Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters

Fabrizio Dell'Acqua

Laboratory for Innovation Science, Harvard Business School

Abstract

I investigate how firms should design human/AI collaboration to ensure human workers remain engaged in their activities. I developed a formal model that explores the tension between AI quality and human effort. As AI quality increases, humans have fewer incentives to exert effort and remain attentive, allowing the AI to substitute, rather than augment their performance. Thus, high-performing algorithms may do worse than lower-performing ones in maximizing combined output. I then test these predictions using a pre-registered field experiment where I hired 181 professional recruiters to review 44 resumes. I selected a random subset of screeners to receive algorithmic recommendations about job candidates, and randomly varied the quality of the AI predictions they received. I found that subjects with higher quality AI were less accurate in their assessments of job applications than subjects with lower quality AI. On average, recruiters receiving lower quality AI exerted more effort and spent more time evaluating the resumes, and were less likely to automatically select the AI-recommended candidate. The recruiters collaborating with low-quality AI learned to interact better with their assigned AI and improved their performance. Crucially, these effects were driven by more experienced recruiters. Overall, the results show that maximizing human/AI performance may require lower quality AI, depending on the effort, learning, and skillset of the humans involved.

1 Introduction

In a growing number of organizational settings, firms want to enjoy the benefits of artificial intelligence (AI) while ensuring that their human workers remain attentive and exert oversight in their tasks. For example, firms increasingly use algorithmic recommendations in recruitment selection decisions but are often reluctant to delegate hiring entirely to machines. Managers prefer humans to be informed by AI but remain engaged enough to override algorithmic advice when justified. However, these goals are in tension. As AI performance improves, human overseers face greater incentives to delegate. If the AI appears too high quality, workers are at risk of "falling asleep at the wheel" and mindlessly following its recommendations without deliberation. In such settings, maximizing combined human/AI performance requires trading off the quality of AI against the potential adverse impact on human effort.

In this paper, I investigate how firms should design human/AI collaborations to ensure that humans remain engaged in their activities. I developed a formal model that highlights the tension between increased AI accuracy and reduced human effort. As AI accuracy increases, humans collaborating with AI may become less incentivized to exercise effort. As a consequence of this behavioral response, combined human/AI performance may decrease, leading higher-performing algorithms to be worse partners for humans than lower-performing algorithms.

I test my novel theoretical perspective on this problem using a pre-registered field experiment in collaboration with a recruitment firm. In the field experiment, 181 professional recruiters collectively reviewed nearly 8000 resumes for a software engineering position.[1] The recruiters' task was to select candidates to call back for interviews. They were all HR professionals with extensive experience in this sector who were incentivized to maximize their performance and select the best candidates.

Each recruiter in the experiment received algorithmic recommendations about job candidates but the quality of these AI recommendations were randomized. Recruiters

[1] Columbia University's IRB Protocol AAAT6077. Pre-registration on the Open Science Framework (OSF) Registry, osf.io/qp8et, currently embargoed. Please contact me for details about OSF pre-registration.

received assistance from one of the following 1) a perfectly predictive AI; 2) a high-performing AI; 3) a lower-performing AI; 4) no AI. Recruiters were aware of the type of AI assistance they would be receiving. I evaluated the accuracy of the recommendations produced by the screeners and compared them across experimental treatment arms.

Measuring accuracy is a difficult problem for studies involving AI, as the underlying correct response is often unknown. To address this problem, the recruiters were given job applications[2] to evaluate based on administrative data (PIAAC) from the OECD Statistical group that contains objective performance measures. In particular, the dataset contains quantitative measures of mathematical performance for people included in the dataset. Because I was recruiting for a math-intensive job, recruiters were asked to evaluate job candidates based solely on their math ability. I used these evaluations and objective performance measures to score the recruiters' accuracy.

The structure of the experiment allows me to collect rich outcome data about decisions and behaviors in the presence of AI. I collected callback decisions from all recruiters for every job candidate. I was also able to observe the recruiters' behavior on the platform on which they performed their tasks. Therefore I could reliably collect measures of effort and time spent on each task. Specifically, what information recruiters uncovered in each job application, and how long they spent on evaluating each job candidate. Additionally, I collected data on whether or not recruiters followed the AI's advice depending on their treatment assignment. Finally, I observed changes in behavior over time.

The results of the experiment provide evidence of "falling asleep at the wheel" in the context of AI-supported HR recruitment. I found that subjects with higher-quality AI assistance were *less* accurate in their evaluation of job applicants than subjects with lower-quality AI assistance were. On average, HR recruiters receiving lower-quality AI were less likely to "fall asleep" as they tended not to automatically select the AI-recommended candidate, and instead exerted more effort and spent more time evaluating resumes than their counterparts did.

Crucially, these effects were driven by more experienced recruiters who were more likely to think effectively on their own and improve on the AI's advice. Their behavior

[2]Figure 1 shows a screenshot from the platform displaying one of the job applications.

was beneficial in subjects assigned to the low-quality AI but proved detrimental when collaborating with perfect algorithms. This finding connects to the debate around skill-biased technical change and job polarization.[3] In this task, more experienced recruiters benefitted from lower quality AI and performed worse than the inexperienced recruiters when receiving higher quality AI, suggesting a complex interaction between skills and technology.

Overall, my theory and my results show that maximizing performance in tasks with human-machine interaction does not equal maximizing AI performance in isolation. The experiment setting allowed me to observe in detail the mechanisms behind this phenomenon. The results indicate that settings exist where AI quality and human effort are substitutes: human workers are more likely to free-ride when the quality of algorithmic advice is too high.

In fact, maximizing human/AI performance may require lower quality AI, depending on effort, learning, and the skillset of the humans involved. Designing effective structures for human/machine collaboration requires careful consideration of the organization's objectives and task features.

The remainder of the paper proceeds as follows. I present a theoretical overview and discussion of related literature in Section 2. Section 3 presents my formal model. Section 4 describes my pre-registered experimental design, it details information about the experimental subjects, the task they performed, and what treatment conditions they were assigned. Section 5 presents my empirical strategy and the main analyses of the paper. Section 6 covers the results, Section 7 discusses some implications of the results, and Section 8 concludes the paper.

2 Theoretical Overview

The introduction of AI into more aspects of work has raised concerns that AI will substitute human workers (Jones, 2013). An alternative view prevalent in the social sciences focuses on the complementary potential of human workers with AI (e.g., Frank et

[3]For example, Acemoglu (2002); Autor (2015); Bessen (2016).

al. (2019); Raisch and Krakowski (2021)). This *augmentation* of human capabilities allows teams of human workers and AI to perform better than AI only.[4] In practice, AI has shown to be imperfect for many job-related tasks; despite growing sophistication, it often fails to make perfect predictions.

In this paper, I take the perspective that while algorithms will continue to improve, they will not fully replace humans. Humans will remain in the loop by continually interacting with AI and learning how to work with new technology. AI itself will become increasingly social (Dafoe et al., 2021), making human/AI interaction even more relevant. Organizations are bound to become "highly-automated man-machine systems" (Simon, 1965) and effective collaboration between humans and AI will therefore become crucial for firms integrating AI into their workflow.

Effective human/AI collaboration faces many challenges. A fundamental one is how to get humans to react positively to collaboration with AI. Mistrust of AI may lead human participants to exert less effort when partnered with AI than they would with a human partner (Dell'Acqua et al., 2021). More broadly, humans have shown numerous instances of low trust in algorithmic prediction, a phenomenon called "algorithmic aversion", which manifests in several different organizational settings (Dietvorst et al., 2015, 2018; Luo et al., 2019) and leads humans to override high-quality AI advice in favor of their own judgment.[5] Even if firms make improvements to the quality of their AI predictions, they may not see improved performances because humans are not incorporating them into their decision-making process.

Additionally, a high-performing AI might incentivize humans to decrease effort and to ultimately fully delegate decisions to AI. As theorized by Athey et al. (2020), AI quality and human effort may be substitutes. Self-driving cars provide a vivid and terrifying example of this: AI performing so well that humans pay little attention to the task of driving, practically "falling asleep at the wheel" while allowing the AI to take over. However, in rare instances where driver engagement is required, delegation to AI can

[4] How combined Human and Computer Intelligence will redefine jobs - TechCrunch

[5] Research around the superior performance of algorithms, and the human tendency to distrust them, has a long history, starting from seminal studies by Meehl (1954) and Dawes (1979).

have deadly consequences.[6] Similar adverse effects of over-delegation to AI also exist in many business settings.

Overall, an AI that is "too good" may induce workers to mindlessly follow algorithmic advice and lead to over-delegation. This suggests a fundamental distinction: maximizing the performance of a human using AI does not equal maximizing AI performance in isolation. Maximizing combined human/AI performance requires trading off the quality of AI recommendations with the potential adverse impacts on human effort provision. These tensions are especially salient in settings where workers may have private information about their task that machines do not have or can use less effectively.[7] Organizations that want to maximize human/AI performance may not want to use the best available AI when advising humans.

This paper contributes to three streams of literature on human/AI interaction; first, to an emerging literature on human/AI teams (e.g., Cowgill (2017); Tong et al. (2021)), presenting an instance of human/AI collaboration and highlighting a potential downside of higher-performing AI when paired with humans. Second, it contributes to the literature analyzing the impact of AI on the economy. The "AI productivity paradox" is a widely discussed issue in economics that addresses the incongruity presented by the rapid improvement of AI performance and its failure to reflect more strongly in productivity statistics (Brynjolfsson et al., 2017). The huge transformative potential of AI technologies (Brynjolfsson and McAfee, 2014) seems inconsistent with the deceleration of productivity growth registered over the last decade (Syverson, 2017). This paper presents micro evidence suggesting human responses are the cause of the limited productivity benefits of AI when compared to its technical accomplishments. Thirdly, this paper contributes to our understanding of AI in an organizational context, showing that "falling

[6]The US National Highway Safety Administration recently opened an investigation on a series of crashes, some fatal, involving Tesla's autopilot, https://static.nhtsa.gov/odi/inv/2021/INOA-PE21020-1893.PDF.

[7]In addition to technical or data limitations of machines, workers' knowledge may be tacit, and knowledge transfer within organizations can be difficult (Polanyi, 1966; Kogut and Zander, 1992; Zander and Kogut, 1995; Hansen, 1999; Reagans and McEvily, 2003). According to some in the computer science community, AI technologies, and in particular deep learning, may be able to overcome these limitations and codify human tacit knowledge (Kambhampati, 2021). In those instances, however, human effort would not be relevant for performance.

asleep at the wheel" is a behavior that firms should take into account when integrating AI.

