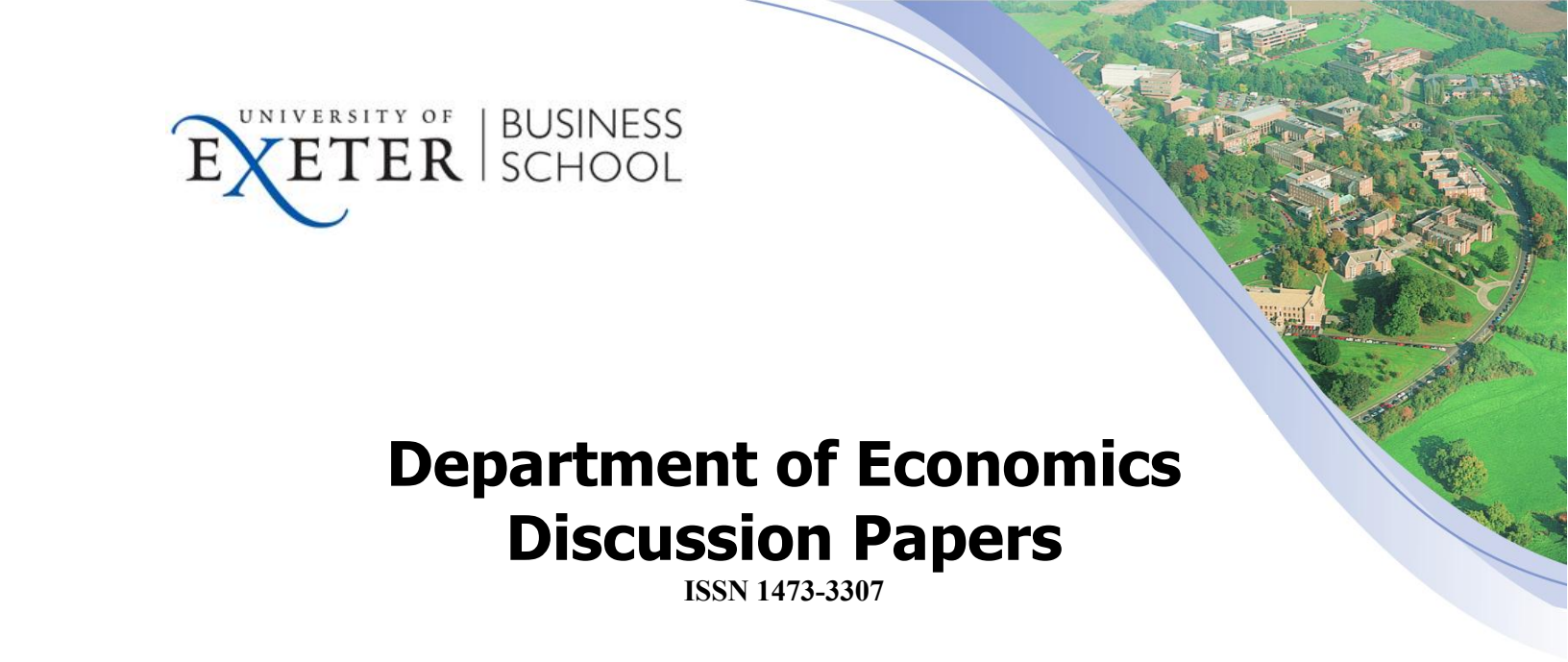


Department of Economics Discussion Papers

ISSN 1473-3307

A Brave New World of Hiring: A Natural Field Experiment on How Asynchronous Interviews and AI Assessment Reshape Recruitment

Mallory Avery, Edwin Ip, Andreas Leibbrandt, and Joseph Vecci

Paper number 26/02

Disclaimer: The views expressed in this discussion paper are those of the author(s) and do not necessarily reflect those of the University of Exeter, its Business School, or the Department of Economics. Discussion Papers are circulated for discussion and comment purposes and have not been peer reviewed.

