the AI recommendation would improve accuracy, but the worker instead overrides whenever reliance is visible.

3.3 Solution and Testable Predictions

For any cutoff $\tau \in [1/n, 1]$, overall AI reliance is

$$r(\tau) = DF(\tau),$$

where $D \equiv \Pr(y^0 \neq \hat{y})$ is the probability of disagreement. Overall expected accuracy is

$$a(\tau) = \theta + D \int_{1/n}^{\tau} (\kappa - p) f(p)\, dp,$$

where θ denotes baseline unaided accuracy. The integral captures the expected gain from following the AI rather than retaining the initial choice on disagreement.

Using Proposition 1, let $\tau_c^* = \kappa$ denote the optimal cutoff in the control regime and let $\tau_t^* = \kappa - \lambda$ denote the optimal cutoff when reliance is observable. Hence, for $j \in \{c, t\}$,

$$r_j^* = DF(\tau_j^*), \qquad a_j^* = \theta + D \int_{1/n}^{\tau_j^*} (\kappa - p) f(p)\, dp.$$

Model predictions. Moving from the control benchmark to a regime in which evaluators observe AI reliance, and given that the worker's anticipated evaluation score $S_t(a, r)$ is strictly increasing in accuracy ($\partial S_t / \partial a > 0$), the model implies:

1. **AI reliance.** If $\partial S_t / \partial r < 0$ at the treatment optimum, then

   $$\lambda > 0, \qquad \tau_t^* < \tau_c^* = \kappa, \qquad r_t^* < r_c^*.$$

   If $\partial S_t / \partial r > 0$, these inequalities reverse.

2. **Accuracy.** Whenever observability shifts the optimal cutoff away from κ,

   full support implies

   $$a_t^* \; < \; a_c^*.$$

   That is, accuracy is strictly lower under any nonzero visibility-induced distortion of the optimal cutoff. Appendix C.2 shows that accuracy can be represented as an inverse-U-shaped function of AI reliance, attaining a maximum at $\tau = \kappa$ (the optimal cutoff under the control regime).

2.a **Perceived accuracy conditional on switching and not switching on disagreement tasks.** Under the maintained no-updating assumption, the worker continues to evaluate the AI recommendation using the fixed assessment κ, regardless of the cutoff that determines when she follows it. Hence, perceived expected accuracy conditional on switching and not switching are, for $j \in \{c, t\}$,

$$\tilde{a_j}^S = \kappa \qquad \text{and} \qquad \tilde{a_j}^{NS} = \mathbb{E}[p \mid p > \tau_j^*].$$

Because this conditional expectation is strictly increasing in the cutoff, an anticipated penalty on AI reliance implies

$$\tilde{a_t}^S = \tilde{a_c}^S = \kappa \qquad \text{and} \qquad \tilde{a_t}^{NS} < \tilde{a_c}^{NS}.$$

Implication (penalty case). If workers anticipate a penalty on AI reliance ($\partial S_t / \partial r < 0$), then making reliance observable yields (i) lower AI reliance, (ii) lower accuracy, and (iii) unchanged perceived accuracy conditional on switching, and lower perceived accuracy conditional on not switching.

4 Experimental Results

Table 1 provides a summary of workers' characteristics. Gender and ethnicity were self-reported by the workers, while the remaining variables, encompassing platform history and previous education, were collected from their public profiles. Two-sided t-tests indicate that treatment assignment was balanced across the available observable characteristics, with no variable showing a statistically significant difference at the 10% level.

Table 1: Descriptive Statistics and Balance Check.

	Private	Public		Public w/ Reassurance	
	Mean (sd.)	Mean (sd.)	t-stat	Mean (sd.)	t-stat
Female	0.69 (0.46)	0.68 (0.47)	0.25	0.67 (0.47)	0.41
Minority	0.43 (0.50)	0.46 (0.50)	-0.58	0.50 (0.50)	-1.22
Previous Earnings	8,262 (30,292)	7,958 (19,757)	0.10	7,635 (24,275)	0.20
Previous Jobs	15.2 (17.1)	17.9 (20.1)	-1.24	15.5 (20.0)	-0.11
Hourly rate	20.3 (9.6)	21.1 (9.3)	-0.75	21.4 (14.7)	-0.77
Badge	0.35 (0.48)	0.39 (0.49)	-0.72	0.40 (0.49)	-1.00
Undergraduate degree	0.65 (0.48)	0.66 (0.48)	-0.24	0.70 (0.46)	-1.07
Graduate degree	0.20 (0.40)	0.27 (0.44)	-1.36	0.27 (0.44)	-1.40
Observations	150	150		149	

Notes: *Minority* is a dummy variable equal to one for workers identifying with any non-white ethnicity. *Previous Jobs* was winsorized at the 95th percentile, with values above this percentile set to the 95th-percentile value. *Badge* is a dummy variable equal to one for workers holding any of the platform badges: 'Rising Talent', 'Top Rated', or 'Top Rated Plus'. The t-statistics reported correspond to two-sided t-tests comparing the private (control) group to either of the public (treatment) groups.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.1 Treatment Effects

For ease of exposition, I pool the two treatment groups in which AI reliance was public. These two conditions produced very similar effects, with no statistically significant differences across any outcome variable.[18] Section 5 discusses in greater detail the implications of not finding differences between the two *public* treatments. Unless otherwise noted, the control group refers to the *Private* condition, and the treatment group combines the *Public* and *Public with Evaluator Reassurance* conditions.

[18]Table A.1 reports results with the treatments separated, showing nearly identical and statistically significant estimates when each treatment is compared separately to the control group.

The empirical analysis relies on the following simple regression framework:

$$Y_{ij} = \alpha + \beta T_i + \gamma X_j + \epsilon_{ij} \tag{1}$$

where Y_{ij} denotes the outcome variable for worker i on image j. The treatment indicator $T_i \in \{0, 1\}$ equals 1 if worker i was assigned to either of the public treatment groups, and 0 otherwise. X_j represents image fixed effects. The coefficient of interest, β, captures the average treatment effect of being assigned to a condition in which AI reliance was visible to the HR specialist, relative to the control group, where AI reliance was not visible. All specifications include image fixed effects, and standard errors are clustered at the worker level.[19] Treatment effects for the six pre-registered outcome variables are reported in Table 2.

Table 2: Treatment Effects.

	AI recommendation reliance		Correct answer		Response time	
	All	Conditional	Initial	Final	Initial	Rec. stage
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	**-0.043*****	**-0.080*****	**0.008**	**-0.027*****	**2.10***	**-0.72**
	(0.014)	(0.022)	(0.012)	(0.008)	(1.10)	(0.57)
Constant	0.305	0.640	0.553	0.791	21.3	10.1
Observations	22,398	10,554	22,398	22,398	22,398	10,554

Notes: Each dependent variable is regressed on a treatment group indicator (equal to 1 if the worker was assigned to any of the *public* treatment groups), with image-specific fixed effects included. Standard errors are clustered at the worker level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The first main result is that assignment to a public treatment reduced workers' reliance on AI recommendations. Columns 1 and 2 of Table 2 report the treatment effects on AI reliance.[20] Across both measures, we observe a statistically significant decrease in AI reliance. Using Column 1 as the main benchmark, workers in the treatment groups followed the AI recommendation 4.3 percentage points (pp) less often, significant at the 1% level. This effect represents a 14% reduction relative to the 30.5% AI reliance rate observed in the control group, where workers had no concern about an HR specialist observing their AI use.

[19] Results are robust to excluding image fixed effects.

[20] Column 1 includes all rounds (the unconditional reliance rate), which is also the measure observed by HR evaluators and therefore the outcome most relevant for workers' image concerns. Column 2 restricts to only rounds in which workers had the opportunity to switch from their initial answer to the AI recommendation. When the AI recommendation matched a worker's initial choice, this does not count as relying on AI.

In Column 2, which focuses only on rounds where workers faced a genuine choice between their own answer and the AI recommendation, the estimated effect is naturally larger at 8 pp.

Making AI reliance rates visible to HR specialists reduced workers' reliance on AI, which in turn led to worse performance in the categorization task, even when that performance itself was observable to HR. Column 4 of Table 2 reports the treatment effects on the percentage of correct answers after considering the AI recommendation, our measure of job performance. By ignoring AI recommendations more often, workers in the public conditions experienced a 2.7 pp decline in accuracy, significant at the 1% level. Relative to the 79.1% accuracy rate observed in the control group, this corresponds to a 3.4% performance reduction. Figure A.3 shows that the treatment effects on AI reliance and performance remain stable across the 50 rounds, indicating a consistent change in behavior throughout the task.

These patterns are consistent with the two main model predictions in Section 3, which formalizes a worker's decision in a categorization task, including the option to revise an initial answer after considering an AI recommendation. The model predicts that when AI reliance becomes visible and workers expect a penalty for relying on the AI, visibility reduces reliance and leads to lower accuracy relative to the control condition.

A reduction in performance is not guaranteed following the decrease in AI reliance due to its visibility, as several orthogonal channels not captured by the model may have sustained or even improved accuracy. One plausible scenario is that workers exerted greater effort in their initial choice in anticipation of relying less on AI.[21] Column 5 of Table 2 reports a 10% increase in response time, a natural proxy for effort (significant at the 10% level). On average, treated workers took about two additional seconds per image to make their initial selection. However, column 3 shows that these extra seconds did not have an impact on performance, as initial accuracy was unaffected. This makes it unlikely that changes in initial-choice behavior explain the performance decline.

Another potential mechanism is that treated workers became more discerning in their AI use, forgoing incorrect recommendations more often. However, the evidence does not support this interpretation. First, Column 6 shows a negative but statistically insignificant effect on response times at the recommendation stage, making it unlikely that treated workers engaged in more careful deliberation. The model's no-updating assumption implies that workers continue to evaluate adopted AI recommendations using the same fixed accuracy assessment. Consistent with this benchmark, realized accuracy is unchanged when workers follow the recommendation and lower when they do not, as shown in Figure A.4.

[21] Almog, Gauriot, et al. (2026) provide field evidence that public overruling of human decisions by AI can trigger effort adjustments, while Agarwal, Alexander Moehring, and Wolitzky (2025) show how such effort responses can be incorporated into the design of human–AI collaboration systems.

I also conducted a counterfactual exercise to test whether the induced reduction in AI reliance reflected a deliberate change in engagement rather than merely a mechanical decrease. Specifically, I simulated a scenario in which, for every round where control workers followed the AI recommendation, I randomly forced 14% of those choices to be overridden, thereby reducing their AI reliance to the rate observed in the treatment group. Figure A.5 shows that, under this adjustment, the average accuracy gains for control workers declined from 23.8 pp to 20.4 pp, virtually identical to the gains among treated workers (20.3 pp).

Taken together, these findings show that the decline in AI reliance when it was visible did not reflect improved discretion in following AI recommendations. Instead, it was comparable to randomly overriding additional AI recommendations, providing clear evidence that workers in the treatment group did not improve the quality of their decisions about when to rely on AI.

A recurring question in workplace AI adoption is whether its effects differ across groups in ways that widen or narrow existing inequalities (Cruces et al. 2026). To examine this possibility, I extend the empirical analysis to test whether workers respond differently to the observability of their AI reliance depending on gender, ethnicity, education, or prior platform experience. Figures A.6 and A.7 show broadly similar patterns across the demographic and experience groups considered, and I find no statistically significant evidence that the treatment effects differ systematically across these groups.

4.2 Scope for AI-human Collaboration

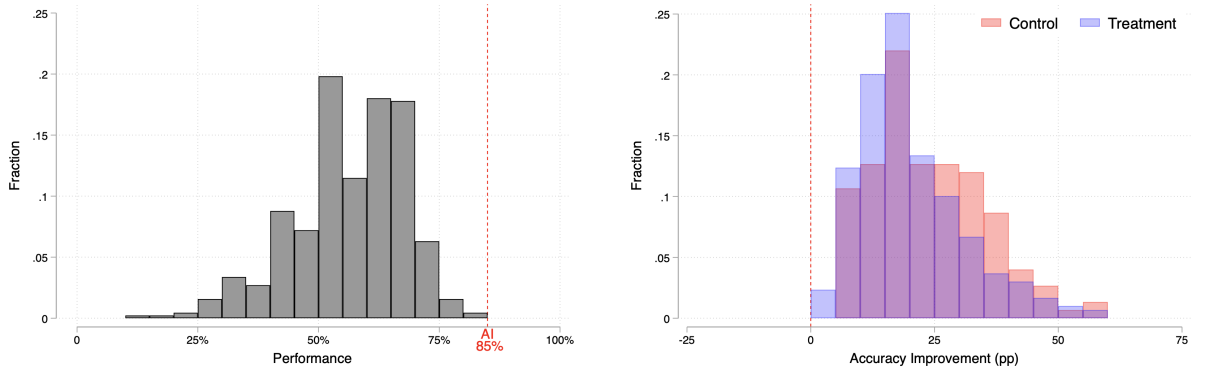
The potential for AI-human collaboration lies in what humans and AI can achieve together. Although no worker in this setting was individually more accurate than the AI (Figure 5a), this does not imply the absence of collaborative gains. Figure 5b shows the distribution of performance improvements per worker after considering the AI recommendations, highlighting two key findings: (i) every worker benefited from AI assistance, and (ii) the distribution of performance gains is more skewed to the right in the control group than in the treatment group, consistent with the larger improvements when AI reliance was private.

A natural benchmark for successful collaboration is whether workers, with AI assistance, surpass the AI's standalone accuracy of 85%. Figure A.8 shows that the share of workers meeting this benchmark falls from 24.7% to 18.5% when AI reliance becomes visible to the HR specialist. Although this 6.2 percentage-point difference is imprecisely estimated ($p = 0.129$), the point estimate corresponds to a 25% reduction in the share of workers meeting the benchmark.

A meta-analysis by Vaccaro, Almaatouq, and Malone (2024) shows that AI-human collaboration is especially likely to underperform in two kinds of settings: decision-making tasks

(rather than content creation) and environments in which AI already outperforms humans working alone. In this sense, this setting presents a particularly common but difficult test for successful collaboration. Consistent with that broader pattern, workers in the data struggle to reap gains from collaborating relative to the AI alone. The results suggest that image concerns may be one important friction behind this prevalent difficulty in successful AI-human collaboration.

Figure 5: Accuracy Distributions



(a) Initial Accuracy (before AI). (b) Accuracy Gains from AI Recommendations.

Notes: Panel (a) shows workers' accuracy before receiving AI recommendations, with the red dashed vertical line indicating the AI's accuracy. Panel (b) displays the distribution of accuracy improvements from incorporating AI recommendations relative to workers' initial answers, distinguishing control workers (red) from treatment workers (blue).

4.3 What Workers Fear AI Reliance Communicates

Understanding why workers view AI reliance as damaging to their image is crucial, since addressing this perception is key to recouping the productivity losses it generates. To investigate the underlying mechanism, I developed a novel incentive-compatible method that leverages a core feature of Upwork, *public feedback*, which is also common across many online platforms.