The technical performance of AI has been a common focus in studies of AI. However, if AI primarily augments humans rather than replacing them, then the relevant performance metric becomes combined human/AI performance. Much research has already explored how the best workers are not necessarily the best work peers in human teams, as social skills are an important driver of performance (Deming, 2017; Weidmann and Deming, 2020). Additionally, the best workers may not be the best managers (Benson et al., 2019). This paper extends this line of thought to machine augmentation: AI may be high performing in isolation, but there is a social component that is reflected in human reactions to AI that should be taken into account.

AI in the form of machine learning is inextricably linked to human learning as humans need to learn to interact with machines. In fact, there may be trade-offs between human learning and machine learning (Barach et al., 2019). Over-delegation can lead to skill deterioration in humans (Beane, 2019). More broadly, delegating to AI may lead humans to learn differently, therefore affecting human/AI interaction. On the other hand, noisy decision-making by humans can introduce experimental variation into learning datasets and therefore improve algorithms (Cowgill, 2017).

This paper relates to emerging work in economics and management that discusses the impact of AI on the economy and on organizations (e.g., Agrawal et al. (2018, 2019)), acknowledging that there may be complementarities between AI and human workers (Autor and Dorn, 2013; Acemoglu and Restrepo, 2019). However, recent data suggests that firms may not realize the full gains coming from AI's potential (Brynjolfsson et al., 2017). This may be due to issues of AI integration with complementary assets in organizations, such as human capital, leading to a productivity J-curve: an initial productivity slowdown when a new technology, like AI, is introduced, and an increase in productivity later (Brynjolfsson et al., 2021). Past experience with "general purpose technologies" shows that integration may take an especially long time, even when productivity benefits are very pronounced (Brynjolfsson and Hitt, 2000; Henderson, 2006). This paper explores one of the elements that hinders successful integration

between humans and AI inside of firms. My results suggest that human responses in human/AI collaboration, and in particular their responses to high versus low performing AI, may be one factor behind the productivity J-curve for organizations adopting AI.

A growing literature looks into organizational decisions and responses when automation is introduced in firms (e.g., Raj and Seamans (2019); Puranam (2021)). Regarding human/AI collaboration, specific algorithmic features, such as transparency and reliability, can increase human trust in algorithms, and therefore potentially improve collaboration (Glikson and Woolley, 2020), while features such as opacity may decrease adoption (Lebovitz et al., 2021). There is also much interest in the impact of AI on hierarchical situations where, for example, algorithms give feedback to humans (Tong et al., 2021). Overall, it is important to know where authority is allocated within organizations to understand whether algorithms can translate into improvements in managerial decisions (Athey et al., 2020; Glaeser et al., 2021).[8] This paper introduces the potential concern of over-delegation to algorithms which firms should take into account when incorporating AI in their organization.

3 Model

This section introduces a model that formalizes the intuition behind "falling asleep at the wheel". I am keeping mathematical complexity to a minimum, in order to focus on the core tension that explains why "Bad AI" may perform better than "Good AI" within a human/AI collaboration.

The starting point is the quality of the candidate, q, which can be either 1 or 0. $q = 1$ has ex-ante probability p.

A machine sends a continuous signal s, distributed according to:

$$f(s) = as^{a-1}, \quad s \in [0, 1]$$

[8] Glaeser et al. (2021) show that, in the context of restaurant inspections, using algorithms is beneficial, but there are no additional benefits from algorithmic sophistication. Their results are consistent with the framework presented in this paper in that the lack of benefit from more sophisticated algorithms could be derived from human decision-makers falling asleep at the wheel.

when $q = 1$ and, symmetrically, according to

$$f(s) = a(1-s)^{a-1}, \;\; s \in [0,1]$$

when $q = 0$.

a parametrizes the machine's precision. When $a = 1$, the signal is fully uninformative ($f(s) = 1$ for every $s \in [0,1]$). When $a \to \infty$, on the other hand, the signal is fully informative: $s = q$ with probability 1.

The agent can exert costly effort to ensure that they make the correct decision about the candidate's quality. To avoid corner solutions, I assume that the cost of effort, C, is between 0 and $1/2$. Alternatively, without paying this cost, the agent can base their decision d on the machine's input.

The decision is binary: hire ($d = h$) or reject ($d = r$). $d = q$ indicates a correct decision (i.e., hire high quality candidates and reject low quality ones), while $d \neq q$ indicates the opposite.

The agent draws utility from making the correct decision. I start with symmetric utility functions which simply reflect the probability of making the correct choice.

$$\mathbb{E}(U(d)) = Prob(d = q)$$

First, I compute the posterior probability of the candidate being high-quality for an agent observing a signal $s \in [0,1]$.

Using Bayes' rule, we have that

$$\begin{aligned} P(q=1|s) &= \frac{P(s|q=1) \cdot P(q=1)}{P(s)} \\ &= \frac{as^{a-1} \cdot p}{as^{a-1} \cdot p + a(1-s)^{a-1} \cdot (1-p)} \\ &= \frac{s^{a-1} \cdot p}{s^{a-1} \cdot p + (1-s)^{a-1} \cdot (1-p)} \end{aligned}$$

We start with the following Lemma, which tells us that a more precise machine (higher a) directly translates into more polarized posteriors for the agent.

Lemma 1. *For every $p > 0$, $P(q = 1|s)$ is increasing in machine precision a whenever $s > 1/2$, and decreasing whenever $s < 1/2$.*

Proof. Taking the partial derivative with respect to a,

$$\frac{\partial\left(\frac{s^{a-1}\cdot p}{s^{a-1}\cdot p+(1-s)^{a-1}\cdot(1-p)}\right)}{\partial a} = ln(s)s^{a-1}p\cdot\left(s^{a-1}\cdot p+(1-s)^{a-1}\cdot(1-p)\right) - s^{a-1}p\cdot\left(ln(s)s^{a-1}\cdot p+ln(1-s)(1-s)^{a-1}\cdot(1-p)\right)$$

Some straightforward algebraic manipulation shows that this is equal to

$$ln(s/1-s)\cdot K(a,p,s),$$

where $K(a,p,s)$ is a non-negative function $\forall a,p,s$. Therefore, the derivative has the same sign as $ln(s/1-s)$. This is positive whenever $s/(1-s) > 1$, or $s > 1/2$, and negative otherwise, which concludes the proof.

■

To reiterate, this Lemma formalizes a simple notion: the more precise the AI, the higher the posterior upon observing a high signal (and the lower the posterior upon observing a low signal).

Because additional effort – which comes at a cost C – guarantees the agent a correct decision, and thus a utility $1-C$, the agent will exert effort if and only if $P(q = 1|s) \in [C, 1-C]$. That is, the agent will slack if the posterior is close to 0 or 1, and exert additional effort otherwise. Because the agent's posterior is monotonically increasing in the signal s, this maps directly into two thresholds $\underline{s}$ and $\bar{s}$ for the signal.

This leads me to my central theoretical result:

Theorem 1 (More Precise AI Can Result In Poorer Decisions)**.** *A more precise AI can result in less precise human/AI decision-making.*

Proof. We have that

$$P(q = 1|s) = P(a, s, p) = \frac{s^{a-1} \cdot p}{s^{a-1} \cdot p + (1-s)^{a-1} \cdot (1-p)}.$$

Ex-post precision is given by $P(a, s, p)$ whenever it is below C or above $1 - C$, and 1 otherwise. More formally, we have that ex-ante precision equals

$$2Prob(P(a, s, p) < C) \cdot \int_0^C P(a, s, p) dP(a, s, p) + Prob(P(a, s, p) \in (C, 1 - C)) \cdot 1.$$

What happens with a is important. Higher a does two things here; first, it makes posteriors more informative point-wise, as we've seen in the Lemma. Second, it increases the probability that the agent fully delegates to the machine, that is, it increases $Prob(P(a, s, p) < C$. Intuitively, increasing machine precision increases the probability that the agent simply follows the machine's advice and exerts no effort.

I want to show that there exist values of C such that this expression is non-monotonic in a. To this end, note $a = 1$ and $a \to \infty$ achieve perfect precision. The former does so because the machine is totally inaccurate, leading the agent to always exert effort. The latter does so without human effort.

Now consider $a \in (1, \infty)$. For each of these values, we have that $Prob(P(a, s, p) < C) > 0$: take s close enough to 0, and $P(a, s, p)$ will be close to 0 as well (as $P(a, 0, p) = 0 \; \forall a > 1$). This means that inaccurate decision making will occur whenever $a \in (1, \infty)$.

That is, ex-ante precision is U-shaped in the AI's precision.

■

This provides a theoretical foundation as to why more precise AI may lead to less precise human/AI decision-making. An important remark is in order. I have assumed that the agent can achieve perfect accuracy by exerting effort. This is to emphasize the core tension of this paper with mathematical simplicity. One could think of a (more realistic, but also mathematically cumbersome) model in which the agent's effort increases precision, while not achieving perfection. The core tension I describe would remain intact – only the ex-ante precision would be J-shaped (and not U-shaped) in the

AI's precision.

While not the focus of my experiment, I conclude the theoretical section with a remark on the potential unintended (positive) consequences of algorithmic aversion, that is, the idea that humans misperceive AI's prediction accuracy to be lower than what it is (see, e.g., Dietvorst et al. (2015)).

In line with the literature, I model algorithmic aversion as an excessively negative belief $\hat{a}$ of machine accuracy: $\hat{a} < a$. We have the following result.

Theorem 2. *Algorithm Aversion leads to more precise Human / AI combined decisions.*

Proof. I prove this result constructively. Consider a combination (a, s) such that $p(q = 1|s = 0) = C$.[9]

Then, if the agent believes $\hat{a} < a$, they will conclude $p(q = 1|s = 0) > C$. This is pivotal in terms of effort provision: the agent will exert effort when believing $\hat{a} < a$, whereas they would not have if they believed $\hat{a} = a$.

Therefore, in these instances, the agent will achieve perfect (combined) precision, if and only if, they are algorithmic averse. Because algorithmic aversion never leads to strictly less precise decisions, this concludes the proof: ex-ante expected precision increases in algorithmic aversion.

■

Note that the agent being perfectly accurate whenever they exert effort implies that their algorithmic aversion is never detrimental to decision making. This is because the agent's underestimation of the machine may be pivotal and lead to effort that a rational agent would not exert; extra effort that also leads to perfect accuracy. So, this modeling choice rules out situations in which the AI is actually more accurate than the agent, but the agent's (relative) overconfidence leads them to underweight the machine signal, possibly leading to a worse combined decision.

[9] To be perfectly precise, consider the case of $p(q = 1|s = 0) = C - \epsilon$, for a very small $\epsilon > 0$.