RePEc page: https://ideas.repec.org/s/exe/wpaper.html

A Brave New World of Hiring: A Natural Field Experiment on How Asynchronous Interviews and AI Assessment Reshape Recruitment*

This version: 25 March 2026

Mallory Avery†, Edwin Ip‡, Andreas Leibbrandt§, Joseph Vecci**

Recent technological advancements are reshaping pathways to employment by automating the interview process. Asynchronous interviews, in which job applicants submit answers to interview questions via an online platform without interacting with an interviewer, are replacing more traditional face-to-face job interviews. At the same time, AI algorithms are now widely used to assess these interview answers. In this paper, we use a field experiment to comprehensively study how these new technologies affect applicants and employers in the recruitment process. Over 3,000 job applicants are randomized into asynchronous audio or video interviews, live online interviews, and a control group. Their job interviews are then assessed by both professional recruiters and a commercial AI recruitment tool used by most Fortune 100 companies. We find that asynchronous interviews cause an over 50% decrease in application continuation, including among the most qualified applicants, and that this decline is largest for women. A complementary vignette experiment provides evidence that this deterrence is driven by perceptions about the competitiveness and fairness of the recruitment process. In terms of assessments, we find that the AI evaluation tool scores women and underrepresented racial minorities higher than human evaluators, while the opposite is true for men, Whites and Asians. We track our applicants' subsequent labor market outcomes and find that the AI assessment tool predicts subsequent employment success substantially better than human recruiters, suggesting that AI captures soft skills and potential that humans overlook. In addition, we provide evidence that, unlike AI, human recruiters' assessments suffer from multiple cognitive biases. Our findings provide some of the first key evidence on how recent technological advances are transforming the hiring process.

Key Words: Technological Change, Artificial Intelligence, Gender, Field Experiment
JEL codes: C93, J23, J71, J78

---

* We thank Alexander Cappelen, Mitch Hoffman, Paula Scholz, Josh Liff and participants at conferences including AFE 2025, ANZWEE 2025, BEE-UK 2025, CSEE 2025, SABE 2025, SIOE 2025, and the Economics Network's Symposium on Ethical Use of AI in Economics, and at seminars at His Majesty's Treasury, University of Arizona, Florida International University, University of Exeter, University of Gothenburg, University of Heidelberg, University of Insubria, University of Konstanz, Norwegian School of Economics, University of Technology Sydney, University of Venice and University of Western Australia for their helpful feedback. The study was pre-registered at the AEA registry (AEARCTR- 0013356) and received funding from the Australian Research Council project ID FT190100298.
† E-mail: Mallory.avery@monash.edu, Monash University, Melbourne, Australia
‡ E-mail: e.ip@exeter.ac.uk, University of Exeter, United Kingdom
§ Corresponding author. E-mail: andreas.leibbrandt@monash.edu, Monash University, Melbourne, Australia
**E-mail: joseph.vecci@economics.gu.se, University of Gothenburg, Sweden

1. Introduction

For much of human history, whether in ancient guilds, Roman apprenticeships, or early religious communities, access to jobs was typically determined through kinship ties, personal recommendations, and local reputation (e.g., de la Croix et al., 2018). The twentieth century marked a major shift in how employees were selected. The development of large-scale firms and industrial psychology gave rise to the structured job interview (McDaniel et al, 1994; Autor, 2001). Their widespread use has continued into the twenty-first century despite being costly and time-intensive, and despite the fact that a growing body of research shows that human evaluators are often inconsistent, overconfident, and vulnerable to cognitive biases (Kausel et al., 2016; Gabaix, 2019; Hirshleifer et al., 2019).[1] These limitations pose fundamental questions about their efficiency and fairness, and raise concerns that traditional job interviews fail to reliably identify high-ability workers and disadvantage minority candidates.