After completing the categorization task, all workers were informed that they would receive feedback intended to be informative for future employers on the platform. The feedback described the AI-assisted categorization task they had performed, ensuring that it provided sufficient context on its own. Workers were then given the opportunity to choose among three positive but mutually exclusive traits to emphasize in their feedback. Specifically, they could select a statement highlighting that they are hard workers, highly skilled at this type

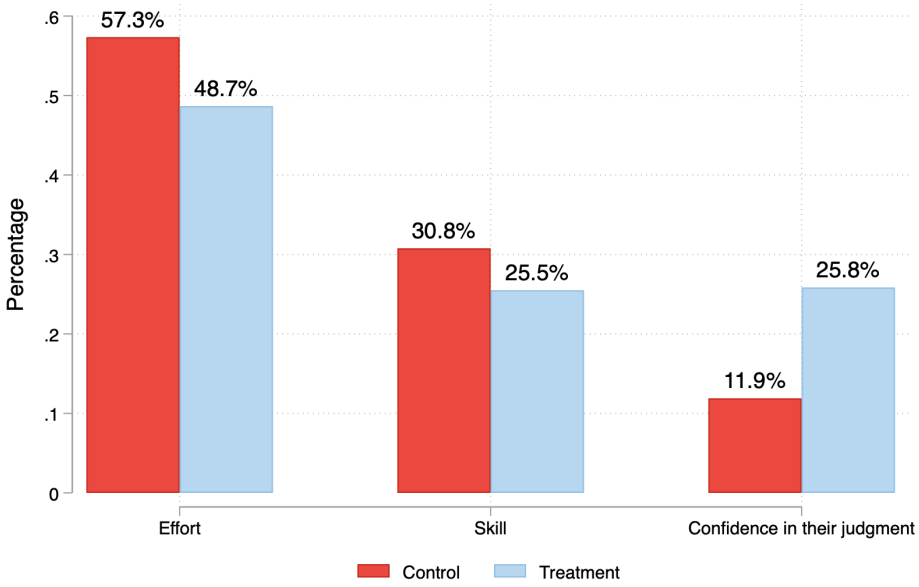
of task, or confident in their own judgment.[22] Here, confidence in one's own judgment refers to trust in the answer a worker formed before seeing the AI recommendation: a willingness to stand by that initial assessment rather than second-guess it or defer to the AI. For workers in the control group, their choice of statement reflects baseline preferences over these traits. By contrast, for workers in the treatment group, the choice reflects which positive signal they most want to reinforce when their AI reliance may be disclosed. Comparing the two distributions therefore shows which positive signal workers most seek to restore when reliance may be disclosed.

Figure 6 presents the distribution of preferred statements by treatment status. In the control group, the majority of workers (57.3%) preferred to emphasize effort, followed by skill (30.8%), with confidence in their own judgment ranking last (11.9%). In the treatment groups, although signaling effort remained the most common choice (48.7%), preferences shifted away from both effort and skill toward signaling confidence in their own judgment (25.8%). The share selecting confidence in their own judgment increased by 13.9 percentage points, equivalent to a 117% increase relative to the control mean (two-sided test of proportions, $p < 0.001$). These results provide two main insights. First, many workers consistently value being perceived as hard-working, regardless of whether AI reliance is disclosed, which is reasonable given that monitoring remote workers is challenging and effort tends to extrapolate well to other tasks. Second, signaling confidence in their own judgment, which is relatively unpopular when AI reliance is not disclosed in the feedback, becomes much more relevant once workers worry that heavy AI reliance may be visible to future employers.

To the best of my knowledge, this is the first use of public feedback to elicit preferences in an incentive-compatible way. To validate these findings, I also included two more direct, though non-incentivized, questions in the second job questionnaire. I asked 283 workers who completed the follow-up questionnaire to report, using a slider ranging from 1 (not important at all) to 10 (extremely important), how important it is to signal each of the same three positive traits to employers on Upwork. The question was asked twice: first in a general setting, and then specifically for tasks involving the use of AI. Figure 7 presents the results. The first question, situated in a general setting, replicates the ranking observed in the control group: effort first, followed by skill, and lastly confidence in judgment. When the second question referred to tasks involving AI, the ranking reversed, consistent with the shift we observe in treated workers' feedback preferences toward signaling confidence in their judgment. These questionnaire results are reassuring, suggesting that the incentive-compatible elicitation method was well understood by the workers and reflective of underlying

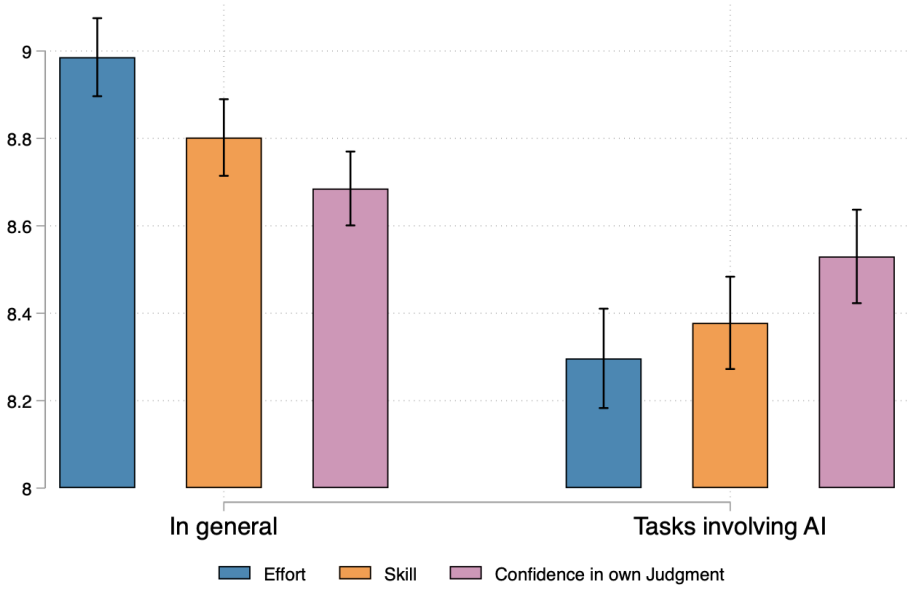
[22]These three traits emerged as the most common in a prior survey and are also prominent in public discourse. Respondents who used the open-ended option typically referred to one or more of these traits, often in overlapping ways. To avoid ambiguity, I restricted the choice to only these three, which ensured clarity and comparability across responses.

Figure 6: Feedback Preference



preferences for these signals.

Figure 7: Self-reported Importance of Signals.



5 Robustness of the Observability Effect and the Limits of Mitigation

This section examines the persistence of the observability effect and the limits of a simple mitigation strategy. I begin by highlighting two features of the evaluation environment that were deliberately chosen to restrict the scope for image concerns. First, although workplace evaluations often reflect attributes beyond contemporaneous output, evaluators in the experiment were instructed to focus exclusively on expected accuracy. Second, workers were evaluated based on their expected performance in another session of the same AI-assisted task, limiting the scope for signaling traits that might be valuable because they extrapolate to other tasks or domains. The fact that observability nevertheless reduced AI reliance and accuracy shows that the main effects persist in an unusually narrow and well-specified evaluation environment. Image concerns may have greater scope when tasks or evaluation criteria are less clearly defined, although their magnitude in such settings remains an open empirical question. I then test whether informing workers that evaluators had been reassured about their general quality attenuates the behavioral response and examine whether workers themselves penalize AI reliance when evaluating others. Together, the results show that generic reassurance does not mitigate the observability effect and that the anticipated penalty is consistent with an evaluation norm that workers themselves apply.

5.1 Limits of Generic Reassurance

Earlier in the paper, I justified pooling the two treatment conditions in which AI reliance was public, as the information intervention had no effect on the main outcomes: response times, accuracy, and AI reliance.[23]

The intervention was motivated by the possibility that AI reliance might reveal something about a worker's type in a context where the HR specialist had very limited information about the workers. In recruitment, HR specialists are often assumed to weigh traits beyond contemporaneous output, particularly in remote-work settings where direct supervision is limited. Attributes such as reliability, integrity, work ethic, accountability, communication, and punctuality are central when direct supervision is limited, and these may plausibly correlate with reliance on AI.

To test whether image concerns were to some degree driven by information gaps, workers in the *Public with Evaluator Reassurance* condition were told that the HR specialist

[23]Table A.1 reports p-values from Wald tests of equality between the *Public* and *Public with Evaluator Reassurance* treatments. All p-values are high, providing no evidence of distinct effects across any of the outcome variables of interest.

had already been assured of their strong track record on the platform and that there was no reason to doubt their quality. As noted in the pre-analysis plan, the expectation was that this information intervention would partially mitigate the treatment effects resulting from making AI reliance visible, leading to levels of AI reliance and accuracy that would fall between those of the control and the standard public-treatment groups. Instead, their behavior was indistinguishable from the latter, indicating that informing the workers about the additional reassurance did not mitigate image concerns. The intervention can also be viewed as approximating one informational feature of a longer employment relationship, in which repeated interactions give evaluators more information about workers' general quality. Its null effect suggests that this type of general reassurance alone is insufficient to attenuate the behavioral response to making AI reliance observable.

Several workers offered a possible explanation for the null effect of the information intervention: they interpreted the reassurance as raising the HR specialist's expectations and sought to meet those expectations by relying less on AI.

As an additional check, I examine whether workers read the instructions carefully. Figure A.9 shows that average reading times for the screen containing the treatment variation were longer in the treatment conditions, and longest in the *Public with Evaluator Reassurance* condition, which included an additional paragraph. Reading times for the other instruction screens did not differ significantly across conditions.

5.2 Workers Themselves Penalize AI Reliance

The reduction in AI reliance reflects workers' beliefs about the HR specialist's evaluation criteria. Specifically, workers anticipate that the HR specialist will assign lower scores to those who rely more heavily on AI. One possible explanation is that workers project their own beliefs onto the HR specialist; another is that workers' beliefs about the HR specialist's actions diverge from what they would themselves do. The latter possibility opens the door for a misperception equilibrium à la Bursztyn, González, and Yanagizawa-Drott (2020), who present evidence that simple information interventions can be effective in addressing such distortions. In what follows, I present evidence consistent with the idea that workers were projecting how they themselves would evaluate others when choosing to rely less on AI when visible.

To better understand where workers' beliefs about HR evaluation come from, I asked returning workers from the treatment groups to score 20 real profiles of other workers who were also selected for the second job.[24] Each profile displayed the worker's accuracy rate

[24]The 20 profiles combined accuracy levels of 60, 70, 80, and 90 with AI use ranging from 10 to 60 in increments of 10 (not all combinations were feasible).

and AI usage from the first job, and participants were monetarily incentivized to score them based on the number of correct answers those 20 workers achieved in the second job.

I find robust evidence that workers penalize AI reliance when assuming the role of an HR specialist and evaluating others' profiles. To illustrate this, I examine how accuracy and AI reliance affect the scores and rankings assigned to profiles. Table 3 reports regression results where the dependent variable is either the score (ranging from 0 to 100) or the ranking of a profile (1 = best, 20 = worst), modeled as a function of the worker profile's accuracy and their AI reliance. The specification includes evaluator fixed effects, and standard errors are clustered at the evaluator level. A one–percentage point increase in accuracy raises the score by about one point, whereas each additional percentage point of AI reliance reduces the score by 0.36 points. Both effects are statistically significant at the 1% level. Because the accuracy coefficient is close to 1, the interpretation is straightforward: The weight placed on AI reliance is roughly 36% of that placed on accuracy, but with the opposite sign. Put differently, evaluators penalize the adoption of three additional AI recommendations more than a single incorrect answer. The results for rankings closely mirror those for scores. Figure 8 shows the 20 profiles sorted by average ranking. Although accuracy is the main driver, the figure illustrates that profiles with 10 percentage points lower accuracy can still achieve better rankings if they rely less on AI.[25]

Table 3: Effect of Worker Accuracy and AI Reliance on Evaluations.

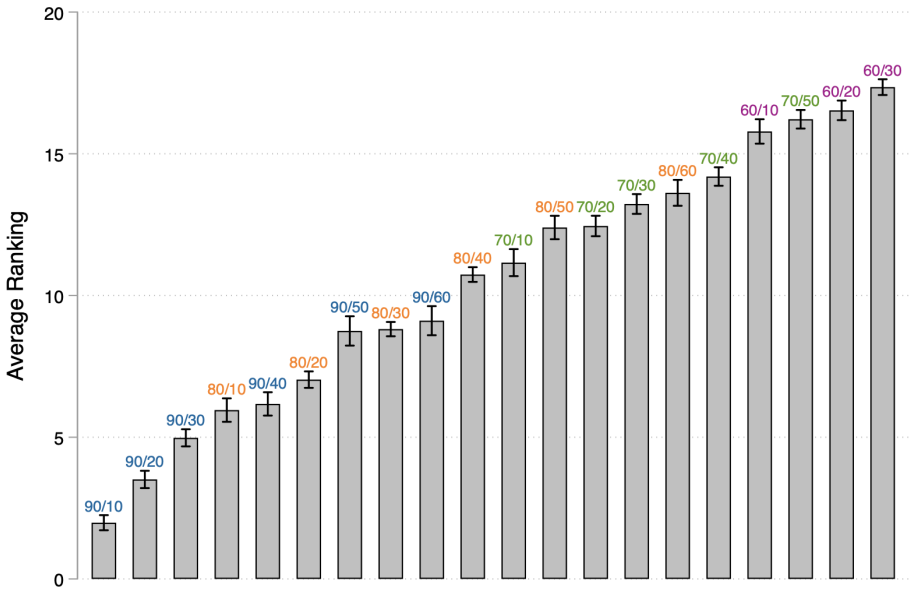
	Score (1)	Ranking (2)
Accuracy (pp)	**1.00*****	**-0.43*****
	(0.044)	(0.015)
AI reliance (pp)	**-0.36*****	**0.142*****
	(0.042)	(0.013)
Evaluator FE	X	X
Constant	-4.35	39.4
Observations	1,860	1,860

Notes: Each dependent variable is regressed on the evaluated worker's accuracy and AI reliance rate, with evaluator fixed effects included and standard errors clustered at the evaluator level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

There is also strong evidence at the individual level that evaluators penalize AI reliance. Following a similar empirical strategy as in the aggregate analysis, I regress each evaluator's

[25] Figure A.11 provides a two-dimensional, intuitive depiction of the average rankings.

Figure 8: Accuracy/AI Reliance Pairs Ordered by Average Ranking.



Notes: Each bar shows the average ranking of a profile with standard errors. Labels display the profile pair (Accuracy/AI reliance) and are color–coded according to accuracy.

assigned scores on accuracy and AI reliance separately for 93 evaluators. Of these, 70 had a coefficient on AI reliance that was statistically significant at the 10% level. Among the 70 evaluators whose estimated coefficients were significant at the 10 percent level, 69 coefficients were negative. Figure A.10 shows the distribution of these statistically significant coefficients associated with AI reliance.

Existing evidence shows that human evaluators penalize work they perceive to have been produced with AI (Awuah, Krenk, and Yanagizawa-Drott 2026; Wang and Qiu 2025), but the results in this paper are especially striking because these workers had themselves already experienced what it feels like to be judged for using AI. Prior research finds that perspective-taking interventions can improve attitudes toward other groups (Alan et al. 2021), with recent work highlighting relatability as an important mechanism (Andries et al. 2026). Even so, directly experiencing the worker role was not enough to eliminate workers' inclination to penalize AI reliance in others.