4 Experimental Design

Hiring is one of the most frequently proposed applications of AI.[10] However, most firms are reluctant to delegate hiring entirely to machines. Managers prefer humans to be guided by AI but engaged enough to override algorithmic advice when justified. Additionally, HR algorithms can be gamed by human applicants. There are websites[11] with the explicit objective of parsing human CVs through an algorithm before submitting them to "get past resume robots.[12] Organizations may want to avoid such gaming and prefer to keep a human-in-the-loop to separate qualified resumes from gamed ones.[13]

In this experiment, I hired HR recruiters and randomly assigned them to receive algorithmic recommendations on job screening decisions. The quality of AI assistance was also randomized: some recruiters collaborated with a high-performing algorithm, and others utilized a lower-performing one. This variation allowed me to measure the responsiveness of the recruiters' behaviors to algorithmic advice.

Each recruiter in the experiment received a packet of 44 job applications. They saw one job application at a time and needed to select whether it was above a certain threshold, that is, whether or not to call this person for an interview. Their objective was to select all candidates above a certain threshold: there were no limits to the number of candidates they could select. The job description specified that recruiters should choose candidates based on their mathematical ability, as they are recruiting for a software engineering job, which is math-heavy. There were incentives for recruiters to respond accurately and truthfully.[14]

[10] According to a Mercer 2019 report, 88% of companies globally use some form of AI in their HR operations, https://www.shrm.org/ResourcesAndTools/hr-topics/global-hr/Pages/Employers-Embrace-Artificial-Intelligence-for-HR.aspx

[11] https://www.jobscan.co/

[12] Jobscan writes on the homepage, "systems analyze resumes and CVs to surface candidates that best match the position, but qualified applicants slip through the cracks. We have researched the top systems and built our resume checker based on their common patterns to help you get noticed".

[13] This is similar to the context studied in Choudhury et al. (2020), where humans with domain expertise are found to successfully complement AI and reduce the bias resulting from strategic behavior by agents.

[14] The structure of this experiment resembles that of a "two-sided audit" (see Cowgill and Perkowski (2020).

4.1 Why an experiment

Studies on AI face an obvious challenge in that the introduction of AI within firms and teams does not happen randomly, leading to potential endogeneity concerns. Additionally, the quality of the AI introduced is not random. An experimental approach is required to address these concerns and focus on the causal impact of the introduction of high or low quality AI.

Measuring accuracy is a difficult problem for studies involving AI, as the underlying correct response is often unknown. Algorithms are built using data from the real world, however, these data may not be a true representation of an algorithm's target because of selection or bias in data collection. This experimental design addresses this issue by using administrative data to create job applications for evaluation. Section 4.8 describes the data and this paper's approach.

The experimental setting allows me to test my hypotheses in a controlled environment that is a realistic setting, with a subject population of real workers performing their standard work activities. Additionally, it allowed me to collect detailed behavior of HR recruiters working on their task. The experiment was pre-registered: specifically, the hypotheses, main analyses and variables collected were registered before any recruiter was contacted.

4.2 Subjects

For this experiment, I hired 181 professional recruiters, at market wages, from an online freelancer recruitment platform where many HR recruiters operate. They were selected based on experience and their record of candidate evaluations on the platform. The recruiters all had at least one year of experience in HR recruitment and had a history of performing similar tasks on the platform.

Table 1 Panel A displays a statistical summary of the demographic variables present in the sample group, Panel B shows their representation within each condition grouping: "Perfect prediction", "Good AI", "Bad AI", and "No AI" /"Control". The majority of the subject pool consisted of women, who are similarly overrepresented in the HR sector as a

whole; 60% of the subjects self-identified as Caucasian, 14% as Black, 9% as Latino/a/x, and 7% identified as Asian. More than two-thirds of the subjects held a university degree; of the remaining participants, some had a high school diploma, some received vocational training, or obtained college credits but no degree. I hired recruiters based in OECD countries that had previously worked with US-based customers. The vast majority of recruiters (89%) were themselves US-based. Section 4.4 describes how subjects were paid.

Recruiters were assigned to different treatments with stratified random assignments. I stratified on gender, race, hourly rate, whether they were US-based, how long they had worked on the platform, how many hours they had billed on the platform, and whether they had received feedback on the platform. As the recruiters were hired sequentially in multiple batches (and not altogether at once), I used sequential re-randomization. This allowed me to keep the treatment groups well-balanced enough to produce more precise estimates. Panel B of Table 1 shows there were no relevant differences across treatments on any demographic characteristics.

4.3 Freelance recruiters

For this experiment, I hired 44 HR recruiters on a large online freelancer platform. This is an ideal population to draw from as their jobs consist of various freelance hiring tasks, not dissimilar to the one I offered them. Each recruiter had spent at least one year performing HR activities, and as a group, they completed more than 50,000 hours of work on the platform.[15]

The recruitment process outsourcing (RPO) industry overall is one that has been growing for years - according to some estimates filling 36% of job applications in the US in 2018 - and is projected to grow in the future.[16] As an example, the state of New York hired a firm to do outsourced recruiting and hire over 7,000 contact tracers from about 50,000 applicants in the early stages of the COVID-19 pandemic.[17]

Many of the freelance recruiters openly affirmed that their main goal in a contract

[15] Agan et al. (2021) used a similar pool of subjects to study questions around hiring and labor.
[16] See Cowgill and Perkowski (2020).
[17] https://info.leveluphcs.com/nys-contact-tracing-video

was to satisfy client requirements and to that end, they would follow our instructions faithfully. Consequently, they were careful in following the rubric provided in this task. All recruiters needed to confirm that they understood the task instructions before receiving the job applications to start their evaluations. I also asked the recruiters to share their online platform CV with me, as well as a copy of their offline CV, or LinkedIn profile. This information was used to collect data about the subjects' characteristics that were later used as input variables in my analysis.

4.4 Payment

Recruiters were hired for one hour at their declared hourly wage on the platform. They charged on average $40.25 an hour, but with a very broad variation. The lowest-paid recruiter had an hourly rate of $5, while the highest-paid had an hourly rate of $100.[18] Additionally, recruiters received a bonus between $0 and $20 based on their accuracy. More accurate recruiters received more, regardless of their treatment assignment or their base rate.

4.5 Task

The recruiters were hired for one hour to score 44 job applications each. The work took place on a web application where they were presented with one candidate at a time. They could see the candidate's name, and buttons that when clicked revealed sections of the candidate's bio and CV. Figure 1 shows a screenshot from the platform displaying a job application. Depending on the treatment group recruiters were assigned to, they either received AI assistance of varying qualities or none. The web application measured the number of clicks each recruiter made on the job application, and the time spent on each page. Figure 2 shows the same screenshot after one of the buttons has been clicked, uncovering information about the education of the applicant. I used data about clicks on the job application to make inferences about recruiters' effort level in the task.

For each job application, the recruiter had to choose whether or not to interview the

[18] I capped my maximum expense per recruiter per hour to $100.

candidate. Their task was to select the best candidate based solely on their math ability, without taking into account any other criterion. Their instructions specified that any other hiring goals, for example, skills other than math, or location preferences, would have been incorporated through a separate process. The recruiters' task was to focus on math ability exclusively.

The instructions clarified that each evaluation should happen independently and be based solely on the candidate's quality. There was no minimum or maximum number of candidates that could be selected. The recruiters' binary decision, combined with ground truth (as described in section 4.8), generated a binary accuracy measure for each scored job application. If the recruiter chose to interview a candidate who was above the relevant ability threshold, or chose not to interview a candidate below the relevant ability threshold, then their decision was scored as 1 (i.e., a correct decision). All other decisions were scored as 0.

As recruiters chose whether or not to interview a candidate, they also reported on their confidence levels (on a 1-5 scale) regarding each decision. This generated a measure of 10 potential evaluations per job application; as a recruiter could decide to interview a candidate and rate their confidence level at any point between "extremely confident" and "not confident at all" . They could also decide not to interview a candidate and give the same range of responses regarding their confidence in their decision. These scores, combined with ground truth, produced an accuracy measure on a 1-10 scale for each job application scored. These accuracy measures were both pre-registered, and are the main dependent variables in this study.

In the first twelve iterations, recruiters learned to interact with their AI. They received feedback after each binary interview decision, regarding whether or not their choice was correct. Thus, they could adjust their expectations on the quality of the AI assistance. Additionally, they were able to develop strategies to best interact with their AI, adjusting the time and effort they spent on each job application. These iterations allowed recruiters to experiment with the AI and form beliefs about its quality without impacting their pay. See Appendix E for details about the specific language used for different types of feedback.

After scoring all 44 job applications, HR recruiters answered a series of questions regarding their interaction with the AI and their prior experience with the technology (see Appendix A). Finally, recruiters were asked to complete a demographic section (see Appendix B).

4.6 Partner Company

The details of this task were developed with the assistance of a partner company; a recruitment agency that collaborates with multiple Fortune 1000 companies for their HR needs. They went through the experimental design and made recommendations, importantly, they helped refine details of the study - specifically around language - to make the task, and overall project, more credible to the HR recruiters hired for the experiment.

4.7 Experimental Structure

Each recruiter scored job applications in three phases; the first phase consisted of 4 job applications that the recruiter scored independently of their treatment assignment and without AI support. After scoring each job application, recruiters received feedback on whether or not their evaluation was correct based on ground truth data.[19] This phase provides me with a baseline to improve the precision of the estimates discussed in Section 6: some of the analyses included a baseline accuracy level calculated from recruiters' evaluations of their first 4 job applications.

In a second experimental phase, recruiters received recommendations from the AI they were assigned to: for example, recruiters in the first condition group could see recommendations from the "Perfect prediction" benchmark and made their evaluation based on that information. After submitting their evaluation, recruiters received feedback about whether or not they were correct. This phase lasted for 8 iterations and had the explicit goal of allowing recruiters to familiarize themselves with the AI in their treatment

[19] Recruiters did not know ground truth data existed, but were presented in phase 1 and phase 2 with job applications that they were told the company had already evaluated. This served the purpose of allowing the recruiters to familiarize themselves with the platform using applicants the company could provide feedback on.

group. Recruiters knew the quality of AI they were assigned to (as reported in their instructions) before starting the task, but this familiarization phase helped them test the quality of their AI, confirm that the instructions were accurate, and learn to incorporate it in their evaluations.

Finally, the third experimental phase had 32 iterations and was the main part of the experiment. Recruiters knew that the job applications they scored in this phase would be evaluated without any feedback given. These 32 iterations are the core of the experiment and form the phase that I include in my main specifications. Figure 3 provides a visualization of the experiment's structure.

4.8 Ground Truth Data

The job applications were constructed using data from the OECD's *Programme for the International Assessment of Adult Competencies* (PIAAC, Schleicher, 2008).[20] The PIAAC data contains quantitative measures of math, language, and problem-solving skills for a representative sample of the working-age population in 24 countries. It provides data for international measures and is the canonical dataset for cross-country and within-country comparisons of numeracy and other adult skills (McGowan and Andrews, 2015; Hanushek et al., 2015).