After a century of relative stability in interview practices, automated hiring processes, such as the use of digital platforms that conduct job interviews asynchronously and AI that evaluates candidates, are being adopted at an unprecedented pace and have become standard tools in the recruitment processes of most large firms.[2] These technologies fundamentally alter when interviews occur, what information is collected from candidates, and how that information is processed and evaluated. However, such potentially efficiency-maximizing technologies also have the potential to substantially alter recruitment outcomes, both in terms of the quality of job-candidate matches and the composition of selected candidates across demographic and skill groups.[3] Understanding how this rapid technological shift in the application process affects labor market outcomes and the allocation of talent has therefore become a central empirical question.

---

[1] Many employers estimate that filling a vacancy can cost three to four times the position's salary (Navarra, 2022).

[2] Recent estimates indicate that more than half of U.S. firms now incorporate AI-based tools into their recruitment processes (Jaser et al., 2022, Brookings, 2025), with usage rates rising to 94–98% among Fortune 500 companies (Brookings, 2025; Fuller et al., 2021). This trend extends beyond the private sector – Nawrat (2023) reports that eight of the ten largest U.S. federal agencies also rely on algorithmic screening for some roles, highlighting the growing role of AI in shaping access to employment opportunities across sectors.

[3] Related research on algorithmic bias finds that these systems replicate and amplify human biases against demographic groups (e.g. Nugent and Scott-Parker, 2022; Carvajal et al., 2025; Wilson and Caliskan, 2025), but may also improve the quality and diversity of hiring decisions (Cowgill, 2020; Raghavan et al, 2020; Awad et al., 2023; Avery et al., 2024; Ip, 2025; Li et al., 2025). However, little is known about whether AI evaluations mitigate or amplify other cognitive limitations associated with human-based assessments: human evaluators can be inconsistent,

Against this backdrop of rapid technological development in candidate screening, our paper provides the first systematic field experimental evidence on how automating the interview process, through both asynchronous interviewing and AI-powered interview evaluation, is reshaping access to employment. We partner with one of the world's major recruitment platforms, used by over 60% of Fortune 100 companies at the time, to randomize more than 3,000 real applicants for technology jobs into asynchronous audio and video interviews, live synchronous online interviews, or a control condition without an interview requirement. This design allows us to causally identify how this new interviewing technology influences who continues in the hiring process. Completed interviews are subsequently evaluated independently by professional recruiters and by a leading commercial AI-assessment tool, allowing us to directly compare human and algorithmic judgments of the same applicant pool. To assess the predictive validity of the evaluation methods, we assess whether the evaluation scores predict labor market outcomes one year after their initial job application. By jointly studying the complementary automation of both the collection of interviews and their evaluation, this design allows us to open the black box of this "brave new world" of recruitment and evaluate how emerging technologies are beginning to redefine who approaches – and who passes – the employment gates.

We find that the asynchronous interviews reduce application continuation by around 53% (45pp), whereas a more traditional synchronous online interview sees continuation decrease by only 20% (17pp), compared to the control of no interview screening. Further, we find that asynchronous interviews significantly deter the most qualified applicants compared to both the control and synchronous online interviews treatments, and that women are 5.1pp less likely to complete asynchronous interviews than men. To capture underlying mechanisms of the overall deterrence of the asynchronous interview, we conduct a separate vignette experiment with more than 600 participants who show interest in the advertised jobs. We find evidence that the deterrence effect is driven by beliefs that more candidates are invited to asynchronous interviews than synchronous interviews – leading to a perception of increased competition – as well as perceived unfairness of the process.

overconfident and suffer from cognitive fatigue (Kausel et al., 2016; Gabaix, 2019; Hirshleifer et al., 2019; Houser, 2019; Mobius et al., 2022), whereas algorithms deliver evaluations instantaneously, apply criteria uniformly, and cannot suffer from fatigue.