6 Conclusion

Digital technologies affect organizations not only through the actions they enable, but also through the records they create about how work is done. This paper shows that when those

records enter worker evaluation, they can change behavior in ways that offset some of the technology's productive value. In a field experiment on an online labor market, making workers' reliance on AI recommendations visible to an evaluator reduced reliance by 4.3 percentage points and final accuracy by 2.7 percentage points. This occurred even though the evaluator observed accuracy and was instructed to predict performance in another session of the same AI-assisted task. Treated workers spent more time on their initial choices, but were no more accurate before seeing the recommendation, and their additional overrides were no better targeted. The experiment therefore identifies a response to observability and evaluation, rather than a general reluctance to use AI.

The mechanism evidence points especially to one image concern: that heavy reliance may be read as low confidence in one's own initial judgment. When the possible disclosure of AI reliance in public feedback was made salient, workers became substantially more likely to emphasize this trait. Returning workers also penalized reliance when evaluating peers, despite having incentives tied to those peers' subsequent performance. By contrast, telling workers that the evaluator knew about their strong platform records did not change behavior. Favorable information about overall worker quality was therefore not enough to remove the concern, and workers applied a similar penalty when they took the evaluator's role.

The results speak most directly to the design of evaluation systems around recommendation-based AI. They do not imply that firms should hide process data or seek to maximize AI reliance. Firms may still need these data for operational or safety reasons. What matters is whether the same data are also used to judge workers. Any informational gain from doing so has to be weighed against the behavioral distortion documented here. Where output is observable, treating a raw reliance rate as a general measure of worker quality can induce workers to reject useful recommendations. The relevant benchmark is not a low reliance rate, but whether reliance is appropriate for the task and the quality of the recommendation. Organizations might therefore keep operational records separate from personnel evaluation or assess reliance against task-specific standards. The experiment does not test these alternatives. It does show that reassurance about workers' general quality was not enough, suggesting that credibly changing the expected evaluation rules may matter more than providing favorable background information.

The unresolved question is when reliance contains useful information about subsequent performance and when making it visible mainly induces workers to manage impressions. That distinction determines whether process transparency improves worker selection or simply lowers output. As digital technologies make production methods increasingly traceable, their productivity effects will depend not only on technical quality, but also on how workers expect their technology use to shape others' perceptions of them and, through those perceptions, their future opportunities.

References

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz (2023). "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology". *Working paper 31422, National Bureau of Economic Research*.

Agarwal, Nikhil, Alexander Moehring, and Alex Wolitzky (2025). "Designing Human-AI Collaboration: A Sufficient-Statistic Approach". *NBER Working Paper 33949*.

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay (2021). "Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking". *The Quarterly Journal of Economics* 136.4, pp. 2147–2194.

Albright, Alex (2024). "The Hidden Effects of Algorithmic Recommendations". *Working paper*.

Almog, David (2025). "AI Recommendations and Non-instrumental Image Concerns". *SSRN working paper 5232232*.

Almog, David and Ari Bronsoler (2025). "Texting to Save Lives: Evidence from Cardiovascular Treatment Reform in Mexico". *SSRN working paper 5123130*.

Almog, David, Romain Gauriot, Lionel Page, and Daniel Martin (2026). "Human Responses to AI Oversight: Evidence from Centre Court". *Working paper*.

Andries, Marianne, Leonardo Bursztyn, Thomas Chaney, Milena Djourelova, and Alex Imas (2026). "In their Shoes: Empathy through Information". *The Quarterly Journal of Economics*.

Angelova, Victoria, Will Dobbie, and Crystal S. Yang (2024). "Algorithmic Recommendations When the Stakes Are High: Evidence from Judicial Elections". *AEA Papers and Proceedings* 114, pp. 633–637.

— (2026). "Algorithmic Recommendations and Human Discretion". *The Review of Economic Studies* 93.4, pp. 2250–2283.

Ash, Elliott, Yichuan Lou, and Lena Song (2026). "Sharing Opinions in the Shadow of AI". *SSRN Working paper 6432959*.

Atkin, David, Azam Chaudhry, Shamyla Chaudry, Amit Khandelwal, and Eric Verhoogen (2017). "Organizational Barriers to Technology Adoption: Evidence from Soccer-Ball Producers in Pakistan". *The Quarterly Journal of Economics* 132.3, pp. 1101–1164.

Awuah, Kobbina, Urša Krenk, and David Yanagizawa-Drott (2026). "Augment or Automate? An Early Field Experiment with Generative AI in Hiring". *Working paper*.

Balakrishnan, Maya, Kris Johnson Ferreira, and Jordan Tong (2025). "Human-Algorithm Collaboration with Private Information: Naïve Advice-Weighting Behavior and Mitigation". *Management Science* 72.1.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2005). "Social preferences and the response to incentives: Evidence from personnel data". *The Quarterly Journal of Economics* 120.3, pp. 917–962.

Barach, Moshe A. and John J. Horton (2021). "How Do Employers Use Compensation History? Evidence from a Field Experiment". *Journal of Labor Economics* 39.1, pp. 193–218.

Bursztyn, Leonardo, Georgy Egorov, and Robert Jensen (2019). "Cool to be Smart or Smart to be Cool? Understanding Peer Pressure in Education". *The Review of Economic Studies* 86.4, pp. 1487–1526.

Bursztyn, Leonardo, Bruno Ferman, Stefano Fiorin, Martin Kanz, and Gautam Rao (2018). "Status Goods: Experimental Evidence from Platinum Credit Cards". *The Quarterly Journal of Economics* 133.3, pp. 1561–1595.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott (2020). "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia". *American Economic Review* 110.10, pp. 2997–3029.

Bursztyn, Leonardo and Robert Jensen (2017). "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure". *Annual Review of Economics* 9.1, pp. 131–153.

Caplin, Andrew, David Deming, Shangwen Li, Daniel Martin, Philip Marx, Ben Weidmann, and Kadachi Jiada Ye (2025). "The ABCs of Who Benefits from Working with AI: Ability, Beliefs, and Calibration". *Management Science*.

Celhay, Pablo, Bruce D. Meyer, and Nikolas Mittag (2025). "Stigma in Welfare Programs". *The Review of Economics and Statistics*, pp. 1–37.

Chandrasekhar, Arun G., Benjamin Golub, and He Yang (2018). "Signaling, Shame, and Silence in Social Learning". *NBER Working paper 25169*.

Coffman, Katherine B., Manuela R. Collis, and Leena Kulkarni (2024). "Whether to Apply". *Management Science* 70.7, pp. 4649–4669.

Cowgill, Bo, Pablo Hernández-Lagos, and Nataliya Wright (2026). "Does AI Cheapen Talk? Theory and Evidence from Global Entrepreneurship and Hiring". *Management Science*.

Cruces, Guillermo, Diego Fernández Meijide, Sebastian Galiani, Ramiro H. Gálvez, and María Lombardi (2026). "Does Generative AI Narrow Education-Based Productivity Gaps? Evidence from a Randomized Experiment". *NBER Working paper 34851*.

Cui, Jingyi, Gabriel Dias, and Justin Ye (2025). "Signaling in the Age of AI: Evidence from Cover Letters". *arXiv preprint arXiv:2509.25054*.

Cullen, Zoë, Danielle Li, and Shengwu Li (2026). "Labor as Capital: AI and the Ownership of Expertise". *Harvard Business School Working Paper 26-063*.

Dai, Tinglong and Shubhranshu Singh (2020). "Conspicuous by Its Absence: Diagnostic Expert Testing Under Uncertainty". *Marketing Science* 39.3, pp. 540–563.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao (2017). "Voting to Tell Others". *The Review of Economic Studies* 84.1, pp. 143–181.

Dewatripont, Mathias, Ian Jewitt, and Jean Tirole (1999). "The Economics of Career Concerns, Part I: Comparing Information Structures". *The Review of Economic Studies* 66.1, pp. 183–198.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey (2015). "Algorithm aversion: people erroneously avoid algorithms after seeing them err". *Journal of experimental psychology: General* 144.1, pp. 114–126.

Dreyfuss, Bnaya and Raphaël Raux (2024). "Human Learning about AI". *Working paper*.

Ellingsen, Tore and Magnus Johannesson (2008). "Pride and Prejudice: The Human Side of Incentive Theory". *American Economic Review* 98.3, pp. 990–1008.

Friedrichsen, Jana, Tobias König, and Renke Schmacker (2018). "Social image concerns and welfare take-up". *Journal of Public Economics* 168, pp. 174–192.

Galdin, Anais and Jesse Silbert (2025). "Making Talk Cheap: Generative AI and Labor Market Signaling". *Working paper*.

Gans, Joshua (2024). "How Will Generative AI Impact Communication?" *Economics Letters* 242, p. 111872.

Goldberg, Samuel and H. Tai Lam (2025). "Generative AI in Equilibrium: Evidence from a Creative Goods Marketplace". *Working paper*.

Goldfarb, Avi and Catherine Tucker (2012). "Privacy and Innovation". *Innovation Policy and the Economy* 12.1, pp. 65–90.

— (2019). "Digital Economics". *Journal of Economic Literature* 57.1, pp. 3–43.

He, Joyce C., Sonia K. Kang, and Nicola Lacetera (2021). "Opt-out choice framing attenuates gender differences in the decision to compete in the laboratory and in the field". *Proceedings of the National Academy of Sciences* 118.42.

Hoffman, Mitchell, Lisa Kahn, and Danielle Li (2018). "Discretion in Hiring". *The Quarterly Journal of Economics* 133.2, pp. 765–800.

Holmström, Bengt (1999). "Managerial Incentive Problems: A Dynamic Perspective". *The Review of Economic Studies* 66.1, pp. 169–182.

Horton, John J. (2017). "The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment". *Journal of Labor Economics* 35.2, pp. 345–385.

Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011). "The online laboratory: conducting experiments in a real labor market". *Experimental Economics* 14, pp. 399–425.

Houeix, Deivy (2026). "Asymmetric Information and Digital Technology Adoption: Evidence from Senegal". *Working paper*.

Jee, Edward, Anne Karing, and Karim Naguib (2025). "Social Image and the Social Multiplier: Experimental Evidence from Community Deworming in Kenya". *Working paper*.

Karing, Anne (2024). "Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone". *The Quarterly Journal of Economics* 139.4, pp. 2083–2133.

Ling, Yier, Alex Kale, and Alex Imas (2026). "Underreporting of AI use: The role of social desirability bias". *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*.

Mas, Alexandre and Enrico Moretti (2009). "Peers at Work". *American Economic Review* 99.1, pp. 112–145.

McLaughlin, Bryce and Jann Spiess (2022). "Algorithmic Assistance with Recommendation-Dependent Preferences". *arXiv Preprint, arXiv:2208.07626*.

Pallais, Amanda (2014). "Inefficient Hiring in Entry-Level Labor Markets". *American Economic Review* 104.11, pp. 3565–3599.

Perez-Truglia, Ricardo and Guillermo Cruces (2017). "Partisan Interactions: Evidence from a Field Experiment in the United States". *Journal of Political Economy* 125.4, pp. 1208–1243.

Prat, Andrea (2005). "The Wrong Kind of Transparency". *American Economic Review* 95.3, pp. 862–877.

Reif, Jessica A., Richard P. Larrick, and Jack B. Soll (2025). "Evidence of a social evaluation penalty for using AI". *Proceedings of the National Academy of Sciences* 122.19.

Spence, Michael (1973). "Job Market Signaling". *The Quarterly Journal of Economics* 87.3, pp. 355–374.

Stanton, Christopher and Catherine Thomas (2016). "Landing the First Job: The Value of Intermediaries in Online Hiring". *The Review of Economic Studies* 83.2, pp. 810–854.

Stevenson, Megan T. and Jennifer L. Doleac (2024). "Algorithmic Risk Assessment in the Hands of Humans". *American Economic Journal: Economic Policy* 16.4, pp. 382–414.

Steyvers, Mark, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth (2022). "Bayesian modeling of human–AI complementarity". *Proceedings of the National Academy of Sciences* 119.11.

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone (2024). "When combinations of humans and AI are useful: A systematic review and meta-analysis". *Nature Human Behaviour* 8, pp. 2293–2303.

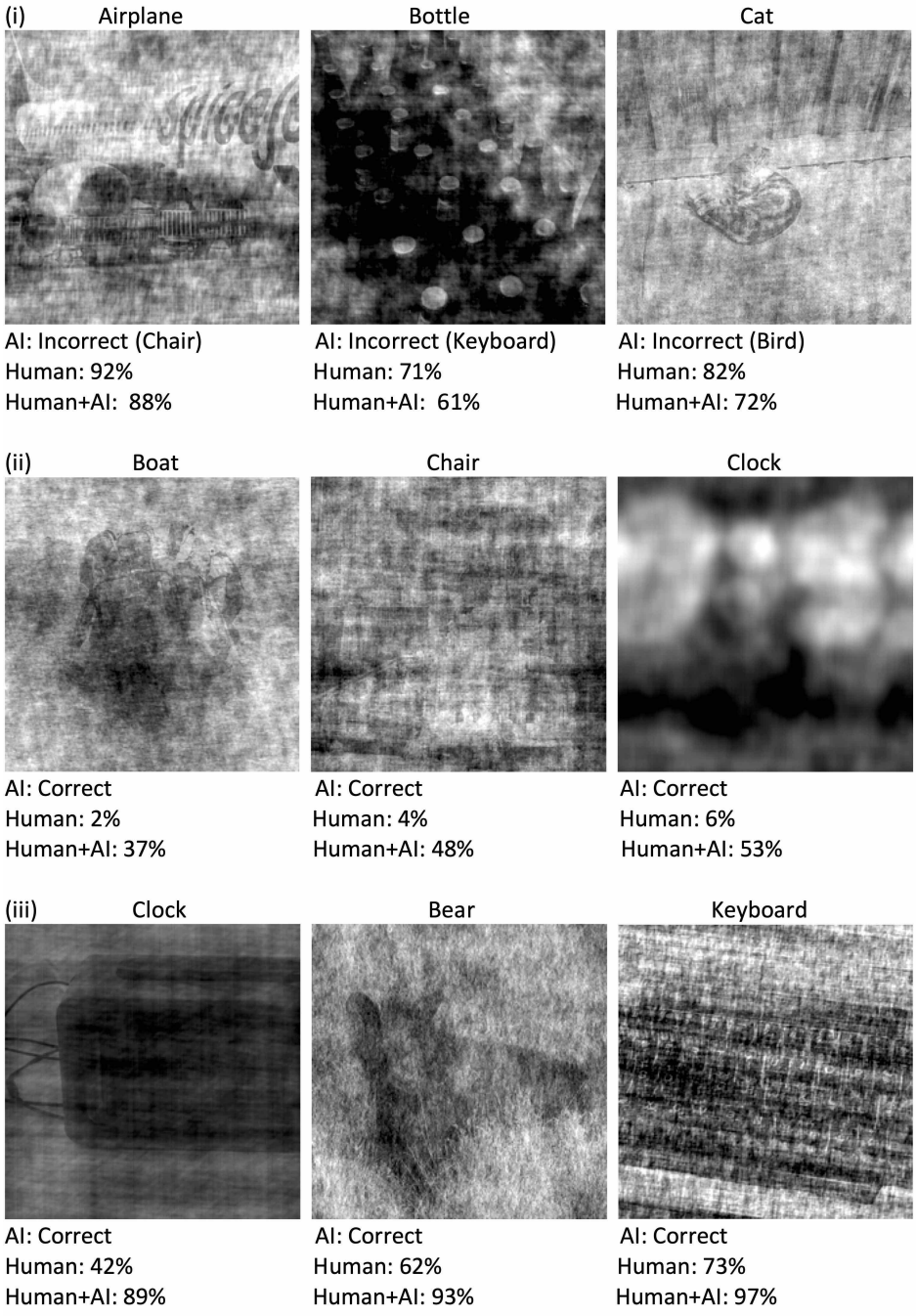
Wang, Sunny and Jingyi Qiu (2025). "Perceived AI Use and Hiring Penalties: Experimental Evidence". *SSRN working paper 5675322*.