This dataset closely resembles the "ground truth" ability of candidates and can therefore be used to score the recruiters' accuracy. In particular, I compare the recruiters' assessments about math ability with PIAAC data about math ability: this allows me to create a measure of accuracy for the recruiters' evaluations, as described in Section 4.5.

4.9 Experimental Treatments

Recruiters were randomly assigned to one of four experimental conditions. In three of the conditions they were treated by receiving AI support while analyzing their set of job applications; they received AI-generated recommendations about each jobseeker's

[20] Cowgill et al. (2020) uses the same data, which is a cleaned and merged version of the PIAAc dataset and can be accessed on dataverse.org, https://doi.org/10.7910/DVN/JAJ3CP

quality. The fourth condition was used as a control condition and had no AI support assigned ("No AI").

The first three AI assistance conditions were determined by their level of accuracy; the first condition was a "Perfect prediction" benchmark that correctly predicted the applicants' quality and whether or not they should be interviewed. The second condition was a high-performing AI with accuracy of around 85% ("Good AI"), and the third condition was a low-performing AI with accuracy of approximately 75% ("Bad AI").[21]

The "Perfect prediction" condition served as a baseline for the experiment. The optimal strategy under this unrealistic condition would be a full delegation: no use of human judgment with all AI recommendations followed. In my analyses, I compared each treatment group to the control condition of "No AI", additionally I compared the effects of being assigned to "Good AI" versus the effects of being assigned to "Bad AI".

"Good AI" and "Bad AI" were the central manipulations of the study. They allowed me to address the question of whether, and to what extent, humans incorporate AI predictions with varied performance levels, and likewise compensate for the AI's shortcomings through personal effort. Do lower levels of performance induce humans to "stay awake"; conversely, does too-accurate AI lead to them "falling asleep at the wheel"?

Note that all recruiters were aware of their assigned AI's quality, which is reported in their instructions (see Appendix D.2 for details). They were also able to learn how to interact with their AI assistance in the initial iterations: these were not scored and were not counted in the recruiters' accuracy scores as they only served for training purposes. This experience was intended to allow recruiters to calibrate the accuracy of their AI assistance and help them decide whether to rely on AI recommendations.

4.10 Algorithms

Three of the four treatment conditions assigned recruiters to collaborate with AI with different algorithms. Recruiters assigned to "Perfect prediction" received assistance from an algorithm that reported the job candidates' exact underlying ground truth math ability:

[21] That is, "Good AI" gave the correct decision about whether to interview a candidate or not for about 85% of candidates, while "Bad AI" gave the correct decision for about 75% of candidates.

this condition served as the benchmark. In very few real-world applications can we count on perfect algorithms. In fact, in my experiment, this was not a prediction algorithm, but rather an algorithm that delivered precise ground truth values to recruiters. Wherever this type of algorithm exists, the correct response is for humans not to take part in the task, and to simply allow the AI to perform it. Ideally, we would have full delegation to AI and would be in the context of human substitution, not human augmentation.

The central conditions for this study, "Good AI" and "Bad AI" algorithms, were both developed by software engineers in a previous experiment that I ran using the PIAAC data (Cowgill et al., 2020). In that study, we asked approximately 400 software engineers to develop an algorithm to predict math ability for people in the PIAAC dataset. "Good AI" is based on the very best AI developed for that experiment, that is, the AI with the lowest mean squared error (MSE) in the test set across all participants. This algorithm was the most accurate one developed by any of the software engineers and had an accuracy of 85% when scoring job applications in this experiment. "Bad AI" is based on a combination of the two median-performing algorithms (based on MSE scores) produced in the same experiment. This combination had a MSE similar to the algorithms at the 35th percentile of accuracy in the Cowgill et al. (2020) experiment. This algorithm had an accuracy of 75% when scoring job applications.[22]

These algorithms are perfectly suited for this experiment, as they were developed in a realistic setting, at the request of a company using the same dataset, by a population of highly skilled software engineers, many of whom went on to work for some of the world's top tech companies.

5 Main Specifications

My goal is to estimate the effect of different AI treatments on performance (i.e., accuracy in job application evaluations). I am also interested in how AI assignments affect adherence to AI recommendations, how long subjects spent on each task, and how much

[22]Subjects' instructions reported the percentage accuracy for the algorithm they were working with, not the MSE, as this was much easier for HR recruiters to interpret.

effort they exerted in performing that task. Below is reported my main regression:

$$y_{s,c,r} = \beta_0 + \beta_1 * PerfectAI_s + \beta_2 * GoodAI_s + \beta_3 * BadAI_s + \theta_c + \tau_r + \chi_s + \epsilon_s \quad (1)$$

where s indexes recruiters, c indexes job candidates, and r indexes rounds. θ_c is a vector of candidates' job application fixed effects, τ_r is a vector of round fixed effects, and χ_s is a set of control variables with recruiter characteristics such as gender, and ϵ_s is the error term. I estimate equation 1 using standard errors clustered at the recruiter's level.

$PerfectAI_s$, $GoodAI_s$, and $BadAI_s$ are binary indicators taking the value 1 if subject s was assigned to their treatment condition, and 0 otherwise. β_1 captures the effects of receiving the most accurate AI, β_2 captures the effects of receiving "Good AI", and β_3 captures the effects of receiving "Bad AI".

5.1 Dependent Variables

Follows AI is a binary variable, taking the value 1 if a subject followed the recommendation provided by the AI, and 0 otherwise. It is always 0 for subjects assigned to the "No AI" condition.

My two main dependent variables regard accuracy; both of these were pre-registered measures. The first one, *Accuracy (binary)*, captures whether the job application was scored correctly by the recruiter. A decision was scored as 1, if the recruiter chose to interview a candidate who was above the relevant ability threshold, or if the recruiter chose not to interview a candidate who was below the relevant ability threshold (i.e. the correct choice). All other decisions were scored as 0.

The second one, *Accuracy*, is created using the recruiter's decision with their stated confidence level (measured on a 1-5 scale)[23], creating a 10-point judgment scale for each decision. A recruiter could be "extremely confident" about deciding to interview a candidate, leading to a value of 10, or they could merely be "very confident" about this decision, leading to a value of 9, and so on. At the other end of the scale, they could be "extremely confident" about not interviewing a candidate, leading to a value of 1, or

[23]Recruiters could be "extremely confident", "very confident, "confident", "not particularly confident", or "not confident at all".

"very confident" about the same decision, leading to a value of 2, and so on. Figure 4 displays a screenshot with the recruiters' potential assessments and their numeric values. I used these judgments and combined them with ground truth to calculate accuracy. An always-correct recruiter would make all perfect predictions about candidates and be "extremely confident" about all of those predictions. An always-incorrect recruiter would make all the wrong predictions, and be "extremely confident" about those predictions. Accuracy captures the distance from this ideal, rescaled in a way that higher numbers indicate greater accuracy. The always-incorrect recruiter would have an accuracy of 1, while the always-correct recruiter would have an accuracy of 10.

Time Spent is a variable that reports the number of seconds a subject spent on a given job application. A higher number indicates more time was spent analyzing a job application. *Effort (clicks)* reports the number of times subjects clicked when exploring a given job application. Subjects who clicked more were exerting more effort when evaluating job applicants.

5.2 Observations

181 recruiters evaluated 44 job applications each, with a total of 7964 evaluations. My main analyses report results for the last 32 job applications evaluated by each recruiter, as the first 12 were part of a learning phase where recruiters were familiarizing themselves with the platform and their assigned AI. Additionally, I excluded 19 HR recruiters who failed a simple attention check, resulting in a sample of 5184 evaluations. Most of the analyses use this sample.[24]

[24]The recruiters who failed the attention check were evenly distributed across treatment conditions. Results are consistent if I include these observations.

6 Results

6.1 Impact of AI Quality on following AI's recommendations

I began by showing experimental compliance. Table 2 illustrates the effect of the AI treatments on adherence to AI advice. Subjects who received "Perfect prediction" were most likely to follow AI advice, subjects who received "Good AI" were less likely to follow AI advice, but did so more often than subjects who received "Bad AI". Those subjects who received "Bad AI" were the least likely to follow AI advice. This table shows that subjects complied with what was expected of them given their treatment assignment, as they responded to the quality of the assigned algorithm and were more likely to follow the recommendations of more accurate AI.

6.2 Impact of AI Quality on the Accuracy of recruiters' evaluations

Table 3 displays the effect of all AI treatments on accuracy when compared with a control of "No AI". Columns 1-2 present the effects of receiving AI assistance on *Accuracy (binary)*, while columns 3-4 report the effects for accuracy when measured on a 1-10 scale. Columns 1 and 3 report the results of regressions with a set of recruiters' characteristics as controls. Columns 2 and 4 include recruiters' HR experience, their AI experience, as well as their initial accuracy (i.e., their average accuracy in the first 4 iterations). This table shows that on average, receiving AI assistance (regardless of its quality) improved performance. This is true however we measure performance.

Table 4 reports the main results of the study. I theorized on the impact of receiving different types of AI assistance against a control of "No AI". This table examines the effects of AI treatments on accuracy: columns 1 and 2 report the results by using *Accuracy (binary)* as a dependent variable, while columns 3 and 4 report those for *Accuracy*. Columns 1 and 3 report the results of regressions with a set of recruiters' characteristics as controls.[25] Columns 2 and 4 also include recruiters' HR experience, their AI experience,

[25]Controls included recruiters' racial (Caucasian, Black, Latino/a/x, Asian, Other), gender (Female, Male, Other), and educational (High School, Bachelor, Master, PhD degrees) characteristics, as well as their hourly rate and a dummy variable taking the value 1 if they were US-based and 0 otherwise.

and their initial accuracy. This table also demonstrates that receiving "Perfect prediction" had the greatest impact on performance. This is unsurprising as "Perfect prediction" made no mistakes in its predictions; it is a baseline for an unrealistic AI that would be able to fully substitute humans. Although subjects assigned to this condition did not comply with the "Perfect prediction" every time, they followed the recommendations often enough to have an important boost to their performance and performed better than any other group.

"Good AI" and "Bad AI" both have, on average. positive effects on performance when compared with "No AI". In line with my theory, "Bad AI" outperformed "Good AI" in all specifications. This effect is stronger and more precisely estimated when Accuracy is measured on a 1-10 scale (columns 3 and 4) rather than a binary one (columns 1 and 2). The bottom of the table displays a p-value that tests whether the effects of receiving "Good AI" were equivalent to those of receiving "Bad AI". They report a significant difference in columns 3 and 4. These results show that recruiters who received worse quality AI performed, on average, better than recruiters who received higher quality (but not perfect) AI. In the next sections, I explore some of the potential mechanisms behind this finding.