In terms of assessments, we find that the AI assessment tool scores female and underrepresented racial minorities (URMs) higher than human evaluators, whereas human evaluators score men and non-URMs higher than AI.[4] These evaluation differences translate into notably different shortlists: AI-generated shortlists include over 10pp more women and over 7pp more URM candidates compared to human-generated ones, indicating that AI may promote more diverse candidate pools particularly at the upper end of the evaluation spectrum, i.e. the group of candidates most likely to move onto to the next stage of the hiring process. Further, we find that AI scores are generally at least two times more predictive of actual labor market success, defined by job attainment, employment status, and seniority 12 months later, than either human scores or CV-based metrics. To understand these findings, we examine whether cognitive biases can help explain the observed differences in performance. We focus on three key mechanisms and find: i) humans appear to suffer from time of day effects, with both their evaluation scores and the predictive validity of those scores varying systematically over the course of the working day, whereas AI evaluations remain stable; ii) humans tend to be anchored by an applicant's response to the first interview question and score all subsequent questions similarly, whereas AI appears to evaluate each interview question more independently; and iii) humans have a tendency to compress ratings, which reduces their predictiveness, whereas AI assigns scores that are more dispersed.

This paper makes several important contributions to the study of technological change and labor market entry and informs ongoing debates about the efficiency and fairness of rapidly evolving recruitment practices. *First*, we contribute to a nascent literature on the impact of asynchronous interviewing on applicant behavior, particularly interview completion. Most current work is correlational and focuses on candidate perceptions, which are found to be negative though heterogeneous (Brenner, 2016; Hiestra et al., 2019). Two recent papers, Jabarian and Henkel (2025) and Aka et al. (2025), experimentally study whether applicants interviewed by an AI agent in real time had better outcomes.[5] Our focus is on one-way asynchronous interviews and whether

[4] We define underrepresented racial minorities (URMs) as all ethnic groups other than White and Asians, who are considered as racial majorities in the tech industry. In our sample, URMs are primarily African Americans and Hispanic Americans.

[5] These papers consider how those who underwent interviews with an AI agent who interacted with candidates in real time (two-way interviews) get evaluated by humans later in the recruitment process. Jabarian and Henkel (2025) considers how these applicants perform (e.g later in the recruitment process) compared to those who were interviewed by a human agent, whereas Aka et al (2025) considers how they perform compared to those who did not do an interview.

applicants continue in the recruitment process. Using an asynchronous interview platform used by most Fortune 100 companies, we study whether the use of asynchronous interviews deters applicants from continuing compared to the use of more traditional synchronous interviews with humans, compared to a control group that measures the natural continuation rate. This allows us to provide the first causal evidence on how asynchronous interviewing affects applicant attrition. We are also the first to study why the attrition occurs, finding that applicants perceive asynchronous interviewing to be less fair and more competitive than live interviews with humans. This in turn has significant implications on how employers could improve their recruitment process when using asynchronous interviews.

Second, we contribute to the literature on how AI is changing how job applicants are assessed. Most existing work investigates algorithmic bias in AI resume screeners, finding mixed results (Parasurama & Sedoc, 2021; Wang et al., 2024; Zhang & Kuhn, 2024; Wilson & Caliskan, 2025). Only a small set of studies compare human and AI judgments directly (Cowgill, 2017; Hoffman et al., 2018; Avery et al., 2024; Zhang & Kuhn, 2024; Li et al., 2025; Wilson & Caliskan, 2025), but these studies focus on the evaluation of resumes rather than interviews and most do not examine human and AI's relative ability to predict real future labor market outcomes.[6] Unlike this existing literature, we study the evaluation of interviews by AI and Humans, rather than the evaluation of resumes. Thus, we push the envelope on what AI can do, as interviews – unlike résumés – require the evaluation of open-ended responses to infer soft skills and other metrics important to the success of a hired worker but difficult to quantify. Our paper also contributes to this literature by studying for the first time the impact of AI vs. human evaluation of interviews on labor demand and using industry-standard AI models rather than general-use or researcher-generated algorithms. Furthermore, we provide the first investigation into the mechanisms underlying these differences between human and AI interview evaluation, finding that cognitive biases, such as time-of-day effects, anchoring, and rating compression, affect human evaluators and drive the differences in evaluations. These distinctions allow us to provide new evidence on