Wiles, Emma and John J. Horton (2025). "Generative AI and Labor Market Matching Efficiency". *SSRN Working Paper 5187344*.

Yang, Haiyang, Tinglong Dai, Nestoras Mathioudakis, Amy M. Knight, Yuna Nakayasu, and Risa M. Wolf (2025). "Peer perceptions of clinicians using generative AI in medical decision-making". *npj Digital Medicine* 8.530.

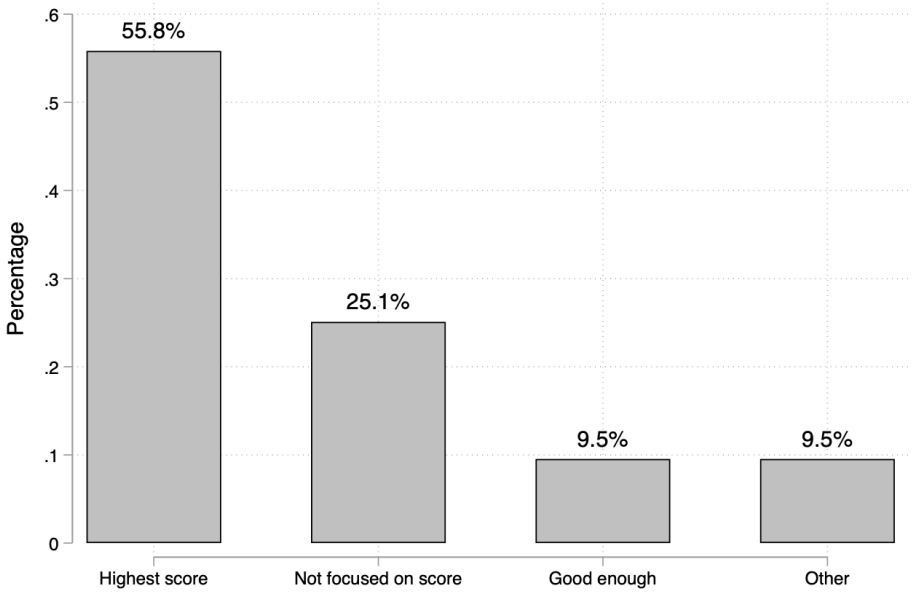
A Additional Tables and Figures

Figure A.1: Illustrative Scenarios of AI-Human Collaboration



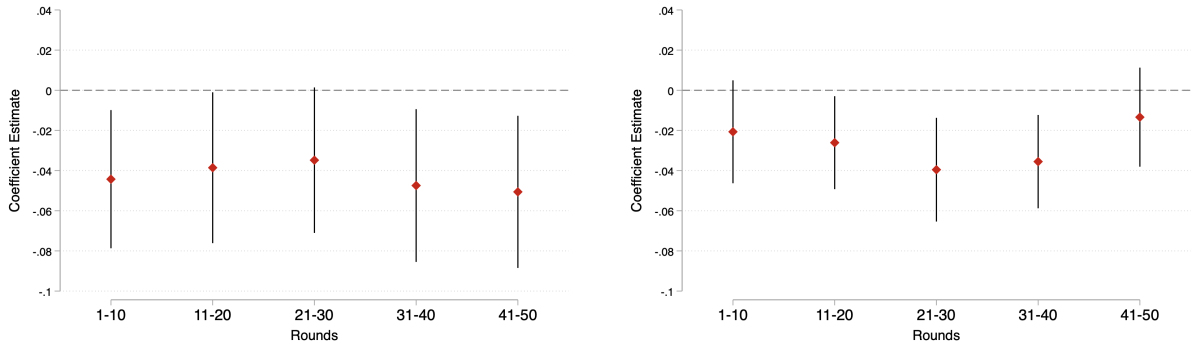
Notes: Images show the ground truth (top), the AI prediction, and humans' accuracy before and after seeing the AI recommendation. Each row illustrates a collaboration scenario: (i) Humans are mostly correct, but the AI is wrong. (ii) The AI is correct, and humans rarely are, so following it is essentially blind delegation. (iii) Humans are often correct, and many improve after reconsidering the image through the AI's suggested lens.

Figure A.2: Self-Reported First-Session Goal About HR Specialist's Score



Notes: The majority of workers who selected "not focused on score" or "other" and provided an explanation gave answers consistent with maximizing (i.e., doing the best possible).

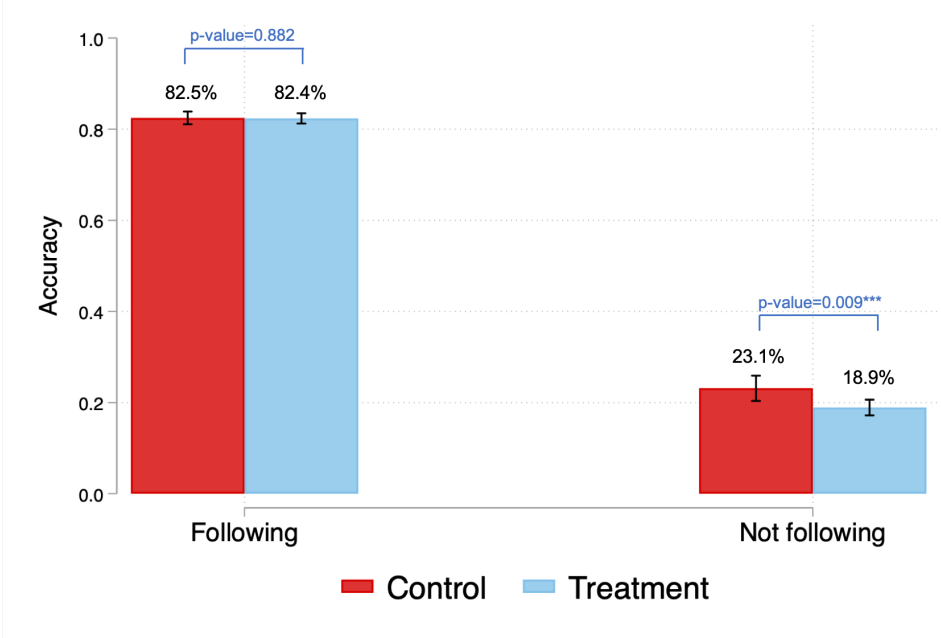
Figure A.3: Treatment Effects by 10-Round Blocks.



(a) AI reliance. (b) Accuracy.

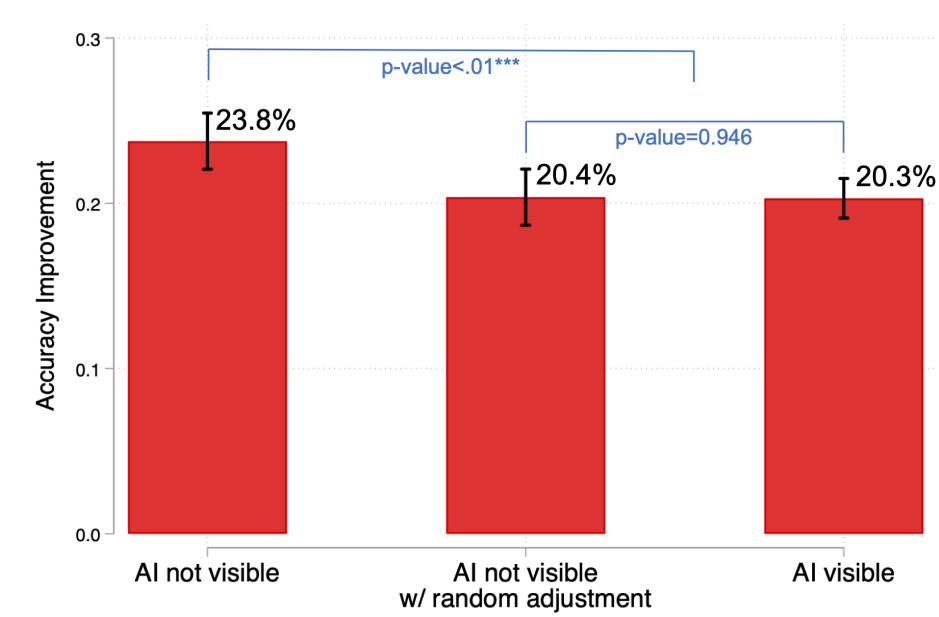
Notes: Each dot represents the coefficient from an interaction between treatment status and a 10-round block indicator. Image-specific fixed effects are included. Standard errors are clustered at the worker level.

Figure A.4: Accuracy When Following vs. Not Following AI Recommendations.



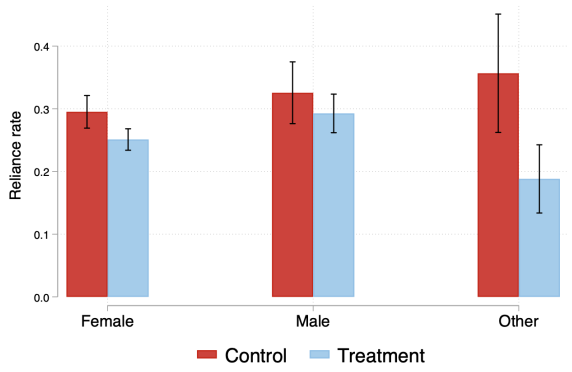
Notes: Image-specific fixed effects are included, and standard errors are clustered at the worker level.

Figure A.5: Gains from Working with AI.

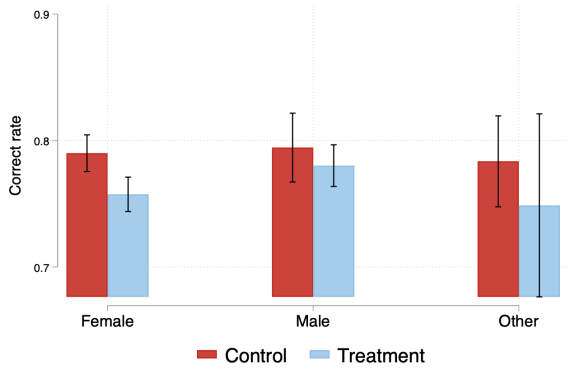


Notes: This graph plots the average accuracy gains per worker from having access to AI recommendations, measured as the improvement from workers' initial choices to their final answers after seeing the AI suggestion. The first bar reports the average gain for the control group, in which AI reliance was not visible to the HR evaluator. The second bar reports the same control group after a random adjustment that lowered their AI reliance to match the treatment group: specifically, 14% of accepted AI recommendations were randomly overruled. The final bar shows the average gain for the treatment group, where AI reliance was visible to the HR evaluator.

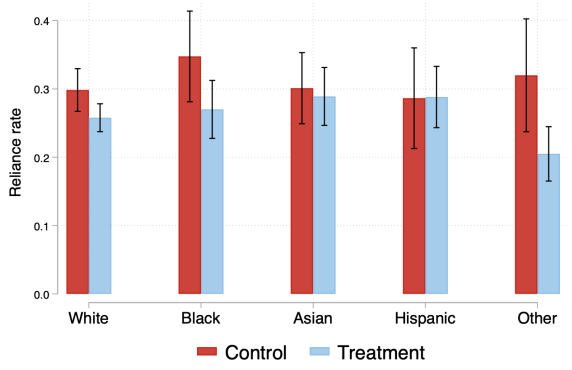
Figure A.6: Heterogeneous Treatment Effects by Demographics



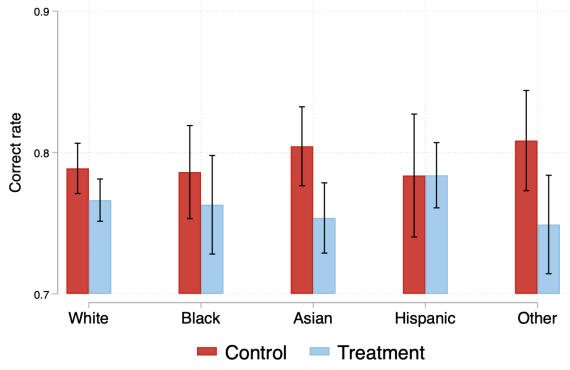
(a) Gender: AI Recommendation Reliance.



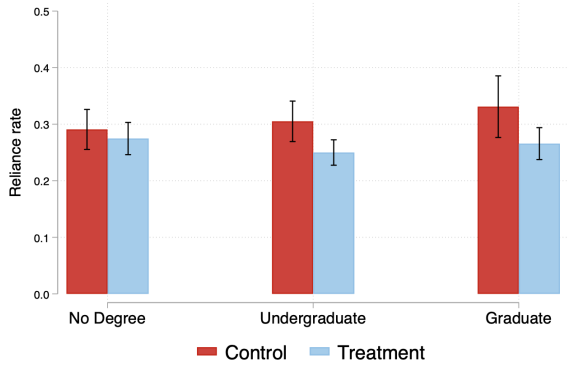
(b) Gender: Accuracy.



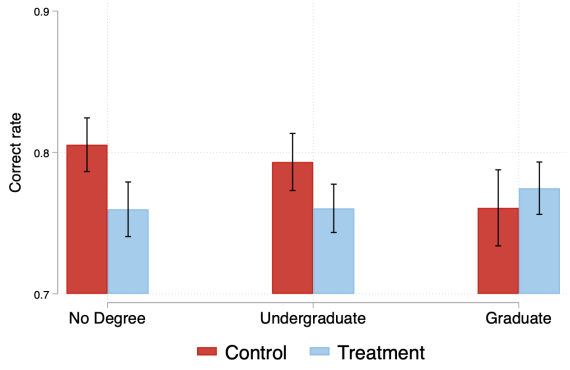
(c) Ethnicity: AI Recommendation Reliance.



(d) Ethnicity: Accuracy.