6.3 Impact of AI Quality on the time spent and the effort of recruiters

Table 5 examines other sources of differential impact between "Good AI" and "Bad AI" and focuses on the effect of AI treatments on the effort subjects exert. In particular, I focus on the time spent by subjects on each job application (columns 1 and 2) and on the number of clicks made on each job application (columns 3 and 4). As in the previous tables, columns 1 and 3 report the results for the basic model, while columns 2 and 4 include recruiters' HR experience, their AI experience, as well as their initial accuracy.

The table shows that subjects in the "Perfect prediction" condition spent less time on their assigned job applications when compared to subjects in the control group ("No AI"). The effects on clicks were minimal. Subjects receiving "Bad AI" spent more time and clicks on each job application when compared to any other condition. In particular, these subjects spent much longer on each job application and clicked more times than subjects

assigned to “Good AI”. The p-values at the bottom of the table show this difference is especially significant for time spent on the task.

These results confirm my predictions about “falling asleep at the wheel”. Subjects receiving better quality AI spend less time on their job applications. Their performance is not affected if they collaborate with the “Perfect Prediction”, and human intervention is not required. However, in the “Good AI” condition, which would require humans to invest more time and effort to improve their performance, subjects neglected to do so. Subjects in the “Bad AI” condition, by comparison, spent more time and effort than subjects in all other conditions.

6.4 IV regressions - effects of the increase in following AI

One of the questions that can be explored with this experimental design is whether following AI more often leads to lower performance in the “Good AI” condition. Table 6 reports instrumental variable regressions where I use treatment assignments to instrument the decision to follow AI recommendations. Columns 1 and 2 report the effects on *Accuracy (binary)* (column 1) and *Accuracy* (column 2) of the increase in following any type of AI (“Perfect prediction”, “Good AI”, or “Bad AI”). The results have shown that an increase in following AI leads to better performance on average. This is unsurprising as AI provides useful information, and is beneficial to subjects when compared to a control of “No AI”.

Columns 3 and 4 focus on the subsample of subjects who received either “Good” or “Bad AI", for a total of 2592 observations. The columns report the effects of the increase in following AI on accuracy as determined by having received “Good AI” rather than “Bad AI”. Receiving “Good AI” makes subjects more likely to follow AI advice than when receiving “Bad AI" advice. I use the treatment assignment to “Good AI” to instrument the decision to follow the AI’s recommendations. These results show that the increase in following “Good AI" recommendations, in fact, leads to lower performance. This is in line with the theoretical framework of this paper: subjects who receive “Good AI” are more likely to simply follow AI advice over their own relying on their own experience, and because of this their performance decreases.

6.5 Heterogeneous effects

Table 7 and Table 8 examine heterogeneity in the treatment effects discussed in Table 4. In particular, they divide the sample in three different ways: columns 1 and 2 compare the accuracy for the first part of the task (1), with accuracy in the second part (2). This analysis includes only the assessments of phase 3; column 1 reports the results of iterations 13-28 and column 2 reports results for iterations 29-44. In theory, we might expect accuracy to decrease over time, as recruiters get more tired, which would have a stronger impact in the treatments where recruiters exert more effort. Instead, when we compare coefficients in column 1 and column 2, we see the differences are minimal across all conditions in both tables.

There is a large body of literature that highlights how new technologies in the past few decades have been skill-biased; favoring more skilled workers to the detriment of less-skilled ones (Goldin and Katz, 1998). However, it is unclear whether AI fits this pattern, as it is directed primarily at middle and even high skill occupations (Autor, 2015; Webb, 2020). In the context of this study, a natural question to ask is whether high-skilled and low-skilled workers respond differently to different types of AI assistance, and who is at greater risk of "falling asleep". The interaction of technology with complementary assets such as human capital is of fundamental importance for firms and teams (Teodoridis, 2017; Choudhury et al., 2020), and is particularly relevant in the context of AI.

Columns 3 and 4 compare the subsample with low HR experience (column 3) and the subsample with high HR experience (column 4).[26] Unsurprisingly, the majority (72%) of HR recruiters report having high HR experience. Here the comparison shows interesting heterogeneities as in both tables more experienced subjects performed worse than less experienced ones when assigned to "Perfect prediction" or "Good AI". In contrast, the high HR experience subgroup performed better when they were assigned to "Bad AI".

Columns 5 and 6 compare the subsample with low AI experience (column 5) with the subsample with high AI experience (column 6).[27] This is about AI experience in the

[26] I defined the subgroup with high HR experience as those who answered "Strongly Agree" or "Agree" to the question "I have considerable experience in the HR/recruiting setting". Results are consistent when I use alternative categories for HR experience.

[27] I defined the subgroup with high AI experience as those who answered "Strongly Agree" or "Agree"

recruiting setting, and only a minority (30%) of recruiters report high AI experience. The effects are very similar to those of columns 3 and 4. Regardless of whether experience comes from a background in HR or from a familiarity with AI technologies; greater experience leads to better performance when subjects are assigned to "Bad AI", and to worse performance when subjects are assigned to "Good AI" but especially "Perfect prediction". Experienced recruiters were, in general, more likely to think independently and not to exclusively follow the AI's advice. This behavior was beneficial for screeners assigned to the low-quality AI, and drove much of the positive effect of "Bad AI", but proved detrimental when collaborating with perfect algorithms.

Table 9 and Table 10 focus on heterogeneity in the treatment effects presented in Table 5. Columns 1 and 2 compare the time spent and the effort exerted, proxied by the number of clicks, in the first part of the task (1) versus the second part (2). The results show that across conditions recruiters spent more time on the first half of their assigned job applications than on the second half. This difference did not impact performance, suggesting recruiters may have become more tired over time, but also learned to perform their task more efficiently. In columns 3-4 and columns 5-6, we see that more experienced recruiters spent more time but not more clicks or effort on their evaluations in comparison to less experienced recruiters. This is especially true for recruiters with AI experience, who spent much longer on evaluations than their less experienced counterparts when they were assigned to "Bad AI". In contrast, they spent less time on evaluations when assigned to the other conditions than less experienced recruiters did. Overall, more experienced recruiters drove much of the positive effect of "Bad AI", and the negative effect of the other AI conditions, on time spent.

7 Discussion

Although the main findings of this paper may seem counterintuitive, ideas related to those presented here have a long lineage in the behavioral sciences. The broader concept

to the question "I have prior experience with AI advice in a recruiting setting". Results are consistent when I use alternative categories for AI experience.

that coarsening information can lead to better decision-making has been analyzed from multiple angles. One good example of this is Goldin and Rouse (2000) "blind" orchestra auditions that removed gender as a factor in the decision-making process; which led to less biased decisions. (Bertrand and Mullainathan, 2004) found that decision-making was impaired when potential employers were faced with almost identical CVs (except for the applicant name) and showed a tendency to reject those whose names sounded more African-American. The idea is not that names on CVs are necessarily useless, but rather that, by removing names decision-makers will focus on more relevant parts of the CV, improving their judgment overall. In the instance discussed in this paper, more imprecise information elicits a response in human behavior and affects overall performance.

A natural follow-on question becomes, "What type of AI tools would be most effective in human/AI interactions". The results of this paper emphasize that human/AI collaboration should be designed with the goal of keeping humans attentive in tasks where their focus is necessary to improve performance. Rather than concentrate on simply "lower-quality AI", organizations may want to develop "custom AI", tailored to ensure that human collaborators remain engaged in their tasks.

The experimental evidence discussed here also sheds some light on the role of experts in human/AI collaborations. The positive effect of receiving "Bad AI" was almost entirely driven by more expert recruiters in the experiment. The same group performed worse than non-experts when receiving ground truth or "Perfect Prediction". This shows that experts are more likely to deviate from AI's recommendations, and when they do, are more likely to add valuable input than non-experts would.

In a world of automation, where machines can simply substitute humans in their tasks, experts may become a sluggish rearguard opposing change. However, in a world of augmentation, experts' habitual tendency to remain in the loop, and use their skills to contribute to their task, proves to be fundamental when interacting with imperfect AI. This adds nuance to the discussion around skill-biased technical change with respect to AI. High-skilled workers may benefit more from AI, especially from AI that is customized to keep them awake by potentially performing suboptimally. The attention of low-skilled workers is less valuable, and their complementarity with AI less clear.

8 Concluding Remarks

The experiment presented in this paper tests human/AI collaboration in a controlled environment and shows that AI assistance that is too precise leads humans to "fall asleep at the wheel"; becoming more reliant on AI and less engaged in their work efforts. Furthermore, maximizing the performance of a human using an AI does not equal maximizing AI performance in isolation. In fact, maximizing human/AI performance may require a lower quality AI. This is dependent on the effort, learning, and skills of the humans involved.

It is important to note that "falling asleep" is an endogenous human behavior that AI developers can incorporate into the design of tools for human/AI collaboration. There may be some design solutions that could attenuate this tendency, but they can be implemented only once the potential problem has been clearly defined. For example, stronger incentives may have an effect in attenuating the tendency to "fall asleep". Behavioral interventions that make tasks more engaging may have similar effects.

Something that easily generalizes from the results of this paper is that the details of human/AI collaboration matter, and these should be taken into account when discussing automation within firms. More broadly, this suggests moving beyond a form of AI determinism, where we exclusively emphasize the technical capabilities and characteristics of new technologies and focus instead on behavioral responses, managerial decisions, and strategic choices.

References

Acemoglu, Daron, "Technical change, inequality, and the labor market," *Journal of economic literature*, 2002, *40* (1), 7–72.

_ **and Pascual Restrepo**, "Automation and new tasks: How technology displaces and reinstates labor," *Journal of Economic Perspectives*, 2019, *33* (2), 3–30.

Agan, Amanda, Bo Cowgill, and Laura Gee, "The Effects of Salary History Bans: Evidence from a Field Experiment," *Working paper*, 2021.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Press, 2018.

_ **,** _ **,** **and** _ , "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction," *Journal of Economic Perspectives*, 2019, *33* (2), 31–50.

Athey, Susan, Kevin Bryan, and Joshua Gans, "The allocation of decision authority to human and artificial intelligence," *AEA Papers and Proceedings*, 2020, *110*, 80–84.

Autor, David, "Why are there still so many jobs? The history and future of workplace automation," *Journal of economic perspectives*, 2015, *29* (3), 3–30.

Autor, David H and David Dorn, "The growth of low-skill service jobs and the polarization of the US labor market," *American Economic Review*, 2013, *103* (5), 1553–97.

Barach, Moshe, Aseem Kaul, Ming D. Leung, and Sibo Lu, "Strategic redundancy in the use of big data: Evidence from a two-sided labor market," *Strategy Science*, 2019, *4* (4), 298–322.