[6] Li et al. (2025) explore applicant outcomes by simulating how different resume-screening algorithms would have selected applicants to be interviewed using historical job applications and find algorithms would have selected candidates with higher hiring potential compared to those actually selected by human recruiters.

the predictive validity, distributional implications, and behavioral consequences of human versus AI screening.

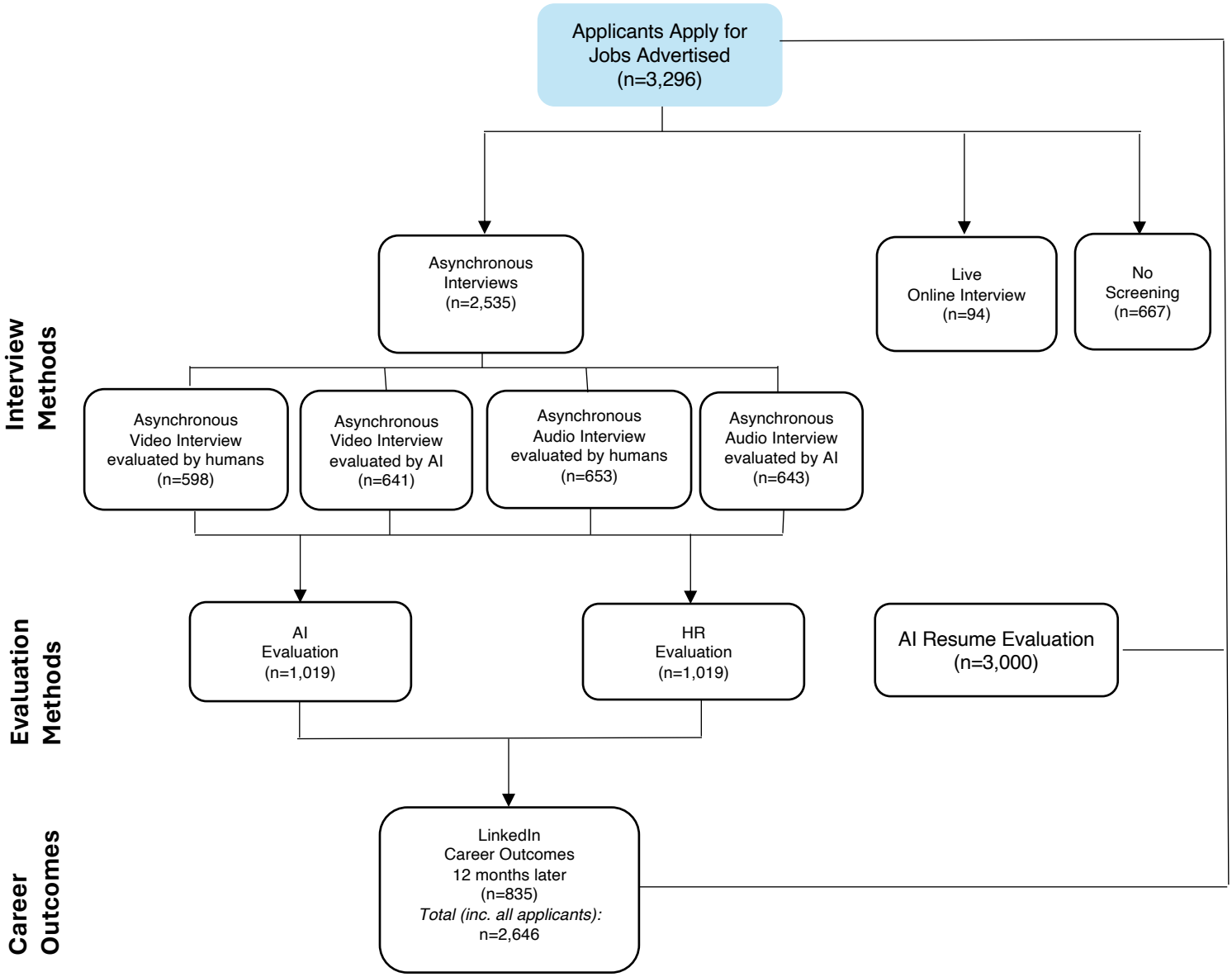
Finally, we contribute to research on how technological changes to the recruitment process affect labor market outcomes for both applicants and recruiters. This research has highlighted the importance of understanding how recruitment procedures, often introduced to improve either the quality or cost of candidate selection, can have unintended and unforeseen consequences both overall and for certain groups. For example, online job application portals have been shown to increase job-finding rates (Fountain, 2005; Suvankulov et al., 2012; Kuhn & Mansour, 2014; Denzer et al., 2021; Zuo, 2021), but also alter labor market participation across demographic groups, particularly married women (Dettling, 2017); employment tests can improve selection, as intended (Bishop, 1988), but can have substantial and mixed impacts on minority group employment outcomes (Neckerman & Kirschenman, 1991; Autor & Scarborough, 2004; Feld et al., 2022; Feld et al., 2023; Amer et al., 2024). In this paper, we are able to capture the impact of a substantial technological shift in recruitment, i.e. the automation of both interviewing and the evaluation of those interviews, (i) on both the behavior of applicants and their outcomes in the recruitment process, (ii) both overall and for underrepresented minority groups, (iii) at the time that the technological change is occurring, rather than retrospectively. This holistic, experimental approach allows us to provide a framework for understanding how this technology is changing who applies and who gets hired, not long after the fact but rather as the change occurs, so policymakers, industry leaders, and academics can make decisions in real time as to how to adapt.