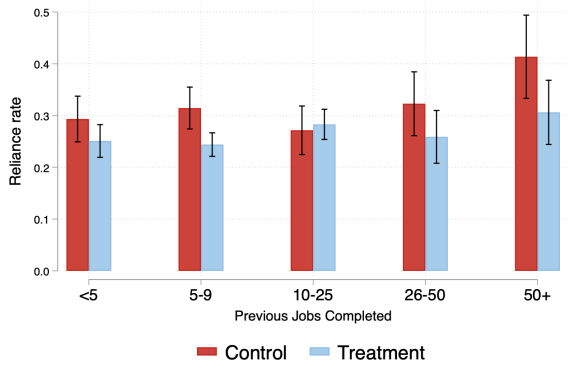
(e) Education: AI Recommendation Reliance.



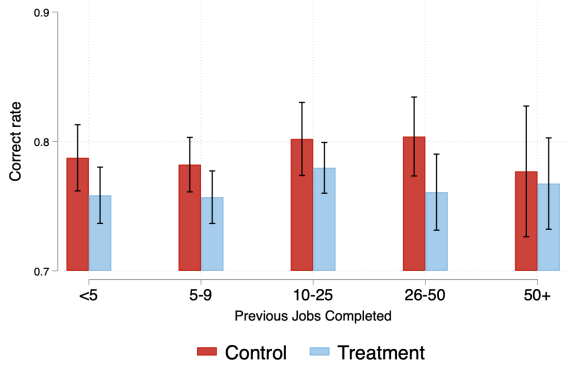
(f) Education: Accuracy.

Notes: Panels (a) and (b) present treatment effects on AI reliance and accuracy by gender, while panels (c) and (d) do the same by ethnicity. Panels (e) and (f) replicate the analysis by workers' highest level of education. Image-specific fixed effects are included, and standard errors are clustered at the worker level.

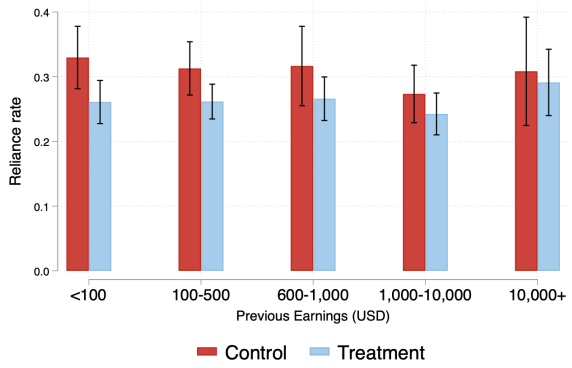
Figure A.7: Heterogeneous Treatment Effects by Platform Experience.



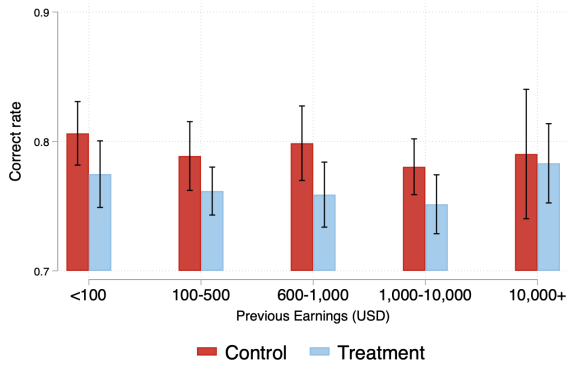
(a) Jobs: AI Recommendation Reliance.



(b) Jobs: Accuracy.



(c) Earnings: AI Recommendation Reliance.



(d) Earnings: Accuracy.

Notes: Panels (a) and (b) present treatment effects on AI reliance and accuracy by buckets of the number of previously completed jobs on the platform, while panels (c) and (d) do the same by grouping workers according to their previous earnings. The correlation between earnings and jobs is 0.55—positive but far from one—so examining both dimensions separately is not redundant. Image-specific fixed effects are included, and standard errors are clustered at the worker level.

Figure A.8: Share of Workers Outperforming AI

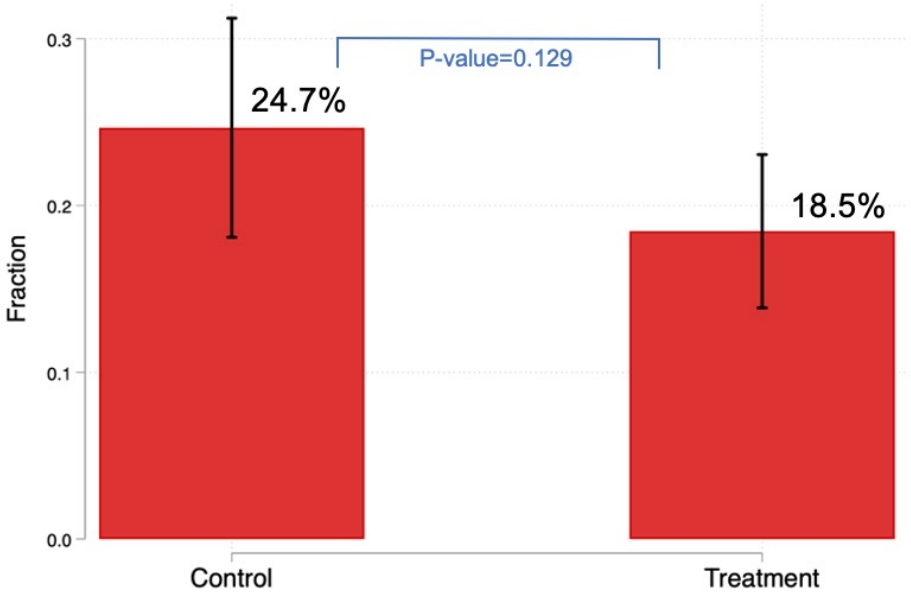
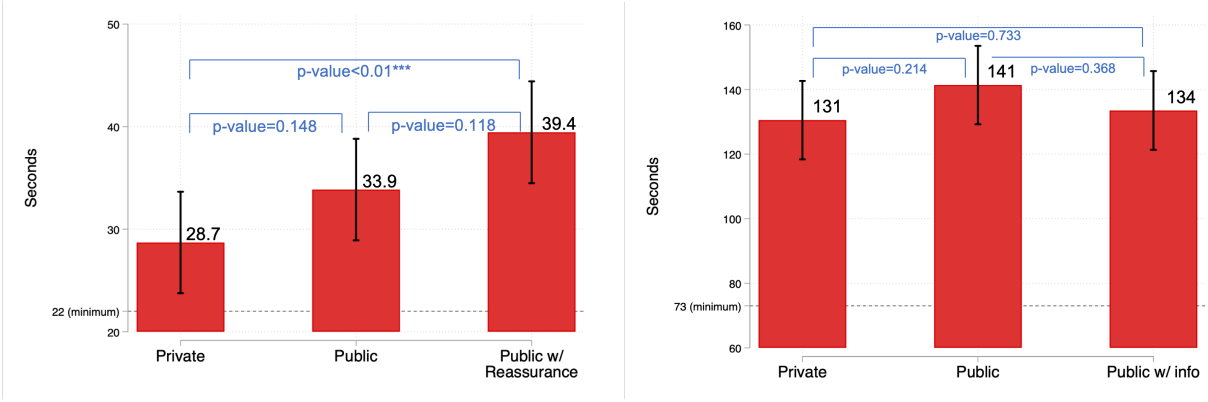


Table A.1: Treatment Effects with Separate Treatment Groups.

	AI recommendation reliance		Correct answer		Response time	
	All	Conditional	Initial	Final	Initial	Rec. stage
	(1)	(2)	(3)	(4)	(5)	(6)
Public	**-0.040****	**-0.080*****	**0.005**	**-0.028*****	**2.14**	**-0.82**
	(0.016)	(0.025)	(0.014)	(0.010)	(1.33)	(0.62)
Public w/ Reassurance	**-0.046*****	**-0.081*****	**0.012**	**-0.026****	**2.05**	**-0.61**
	(0.016)	(0.026)	(0.014)	(0.010)	(1.28)	(0.64)
Equality test (p-value)	0.67	0.97	0.57	0.88	0.95	0.71
Constant	0.305	0.640	0.553	0.791	21.3	10.1
Observations	22,398	10,554	22,398	22,398	22,398	10,554

Notes: Separating treatment groups changes the empirical specification to: $Y_{ij} = \alpha + \beta_1 T1_i + \beta_2 T2_i + \gamma X_j + \epsilon_{ij}$, where each outcome variable is regressed on indicators for treatment assignment. Specifically, $T1_i$ equals 1 if worker i was assigned to the *Public* condition, and $T2_i$ equals 1 if assigned to the *Public with Evaluator Reassurance*. All regressions include image-specific fixed effects, and standard errors are clustered at the worker level. Equality test reports the p-value from a Wald test of equal coefficients for the *Public* and *Public with Evaluator Reassurance* conditions. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.9: Average Time Spent on Instruction Screens

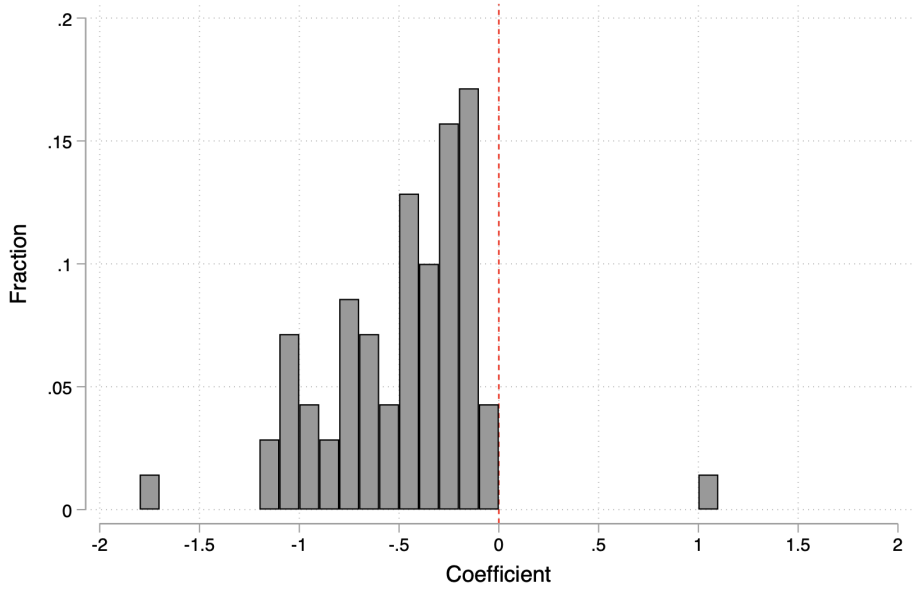


(a) Screen with Evaluation Explanation

(b) Remaining Screens.

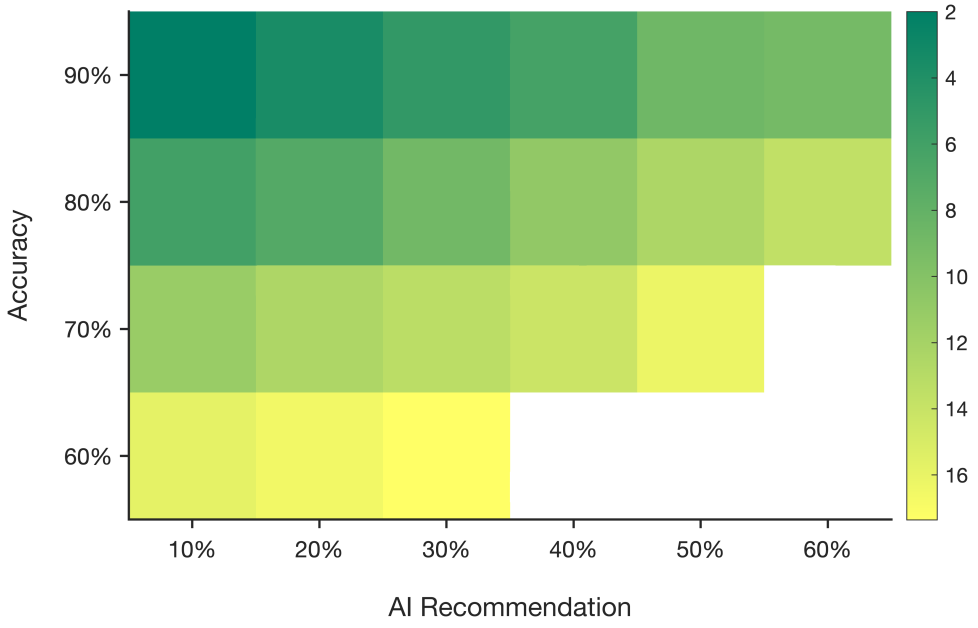
Notes: Minimum time requirements were imposed to ensure careful reading; horizontal dashed lines denote the shortest allowable time per screen. Panel (a) reports time spent on the instruction screen containing the treatment variation, while Panel (b) shows the total time across all other screens.

Figure A.10: Distribution of Significant AI Recommendation Coefficients.



Notes: We estimate individual regressions for each of the 93 returning workers who evaluated 20 worker profiles. The figure shows the distribution of the coefficients on AI reliance for the 70 evaluators with coefficients significant at the 10% level.

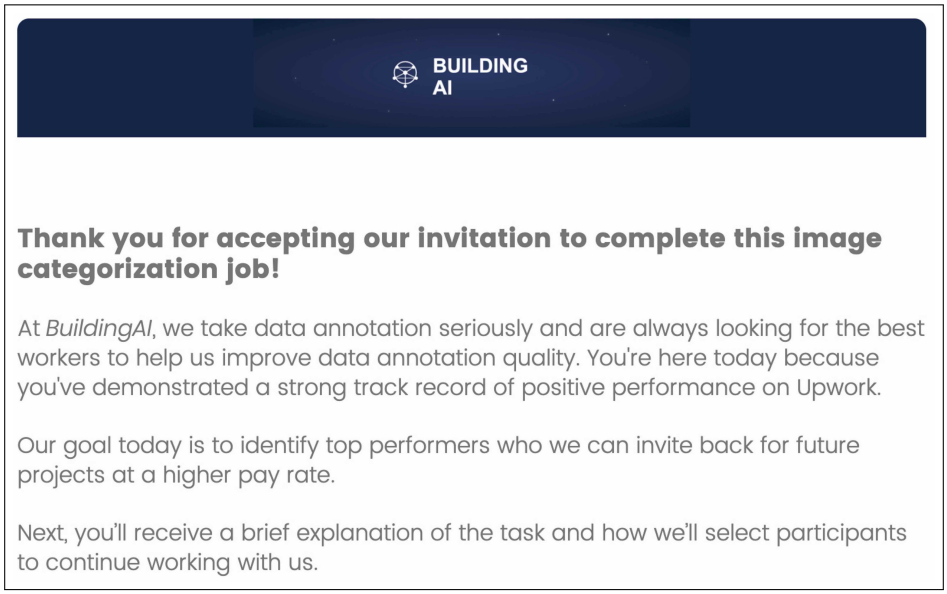
Figure A.11: 2-D Visualization of Rankings.