Beane, Matthew, "Shadow learning: Building robotic surgical skill when approved means fail," *Administrative Science Quarterly*, 2019, *64* (1), 87–123.

Benson, Alan, Danielle Li, and Kelly Shue, "Promotions and the Peter principle," *The Quarterly Journal of Economics*, 2019, *134* (4), 2085–2134.

Bertrand, Marianne and Sendhil Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American economic review*, 2004, *94* (4), 991–1013.

Bessen, James E., "How computer automation affects occupations: Technology, jobs, and skills," *Boston Univ. school of law, law and economics research paper*, 2016, pp. 15–49.

Brynjolfsson, Erik and Andrew McAfee, *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*, WW Norton Company, 2014.

_ **and Lorin M. Hitt**, "Beyond computation: Information technology, organizational transformation and business performance," *Journal of Economic perspectives*, 2000, (14, no. 4), 23–48.

_ , **Daniel Rock, and Chad Syverson**, "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics," *NBER Working Paper #24001*, nov 2017.

_ , _ , **and** _ , "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, 2021, *1* (13), 333–72.

Choudhury, Prithwiraj, Evan Starr, and Rajshree Agarwal, "Machine learning and human capital complementarities: Experimental evidence on bias mitigation," *Strategic Management Journal*, 2020, *41* (8), 1381–1411.

Cowgill, Bo, "Bias and productivity in humans and algorithms: Theory and evidence from resume screening," *Columbia Business School Working Paper*, 2017.

_ **and Patryk Perkowski**, "Delegation in Hiring: Evidence from a Two-Sided Audit," *Columbia Business School Research Paper*, 2020, (898).

_ , **Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau**, "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics," *Proceedings of the 21st ACM Conference on Economics and Computation*, 2020, pp. 679–681.

Dafoe, Allan, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel, "Cooperative AI: machines must learn to find common ground," *Academy of Management Annals*, 2021, *593*, 33–36.

Dawes, Robyn M, "The robust beauty of improper linear models in decision making," *American psychologist*, 1979, *34* (7), 383–397.

Dell'Acqua, Fabrizio, Bruce Kogut, and Patryk Perkowski, "Super Mario Meets AI: The Effects of Automation on Team Performance and Coordination in a Videogame Experiment," *Columbia Business School Research Paper*, 2021.

Deming, David J., "The Growing Importance of Social Skills in the Labor Market," *The Quarterly Journal of Economics*, 2017, *132* (4), 1593–1640.

Dietvorst, B. J., J. P. Simmons, and C. Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114–126.

Dietvorst, Berkeley J, Joseph P. Simmons, and Cade Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, 2018, *64* (3), 1155–1170.

Frank, Morgan R., David Autor, James E. Bessen, Erik Brynjolfsson, Manuel Cebriana, David J. Deming, Maryann Feldman, Matthew Groh, Jose Lobo, Esteban Moro, Dashun Wang, Hyejin Youn, and Iyad Rahwan, "Toward understanding the impact of artificial intelligence on labor," *Proceedings of the National Academy of Sciences*, 2019, *116* (14), 6531–6539.

Glaeser, Edward, Andrew Hillis, Hyunjin Kim, Scott Duke Kominers, and Michael Luca, "Decision Authority and the Returns to Algorithms," 2021.

Glikson, Ella and Anita Williams Woolley, "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, 2020, *14* (2), 627–660.

Goldin, Claudia and Cecilia Rouse, "Orchestrating impartiality: The impact of" blind" auditions on female musicians," *American economic review*, 2000, *90* (4), 715–741.

_ **and Lawrence F. Katz**, "The origins of technology-skill complementarity," *Quarterly Journal of Economics*, 1998, *3* (4), 693–732.

Hansen, Morten T, "The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits," *Administrative science quarterly*, 1999, *44* (1), 82–111.

Hanushek, Eric, Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann, "Returns to skills around the world: Evidence from PIAAC," *European Economic Review*, 2015, (73), 103–130.

Henderson, Rebecca, "The Innovator's Dilemma as a Problem of Organizational Competence," *Journal of Product Innovation Management*, 2006, (23), 5–11.

Jones, Steven E., *Against technology: From the Luddites to neo-Luddism*, Routledge, 2013.

Kambhampati, Subbarao, "Polanyi's revenge and AI's new romance with tacit knowledge," *Communications of the ACM*, 2021, *64* (2), 31–32.

Kogut, Bruce and Udo Zander, "Knowledge of the firm, combinative capabilities, and the replication of technology," *Organization science*, 1992, *3* (3), 383–397.

Lebovitz, Sarah, Hila Lifshitz-Assaf, and Natalia Levina, "To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis," *Organization science*, 2021.

Luo, Xueming, Siliang Tong, Zheng Fang, and Zhe Qu, "Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases," *Marketing Science*, 2019, *38* (6), 937–947.

McGowan, Müge Adalet and Dan Andrews, "Labour Market Mismatch and Labour Productivity: Evidence from PIAAC Data," *OECD Publishing*, 2015, (1209).

Meehl, Paul E, *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.*, Minneapolis, MN: University of Minnesota Press, 1954.

Polanyi, Michael, "The logic of tacit inference," *Philosophy*, 1966, *41* (155), 1–18.

Puranam, Phanish, "Human–AI collaborative decision-making as an organization design problem," *Journal of Organization Design*, 2021, pp. 1–6.

Raisch, Sebastian and Sebastian Krakowski, "Artificial intelligence and management: The automation–augmentation paradox," *Academy of Management Review*, 2021, *46* (1), 192–210.

Raj, Manav and Robert Seamans, "Primer on artificial intelligence and robotics," *Journal of Organization Design*, 2019, *8* (1), 1–14.

Reagans, Ray and Bill McEvily, "Network structure and knowledge transfer: The effects of cohesion and range," *Administrative science quarterly*, 2003, *48* (2), 240–267.

Schleicher, Andreas, "PIAAC: A new strategy for assessing adult competencies," *International Review of Education*, 2008, *54* (5-6), 627–650.

Simon, Herbert A., *The shape of automation for men and management*, vol 13 ed., Harper and Row, 1965.

Syverson, Chad, "Challenges to mismeasurement explanations for the US productivity slowdown," *Journal of Economic Perspectives*, 2017, *31* (2), 165–86.

Teodoridis, Florenta, "Understanding team knowledge production: The interrelated roles of technology and expertise," *Management Science*, 2017, pp. 3625–3648.

Tong, Siliang, Nan Jia, Xueming Luo, and Zheng Fang, "The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance," *Strategic Management Journal*, 2021.

Webb, Michael, "The impact of artificial intelligence on the labor market," *Working paper*, 2020.

Weidmann, Ben and David J. Deming, "Team Players: How Social Skills Improve Group Performance," *National Bureau of Economic Research*, 2020, *w27071.*

Zander, Udo and Bruce Kogut, "Knowledge and the speed of the transfer and imitation of organizational capabilities: An empirical test," *Organization science*, 1995, *6* (1), 76–92.

Table 1: **Summary Statistics**

Panel A: Overall Sample

	Total
Female	74%
Male	24%
Asian	7%
Black	14%
Latinx	9%
White	60%
Uni Degree	76%
US-based	89%
Rate ($ per hr)	40.25
High HR Exp	72%
High AI Exp	30%
Total	181

Panel B: By Treatment Assignment

	Perfect pred.	Good AI	Bad AI	Control
Female	76%	69%	78%	73%
Male	22%	29%	22%	22%
Asian	11%	2%	7%	7%
Black	9%	17%	11%	17%
Latinx	11%	9%	9%	7%
White	60%	58%	63%	58%
Uni Degree	78%	76%	72%	80%
US-based	83%	89%	93%	91%
Rate ($ per hr)	40.05	39.79	41.25	39.9
High HR Exp	76%	73%	65%	73%
High AI Exp	36%	27%	26%	33%
Total	45	45	46	45

Notes: This table displays descriptive statistics for my sample. Panel A displays descriptive statistics for the overall sample, while Panel B displays them by experimental treatment assignment.

Table 2: **Impact on following AI's suggestions**

	Follows AI
Perfect prediction	0.849***
	(0.015)
Good AI	0.739***
	(0.008)
Bad AI	0.531***
	(0.014)
Controls	Y
R2	.544
Observations	5184

Notes: This table examines the effect of the AI treatments on following AI's advice. It shows experimental compliance. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 3: **Impact on Accuracy - AI**

	(1)	(2)	(3)	(4)
	Accuracy (binary)	Accuracy (binary)	Accuracy	Accuracy
AI	0.059***	0.061***	0.375***	0.386***
	(0.012)	(0.013)	(0.104)	(0.102)
Controls	Y	Y	Y	Y
Exp+Baseline		Y		Y
R2	.407	.408	.479	.482
Observations	5184	5184	5184	5184
Control Mean	0.723	0.723	7.08	7.08

Notes: This table examines the effect of all AI treatments combined (compared with a control of no AI) on accuracy. Columns (1) and (3) report the results of regressions with a set of recruiters' characteristics as controls. Columns (2) and (4) also include recruiters' HR experience, AI experience and initial accuracy. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 4: **Impact on Accuracy by AI Quality**

	(1) Accuracy (binary)	(2) Accuracy (binary)	(3) Accuracy	(4) Accuracy
Perfect prediction	0.125***	0.129***	0.762***	0.787***
	(0.018)	(0.019)	(0.130)	(0.126)
Good AI	0.021	0.023	0.079	0.103
	(0.014)	(0.015)	(0.116)	(0.111)
Bad AI	0.031**	0.034**	0.292**	0.314**
	(0.015)	(0.015)	(0.128)	(0.123)
Controls	Y	Y	Y	Y
Exp+Baseline		Y		Y
R2	.416	.417	.487	.489
Observations	5184	5184	5184	5184
Control Mean	0.723	0.723	7.08	7.08
P-values				
Good AI=Bad AI	0.52	0.51	0.06	0.05

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table examines the effect of the AI treatments (compared with a control of no AI) on accuracy. Columns (1) and (3) report the results of regressions with a set of recruiters' characteristics as controls. Columns (2) and (4) also include recruiters' HR experience, AI experience and initial accuracy. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 5: **Impact on Time and Effort**

	(1) Time Spent	(2) Time Spent	(3) Effort (clicks)	(4) Effort (clicks)
Perfect pred.	-1.375	-1.434	0.057	-0.109
	(1.705)	(1.544)	(0.446)	(0.232)
Good AI	-1.436	-1.533	0.170	0.052
	(1.733)	(1.638)	(0.394)	(0.180)
Bad AI	10.034***	8.766***	0.656	0.198
	(2.838)	(2.911)	(0.425)	(0.206)
Controls	Y	Y	Y	Y
Exp+Baseline		Y		Y
R2	.0592	.071	.125	.561
Observations	5184	5184	5184	5184
Control Mean	20.2	20.2	8.06	8.06
P-values				
Good AI=Bad AI	0.00	0.00	0.29	0.49

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table examines the effect of the AI treatments (compared with a control of no AI) on the time subjects spent on each CV (columns 1-2) and on the effort of subjects (columns 3-4). Columns (1) and (3) report the results of regressions with a set of recruiters' characteristics as controls. Columns (2) and (4) also include recruiters' HR experience, AI experience and initial accuracy. *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 6: **IV Regressions**

	(1) Accuracy (binary)	(2) Accuracy	(3) Accuracy (binary)	(4) Accuracy
Follows AI	0.102***	0.600***	-0.041	-1.079*
	(0.017)	(0.136)	(0.078)	(0.591)
Observations	5184	5184	2592	2592

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports IV regressions. Columns 1 and 2 report the effects of the increase in following the AI determined by having received any type of AI. Columns 3 and 4 limit the subsample to subjects receiving Good or Bad AI. They report the effects of the increase in following the AI determined by having received Good AI and not Bad AI *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 7: **Heterogeneity by timing, HR experience and AI experience (Accuracy - binary)**

	(1) Acc (bin)	(2) Acc (bin)	(3) Acc (bin)	(4) Acc (bin)	(5) Acc (bin)	(6) Acc (bin)
Perfect pred.	0.135***	0.120***	0.168***	0.109***	0.143***	0.079**
	(0.021)	(0.022)	(0.030)	(0.023)	(0.023)	(0.037)
Good AI	0.019	0.025	0.025	0.010	0.023	0.006
	(0.019)	(0.020)	(0.038)	(0.017)	(0.017)	(0.027)
Bad AI	0.033*	0.031	0.004	0.031*	0.019	0.046**
	(0.020)	(0.021)	(0.026)	(0.017)	(0.017)	(0.022)
Controls	Y	Y	Y	Y	Y	Y
R2	.429	.418	.405	.437	.414	.456
Observations	2592	2592	1568	3616	3680	1504
Control Mean	0.709	0.738	0.699	0.732	0.719	0.729

Notes: This table compares the treatment effects dividing the sample in three different ways. Columns 1-2 compare the accuracy for the first part of the task (1), with accuracy in the second part (2). Columns 3-4 compare the subsample with low HR experience (3) with the subsample with high HR experience (4). Columns 5-6 compare the subsample with low AI experience (5) with the subsample with high AI experience (6) *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 8: **Heterogeneity by timing, HR experience and AI experience (Accuracy)**

	(1) Accuracy	(2) Accuracy	(3) Accuracy	(4) Accuracy	(5) Accuracy	(6) Accuracy
Perfect pred.	0.794***	0.751***	0.862***	0.724***	0.854***	0.462*
	(0.137)	(0.162)	(0.276)	(0.146)	(0.157)	(0.245)
Good AI	0.069	0.098	0.067	0.060	0.121	0.052
	(0.117)	(0.157)	(0.364)	(0.109)	(0.148)	(0.161)
Bad AI	0.295*	0.302*	0.115	0.292**	0.204	0.362*
	(0.150)	(0.155)	(0.271)	(0.135)	(0.135)	(0.212)
Controls	Y	Y	Y	Y	Y	Y
R2	.498	.490	.479	.509	.490	.519
Observations	2592	2592	1568	3616	3680	1504
Control Mean	7.00	7.17	6.89	7.16	7.05	7.15

Notes: This table compares the treatment effects dividing the sample in three different ways. Columns 1-2 compare the accuracy for the first part of the task (1), with accuracy in the second part (2). Columns 3-4 compare the subsample with low HR experience (3) with the subsample with high HR experience (4). Columns 5-6 compare the subsample with low AI experience (5) with the subsample with high AI experience (6). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 9: **Heterogeneity by timing, HR experience and AI experience (Time Spent)**

	(1) Time Spent	(2) Time Spent	(3) Time Spent	(4) Time Spent	(5) Time Spent	(6) Time Spent
Perfect pred.	-0.545	-2.165	-0.031	-1.959	-1.241	-4.660
	(2.068)	(1.846)	(2.518)	(2.254)	(2.224)	(4.259)
Good AI	-0.807	-2.020	-0.503	-3.256	-1.133	-2.541
	(1.970)	(1.788)	(3.108)	(2.499)	(1.905)	(4.939)
Bad AI	12.470***	7.782***	9.203*	10.869***	7.014**	16.027**
	(3.692)	(2.959)	(4.931)	(3.733)	(2.725)	(6.061)
Controls	Y	Y	Y	Y	Y	Y
R2	.070	.061	.123	.064	.072	.097
Observations	2592	2592	1568	3616	3680	1504
Control Mean	20.32	20.09	18.32	20.89	19.14	22.25

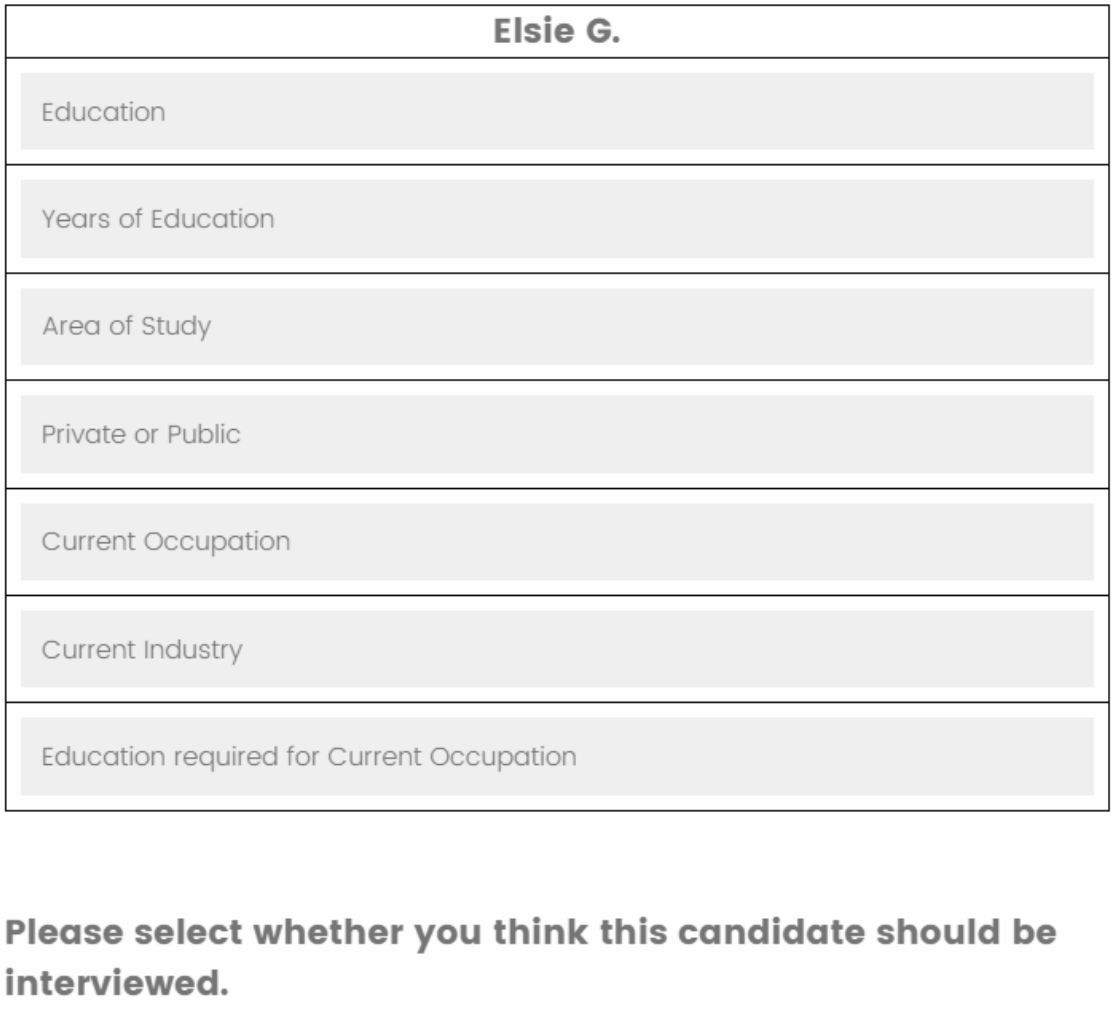
Notes: This table compares the treatment effects dividing the sample in three different ways. Columns 1-2 compare the accuracy for the first part of the task (1), with accuracy in the second part (2). Columns 3-4 compare the subsample with low HR experience (3) with the subsample with high HR experience (4). Columns 5-6 compare the subsample with low AI experience (5) with the subsample with high AI experience (6). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 10: **Heterogeneity by timing, HR experience and AI experience (Effort - Clicks)**

	(1) Effort	(2) Effort	(3) Effort	(4) Effort	(5) Effort	(6) Effort
Perfect pred.	0.155	-0.021	0.073	0.070	0.406	-0.864
	(0.477)	(0.439)	(0.770)	(0.502)	(0.496)	(1.074)
Good AI	0.214	0.123	0.040	0.079	0.136	0.598
	(0.424)	(0.375)	(1.119)	(0.486)	(0.470)	(0.933)
Bad AI	0.610	0.714	0.947	0.860	0.770*	0.544
	(0.431)	(0.437)	(0.963)	(0.533)	(0.447)	(1.007)
Controls	Y	Y	Y	Y	Y	Y
R2	.147	.117	.286	.141	.168	.184
Observations	2592	2592	1568	3616	3680	1504
Control Mean	8.07	8.05	8.36	7.96	7.93	8.32