Supplemental Appendix

B Experiment Implementation

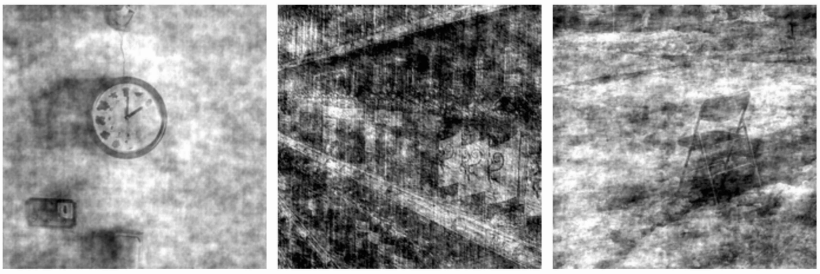
B.1 Experimental Instructions for First Job



BUILDING AI

Image Classification Task

In this job, you'll be classifying blurry images like these ones:



For each image, you job is to choose the correct category from a list of 16 options: Airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven, and truck.

Your objective is to select the category that best matches each image.
There are a total of 50 rounds. The images will change, but the task will always be the same.

BUILDING AI

You will be assisted with AI recommendations.

After making your choice, you will be presented with a prediction from an Artificial Intelligence (AI) algorithm trained on this task, which is **accurate 85%** of the time. This algorithm might perform well on images that are complicated for humans but could also miss some easy images.

If your choice is different from the AI recommendation, you will have the option to change it to the one AI recommends.

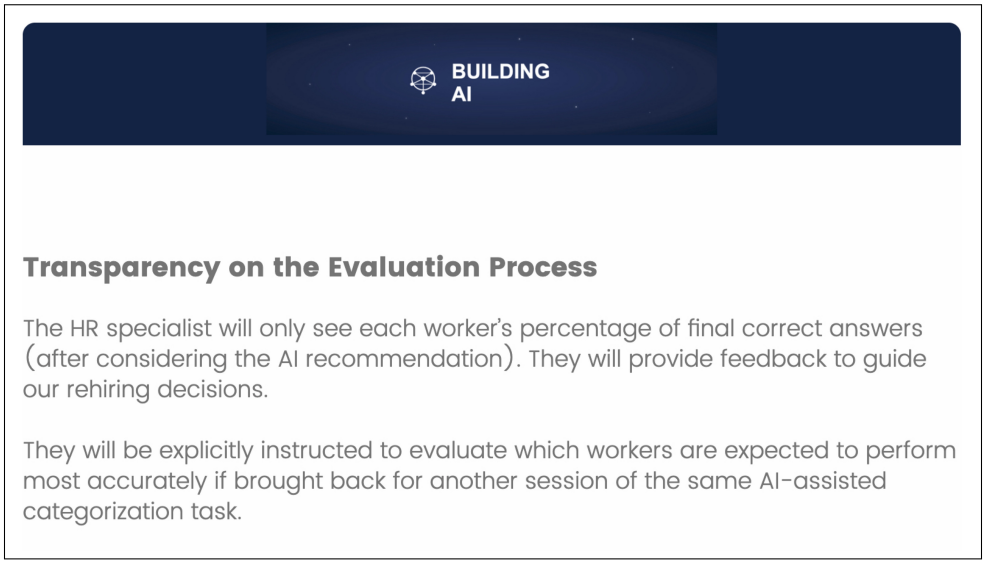


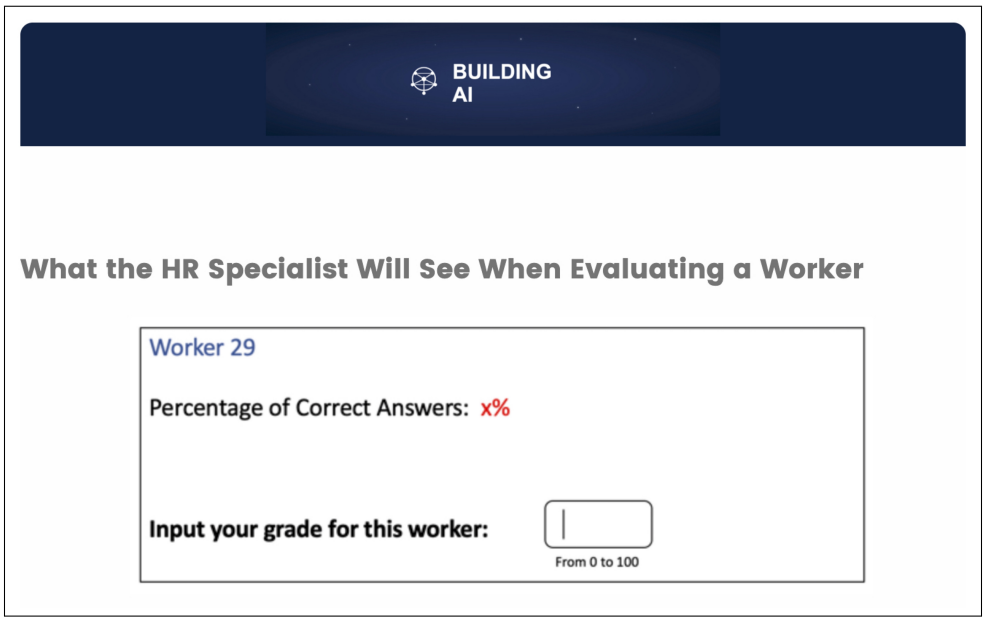
Opportunity to Continue Working with Us

Our goal today is to identify top-performing workers. We plan to invite 30% of the workers back for a similar task, next time at double the pay rate ($20 for another 30 minutes job).

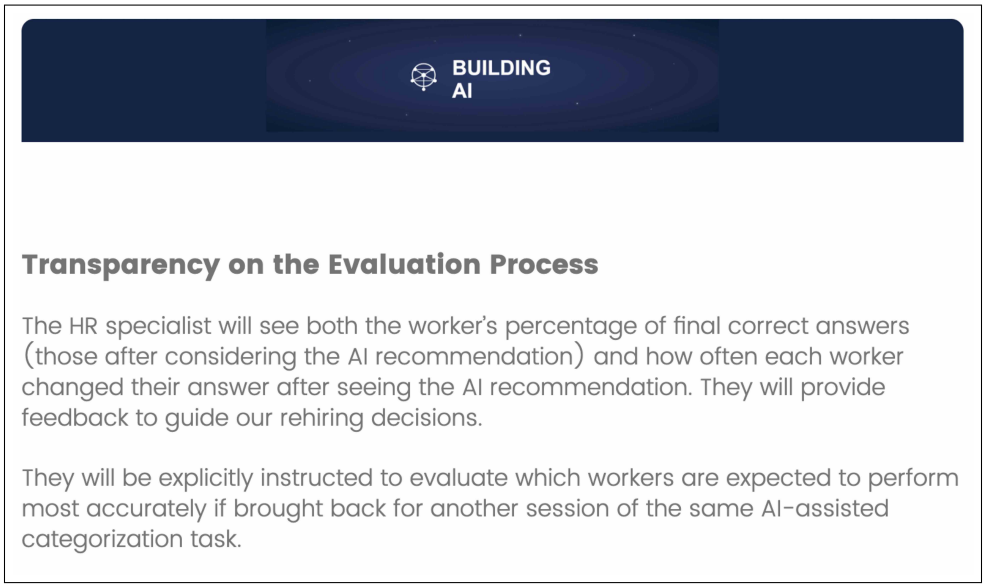
To help us evaluate fairly, an HR specialist will review anonymized participation profiles for all workers.

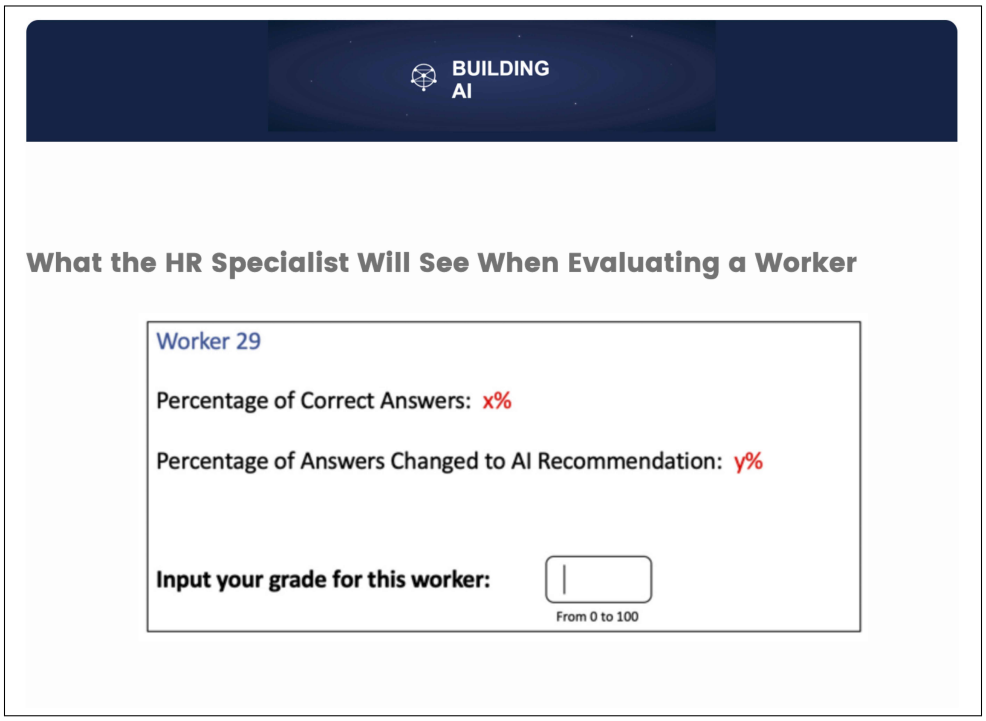
HR Evaluation Process Explained in Control





HR Evaluation Process Explained in Treatment 1





HR Evaluation Process Explained in Treatment 2

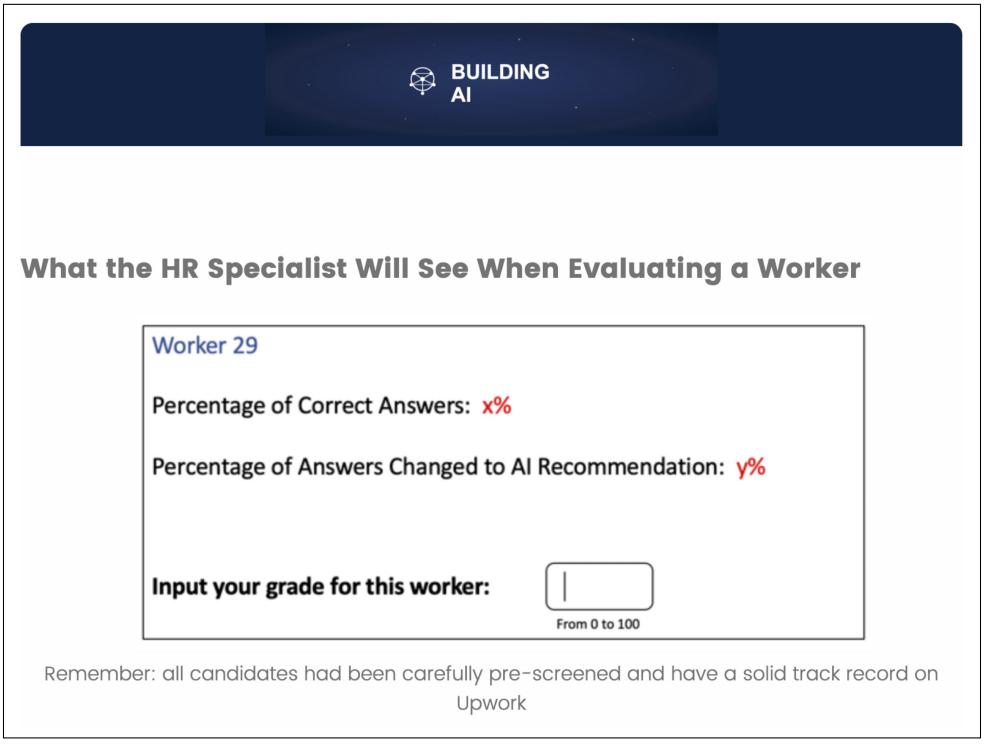
BUILDING AI

Transparency on the Evaluation Process

The HR specialist will see both the worker's percentage of final correct answers (those after considering the AI recommendation) and how often each worker changed their answer after seeing the AI recommendation. They will provide feedback to guide our rehiring decisions.

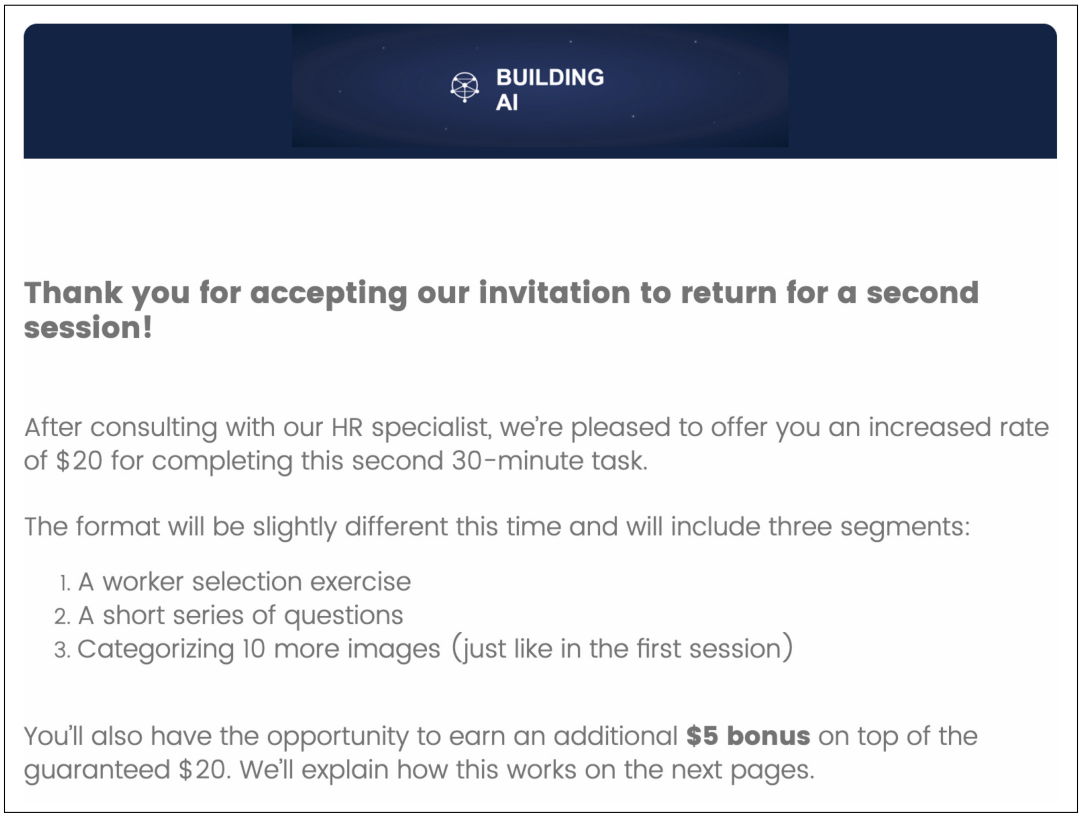
They will be explicitly instructed to evaluate which workers are expected to perform most accurately if brought back for another session of the same AI-assisted categorization task.

We also assured the HR specialist that all candidates had been carefully pre-screened and had a solid track record of positive experiences on Upwork, minimizing the risk of unnecessary concerns about worker quality.



B.2 Additional Questions in Second Job

Instructions shown are for returning treated workers (either in the *Public* or *Public with Evaluator Reassurance* conditions). Returning workers from the control group received similar instructions, except that the "Worker Selection" segment was omitted.





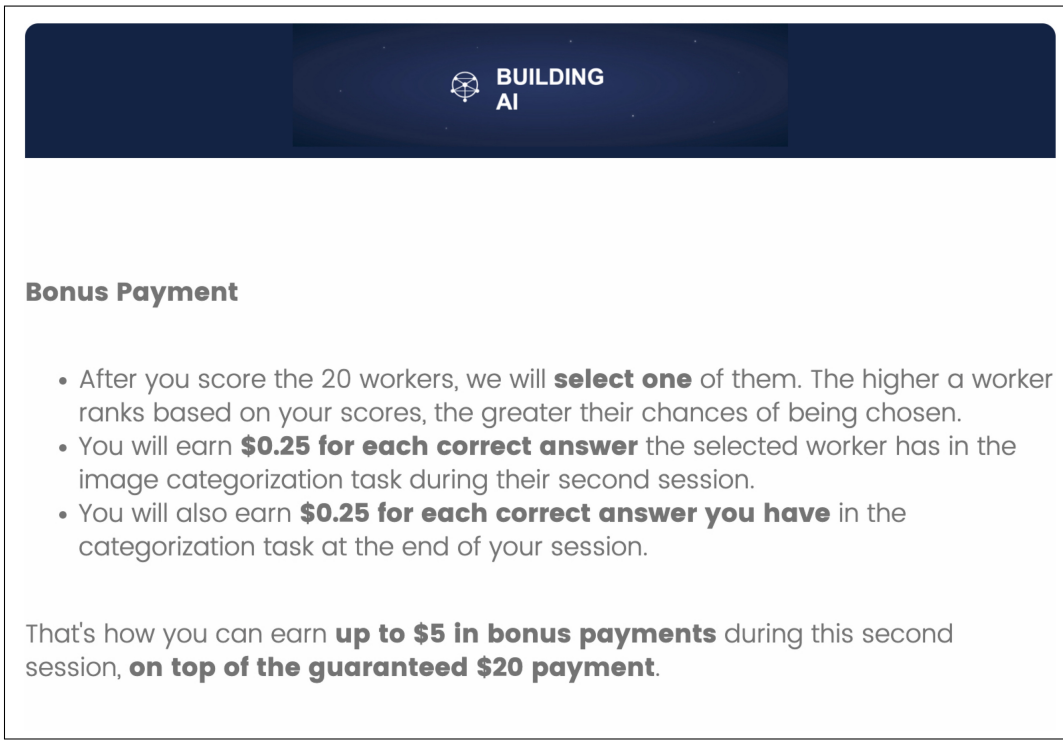
1. Worker Selection Segment: What *you* would have done in the place of the HR specialist?

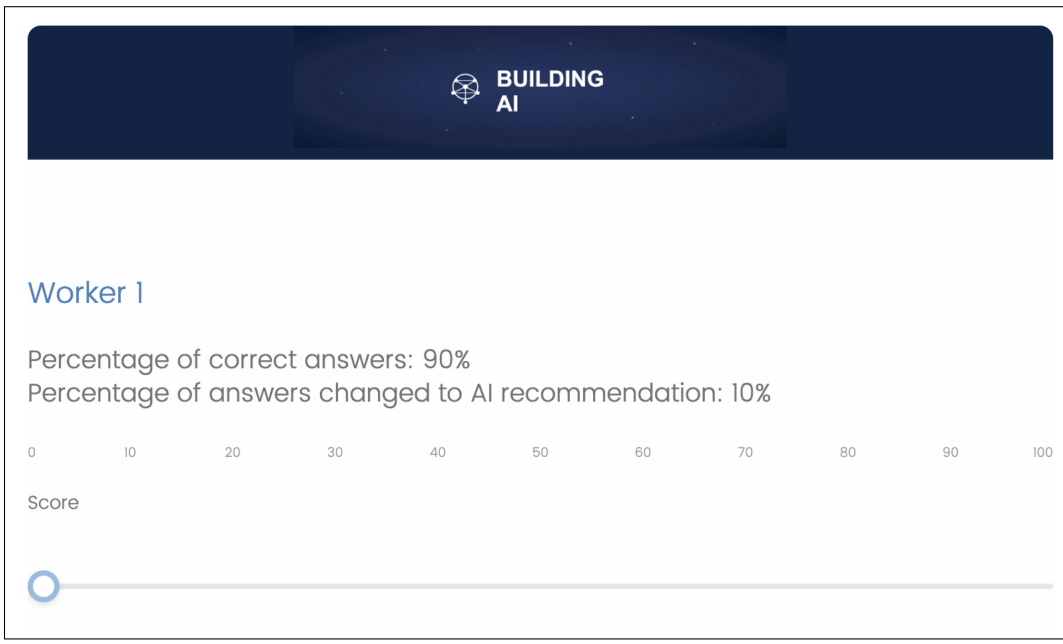
You will see 20 anonymous profiles of other **workers that were also selected to come back for a second session**.

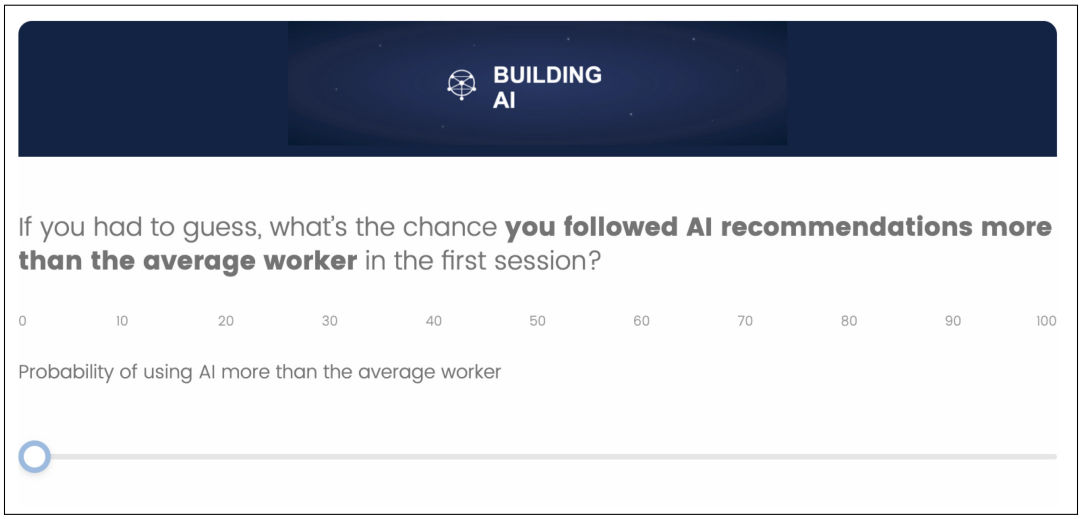
For each worker, you will see the percentage of correct answers and the percentage of answers they changed to match the AI recommendation in their first session, just as the HR specialist did. **Your task is to assign each worker a score from 0 to 100**, with higher scores indicating a stronger preference for selecting that worker.

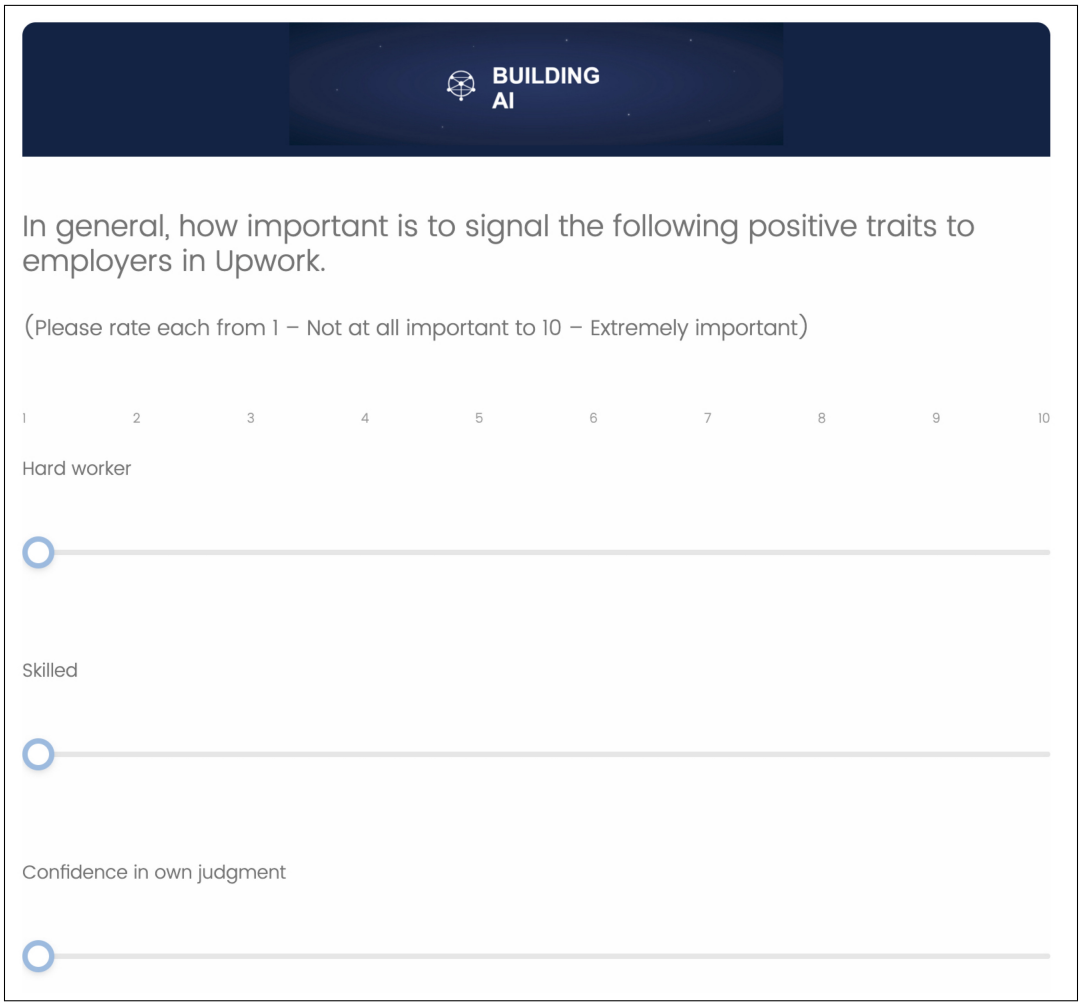
After you complete your evaluations, we will rank the workers based on the scores you gave. One worker will then be selected, with **higher-ranked workers having a higher chance of being chosen**.

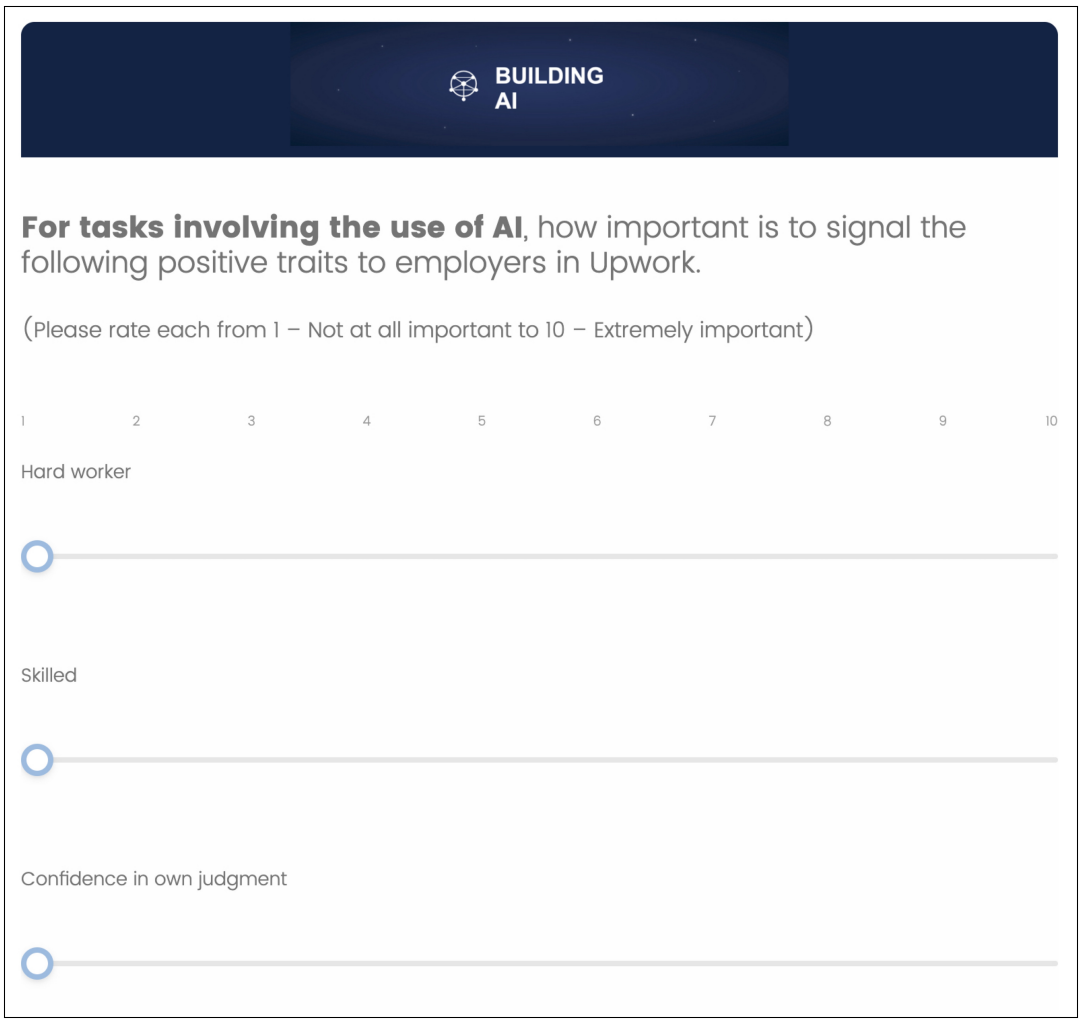
Your bonus will depend on how well the selected worker performs in their second session. So it's in your best interest to score the workers according to your true preferences — this increases the chances that **you'll be rewarded based on the performance of someone you rated highly**.


BUILDING
AI
Bonus Payment
• After you score the 20 workers, we will **select one** of them. The higher a worker ranks based on your scores, the greater their chances of being chosen.
• You will earn **$0.25 for each correct answer** the selected worker has in the image categorization task during their second session.
• You will also earn **$0.25 for each correct answer you have** in the categorization task at the end of your session.
That's how you can earn **up to $5 in bonus payments** during this second session, **on top of the guaranteed $20 payment**.