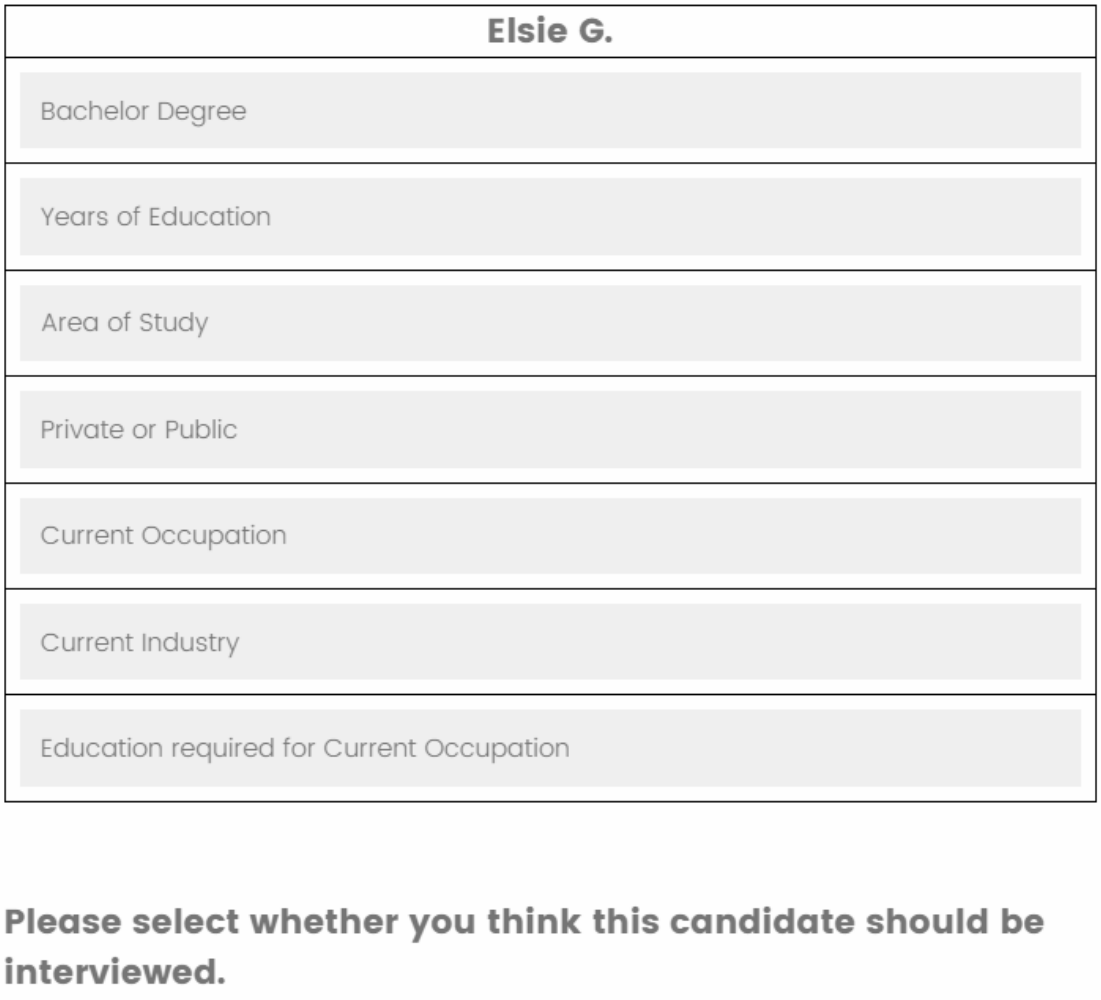
Notes: This table compares the treatment effects dividing the sample in three different ways. Columns 1-2 compare the accuracy for the first part of the task (1), with accuracy in the second part (2). Columns 3-4 compare the subsample with low HR experience (3) with the subsample with high HR experience (4). Columns 5-6 compare the subsample with low AI experience (5) with the subsample with high AI experience (6). *** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Figure 1: **A candidate's job application**



Notes: This figure displays a screenshot from the platform when one of the job applications is presented and before any button is clicked.

Figure 2: **The same job application after the first button is clicked**

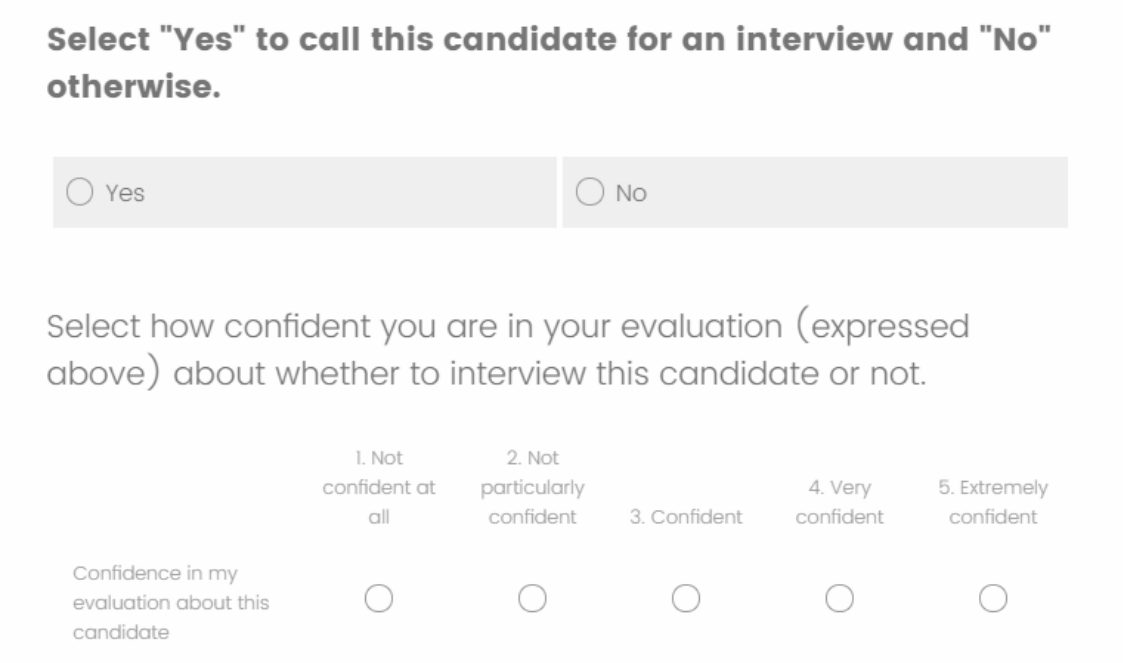


Notes: This figure displays a screenshot from the platform when one of the job applications is presented and the "Education" button is clicked.

Figure 3: **Experimental Structure**

Notes: This figure provides a visualization of the structure of the experiment.

Figure 4: **The recruiters' choices**



Notes: This figure displays a screenshot showing the choice recruiters made about each candidate. They selected whether to interview a candidate, and also how confident they were in their decision.

Appendix

A Survey Questions (1-7 Likert scale)

1. *I believe my assessments of candidates were accurate*
2. *I enjoyed working on this task*
3. *I exerted maximum effort in this task*
4. *I investigated all available information to make correct decisions about candidates*
5. *I felt personally responsible for candidates' callback decisions*
6. *By the end of the task I could understand better which candidates were more likely above the threshold*
7. *I believe the great majority of decisions I made in this task were correct*
8. *I prefer my tried and trusted technologies and products*
9. *I have considerable experience in the HR/recruiting setting*
10. *I have prior experience with AI advice in a recruiting setting*
11. *I am familiar with AI and machine learning tools in multiple contexts*
12. *I work in an industry where AI is frequently used*
13. *I trusted AI's recommendations*
14. *I ignored AI's advice*
15. *Overall, I thought AI's recommendations were unhelpful*
16. *I believe I can improve on the scores produced by AI*
17. *I would like to use the AI assistance again for a similar task in the future*
18. *I knew when to follow what the AI-recommended score suggested and when not to follow it*
19. *Over time, I learned to incorporate AI's recommendations more successfully*
20. *I enjoy the freedom to adopt my own approach to the job*
21. *I have been paying attention during this task*
22. *I enjoy control over the quality of my work*

B Demographic Questions

1. *Your racial/ethnic background is: (select all that apply) White, Black or African American, Hispanic or Latinx, American Indian or Alaska Native, Asian, Native Hawaiian or Pacific Islander, Other*

2. *What is your gender? (Male, Female, Prefer not to say, other)*

3. *Please select your level of education from the list below (Less than High School, High School Graduate,..., Professional Degree, Doctorate)*

C Job posting

I report below my job ad for HR recruiters:

> We are an independent consultancy lead by a group of former executives from large regional companies. We focus on HR, workforce, and personnel issues in our work with clients. We have contracts with Fortune 1000 companies across different industries.
>
> We are currently working with a large organization that hires lots of new workers every year. We received a large number of applications for them, and we would like to have your support in selecting the best candidates.
>
> We have developed a simple platform to perform the initial screening quickly and accurately. You can find attached our task instructions.
>
> You will receive a packet of about 40 CVs, and we would like you to select which candidates should be called for an interview. We developed a simple online platform to manage the CVs. We would also like to ask you a few questions to help us improve our platform.
>
> Note that we don't have lots of information about each candidate, so just do your best.
>
> Information about the candidates will be strictly confidential.
>
> For your assistance with this task, you will be paid hourly, plus a bonus. We expect this would take about one hour.
>
> If you are interested, we might contact you for related tasks in the future.
>
> Let us know if you're available and interested!

D Instructions - AI Treatments

I report below some sections of the instructions that recruiters received.

D.1 Math Competency

Here the instructions clarified that recruiters' goal was to select candidates based on math competency:

> (...)When you follow the link we send you, we will show you some job applicants and data from their job application. We have a broad range of candidates around the math quality level we are seeking, featuring a variety of backgrounds and qualifications. We are showing you only a subset of their job qualifications.
>
> **Please help us identify the candidate which would have the best math competency.**
>
> Insofar as we have any other hiring goals – for example, other skills besides math, or location preferences – we will incorporate these through a separate process.
>
> **Help identify the candidates who are best at math.**

D.2 Algorithm Quality

Recruiters who collaborated with AI received a different description of the algorithm assisting them based on their treatment assignment.

"Perfect" prediction

> The AI tool that will support you has been performing extremely well in prior analysis and we have been very pleased with the candidates selected.
>
> We reviewed the algorithm's recommendations using performance data, and we found that the recommendations about whether to interview a candidate or not were almost always correct (more than 99% of instances).

"Good" AI

> The AI tool that will support you has been performing very well in prior analysis and we have been very pleased with the candidates selected. However, it made a few mistakes for candidates that were close calls.

We reviewed the algorithm's recommendations using performance data, and we found that the vast majority of AI's recommendations about whether to interview a candidate or not were correct (about 85% of cases).

"Bad" AI

The AI tool that will support you has been performing well in prior analysis and we have been pleased with the candidates selected. However, it made some mistakes for candidates that were close calls.

We reviewed the algorithm's recommendations using performance data, and we found that the large majority of AI's recommendations about whether to interview a candidate or not were correct (about 75% of cases).

E Feedback

This is the feedback provided to recruiters after each one of their first 12 iterations. Based on their answer, they saw one of the following four messages.

Positive feedback after the decision to interview

Based on our assessments, your selection about this candidate was correct. We believe **this candidate should be interviewed**.

Positive feedback after the decision not to interview

Based on our assessments, your selection about this candidate was correct. We believe **this candidate should not be interviewed**.

Negative feedback after the decision to interview

Based on our assessments, your selection about this candidate was not correct. We believe **this candidate should not be interviewed**.

Negative feedback after the decision not to interview

Based on our assessments, your selection about this candidate was not correct. We believe **this candidate should be interviewed**.