BUILDING
AI
Worker 1
Percentage of correct answers: 90%
Percentage of answers changed to AI recommendation: 10%
0 10 20 30 40 50 60 70 80 90 100
Score

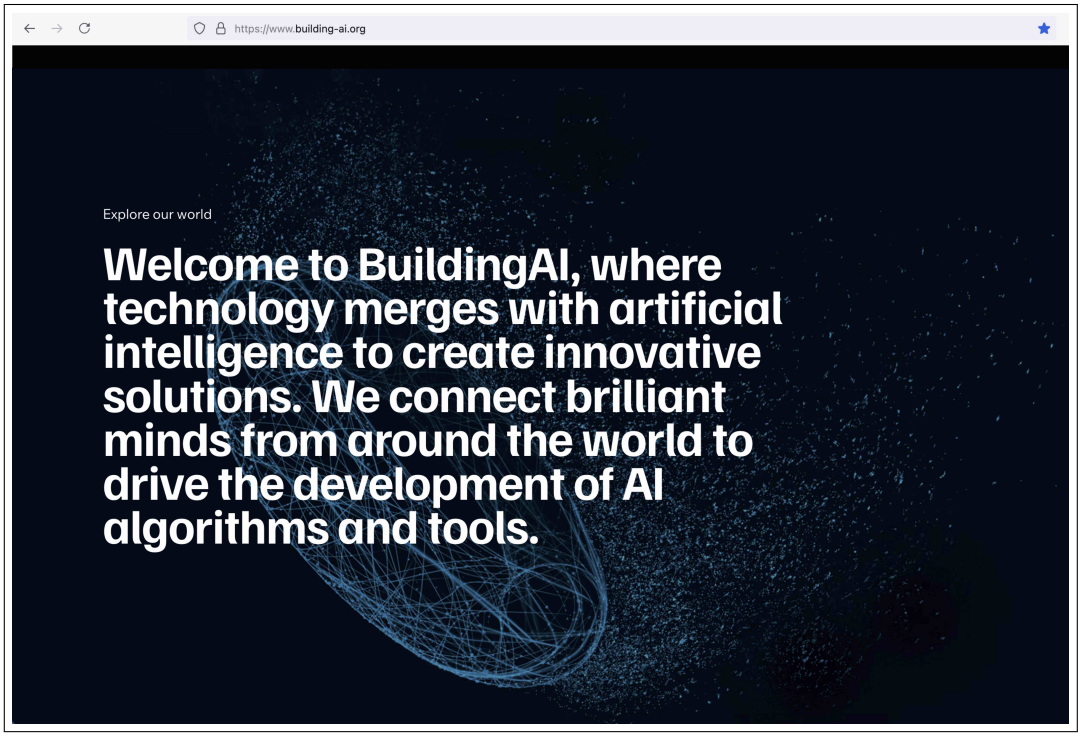

BUILDING
AI
If you had to guess, what's the chance **you followed AI recommendations more than the average worker** in the first session?
0 10 20 30 40 50 60 70 80 90 100
Probability of using AI more than the average worker


BUILDING
AI
In general, how important is to signal the following positive traits to employers in Upwork.
(Please rate each from 1 – Not at all important to 10 – Extremely important)
1 2 3 4 5 6 7 8 9 10
Hard worker
Skilled
Confidence in own judgment

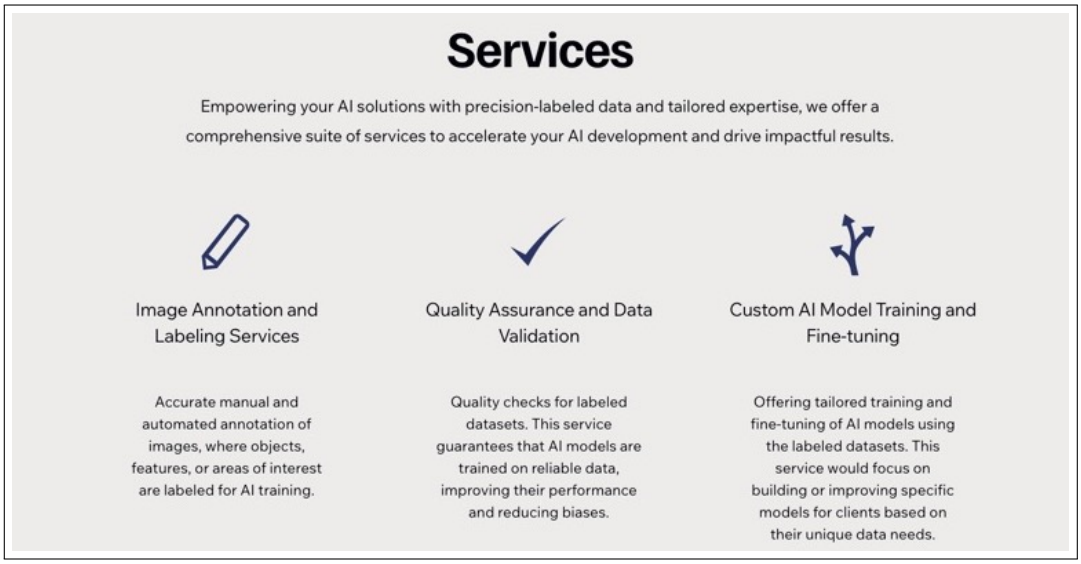

BUILDING
AI
For tasks involving the use of AI, how important is to signal the following positive traits to employers in Upwork.
(Please rate each from 1 – Not at all important to 10 – Extremely important)
1 2 3 4 5 6 7 8 9 10
Hard worker
Skilled
Confidence in own judgment

B.3 *BuildingAI* Webpage

Welcome Screen



Additional Content



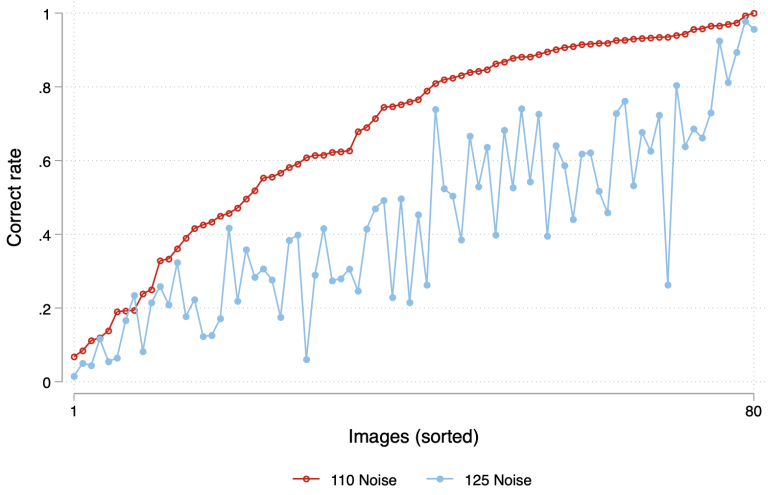
B.4 Task Design

I retrieved 80 images from the dataset made publicly available by Steyvers et al. (2022). These images were selected through multiple rounds of piloting to ensure they were diverse in difficulty yet sufficiently challenging. Each worker was shown a random subsample of 50 images drawn from the pool of 80 images. Images were presented in random order, and the image subsample was drawn independently across workers.

To vary the degree of difficulty in the human and machine classifier experiments, Steyvers et al. (2022) applied four levels of phase noise distortion to the images (80, 95, 110, and 125), which determined the degree of blurriness. This introduced an objective source of variation in image difficulty. In my experiment, I used only the two highest noise levels (110 and 125), with the noise level randomized before an image was shown. Figure B.1 illustrates the variation in task difficulty and highlights a clear within-image increase in difficulty as noise levels rise.

A major advantage of this dataset is that it includes predictions from five pre-trained ImageNet models, along with performance evaluations in collaborative settings: human–machine, two humans, and two machines. This feature enables a non-deceptive experimental design and allows the experiment to draw on algorithms previously tested in studies of human–AI complementarity. In the experiment, I provided participants with predictions from the VGG-19 model, a convolutional neural network with 19 layers. This model was selected because, according to the results in Steyvers et al. (2022), it exhibited the highest complementarity potential among the available options. However, none of the findings in this paper hinge on the use of this specific model.

Figure B.1: Initial Choice Accuracy.



Notes: I compute the accuracy in the first choice as a proxy for image difficulty. Images are then sorted based on this difficulty metric using the 110 noise level (shown in red). The overlaid blue line connects the corresponding accuracy values for each image at the 125 noise level.

C Proofs for Section 3

C.1 Proof of Proposition 1

Consider only tasks on which the worker's initial choice and the AI recommendation disagree. Let

$$D \equiv \Pr(y^0 \neq \hat{y}) > 0,$$

and let $s(p) \in [0,1]$ denote the probability that the worker follows the AI recommendation on a disagreement task when her private confidence is p. A general policy s induces overall AI reliance

$$r(s) = D \int_{1/n}^{1} s(p) f(p)\, dp$$

and expected accuracy

$$a(s) = \theta + D \int_{1/n}^{1} (\kappa - p) s(p) f(p)\, dp.$$

We first show that an optimal policy can be represented by a cutoff rule. Fix an arbitrary policy s, and define

$$q \equiv \int_{1/n}^{1} s(p) f(p)\, dp.$$

Because F is continuous and strictly increasing on its support, there exists a cutoff $\tau \in [1/n, 1]$ such that

$$F(\tau) = q.$$

Consider the cutoff policy

$$s_\tau(p) = \mathbf{1}\{p \leq \tau\}.$$

By construction, s and s_τ induce the same reliance rate:

$$r(s_\tau) = r(s).$$

Moreover, the difference in expected accuracy between the cutoff policy and the original policy is

$$\begin{aligned} a(s_\tau) - a(s) &= D \int_{1/n}^{1} (\kappa - p)\Big[s_\tau(p) - s(p)\Big] f(p)\, dp \\ &= D \int_{1/n}^{1} (\tau - p)\Big[s_\tau(p) - s(p)\Big] f(p)\, dp, \end{aligned}$$

where the second equality follows because the two policies induce the same reliance rate. The final integrand is weakly positive for every p. If $p \leq \tau$, then both $\tau - p$ and $s_\tau(p) - s(p) = 1 - s(p)$ are weakly positive. If $p > \tau$, then both terms are weakly negative. Therefore,

$$a(s_\tau) \geq a(s).$$

Because both S_c and S_t are strictly increasing in accuracy, there exists an optimal policy that is a cutoff rule.

Under a cutoff τ, overall reliance and expected accuracy are

$$r(\tau) = DF(\tau)$$

and

$$a(\tau) = \theta + D \int_{1/n}^{\tau} (\kappa - p) f(p)\, dp.$$

(i) Scoring when AI reliance is not observed. Because $S_c(a)$ is strictly increasing in accuracy, the worker chooses the cutoff that maximizes $a(\tau)$. Differentiating gives

$$a'(\tau) = D(\kappa - \tau) f(\tau).$$

Since $D > 0$ and $f(\tau) > 0$ for every $\tau \in (1/n, 1)$, expected accuracy is increasing when $\tau < \kappa$ and decreasing when $\tau > \kappa$. The unique accuracy-maximizing cutoff is therefore

$$\tau_c^* = \kappa.$$

(ii) Scoring when AI reliance is observed. Define the worker's expected evaluation score induced by cutoff τ as

$$V_t(\tau) \equiv S_t\Big(a(\tau), r(\tau)\Big).$$

Using

$$a'(\tau) = D(\kappa - \tau) f(\tau) \qquad \text{and} \qquad r'(\tau) = Df(\tau),$$

the derivative of the induced objective is

$$\begin{aligned} V_t'(\tau) &= \frac{\partial S_t}{\partial a} a'(\tau) + \frac{\partial S_t}{\partial r} r'(\tau) \\ &= Df(\tau) \left[\frac{\partial S_t}{\partial a} (\kappa - \tau) + \frac{\partial S_t}{\partial r} \right], \end{aligned}$$

where the partial derivatives are evaluated at $\Big(a(\tau), r(\tau)\Big)$.

At an interior optimum τ_t^*, the first-order condition is

$$\left.\frac{\partial S_t}{\partial a}\right|_{(a_t^*, r_t^*)} (\kappa - \tau_t^*) + \left.\frac{\partial S_t}{\partial r}\right|_{(a_t^*, r_t^*)} = 0.$$

Because $\partial S_t / \partial a > 0$, rearranging yields

$$\tau_t^* = \kappa + \left.\frac{\partial S_t / \partial r}{\partial S_t / \partial a}\right|_{(a_t^*, r_t^*)} = \kappa - \lambda^*,$$

where

$$\lambda^* \equiv -\left.\frac{\partial S_t/\partial r}{\partial S_t/\partial a}\right|_{(a_t^*, r_t^*)}.$$

By the single-peaked-objective assumption, $V_t(\tau)$ is strictly quasiconcave in τ. Hence, the first-order condition identifies the unique global maximum among cutoff policies. Since an optimal policy can be represented by a cutoff rule, $\tau_t^* = \kappa - \lambda^*$ is globally optimal.

C.2 The Accuracy–Reliance Curve Exhibits an Inverse-U Shape

Under any cutoff $\tau \in [1/n, 1]$, the worker follows the AI on disagreement tasks whenever $p \leq \tau$. Hence AI reliance is

$$r(\tau) = \Pr(p \leq \tau \mid y^0 \neq \hat{y})\, D = DF(\tau),$$

where $D \equiv \Pr(y^0 \neq \hat{y})$ is the probability of disagreement, which for simplicity let us assume $D > 0$. Since $f(\tau) > 0$ on $(1/n, 1)$, $r(\tau)$ is strictly increasing and therefore invertible, with

$$\tau(r) = F^{-1}(r/D).$$

Accuracy as a function of the cutoff can be written as

$$a(\tau) = \theta + D \int_{1/n}^{\tau} (\kappa - p) f(p)\, dp.$$

Differentiating with respect to τ,

$$\frac{da(\tau)}{d\tau} = D(\kappa - \tau) f(\tau).$$

By the chain rule, the accuracy–reliance curve $a(r) \equiv a(\tau(r))$ satisfies

$$\frac{da(r)}{dr} = \frac{da/d\tau}{dr/d\tau} = \frac{(\kappa - \tau(r)) f(\tau(r)) D}{f(\tau(r)) D} = \kappa - \tau(r) = \kappa - F^{-1}(r/D).$$

Differentiating again,

$$\frac{d^2 a(r)}{dr^2} = -\frac{d\tau(r)}{dr} = -\frac{1}{D\, f(\tau(r))} < 0.$$

Thus $a(r)$ is strictly concave in r. Moreover,

$$\frac{da(r)}{dr} > 0 \iff \tau(r) < \kappa, \qquad \frac{da(r)}{dr} = 0 \iff \tau(r) = \kappa, \qquad \frac{da(r)}{dr} < 0 \iff \tau(r) > \kappa.$$

Therefore accuracy is maximized at $\tau = \kappa$, which is exactly the control-regime cutoff from Proposition 1. Any visibility-induced distortion that shifts the cutoff away from κ lowers accuracy, yielding the inverse-U relationship between expected accuracy and AI reliance.