The rest of the paper is organized as follows: Section 2 provides details on our study design. Section 3 discusses the results on applicants' responses to the use of asynchronous interviews as well as our complementary vignette experiment. Section 4 discusses the results on the evaluation of applicants, shortlisting outcomes, and detailed comparisons of human and AI assessment in terms of predictive power of career outcomes as well as the mechanisms behind these differences. Section 5 concludes the study.

2. Study Design

We conduct a natural field experiment to provide a comprehensive, causal study on the impact of automating the interview process on both 1) labor supply (applicant behaviors) and 2) labor demand (interview evaluation outcomes). The study was pre-registered at the AEA registry (AEARCTR- 0013356) and ethics approval was obtained from Monash University. We summarize our whole study design in Figure 1.

Figure 1: Study Design Overview



2.1 Labor Supply: Applicant Behaviors

We first study how asynchronous screening tools impact job application behavior. To do this, on behalf of an organization, we simultaneously advertised three common jobs in the tech sector – Programmer, Web Designer, and Content Creator – across major job boards in the United States for approximately two months starting in April 2024 (see Appendix A for the content of the job ads). To apply for these real jobs, applicants submitted their résumés and completed a short application form (e.g., years of relevant experience, demographics, contact details). Applicants were required to reside in the United States. In total, we received 3,296 applications, including 1,035 for the content creator role, 1,214 for the web developer role, and 1,047 for the programmer.

The advertised positions are concentrated within the tech sector, a vital component of the U.S. economy that accounts for over 10% of national GDP and supports 12.1 million jobs (U.S. Department of Commerce). The three jobs cover a broad spectrum of the tech industry with varying degrees of gender and racial disparities. Our applicants are 21% female, 19% Asian, and 22% URM for the programmer position, 41% female, 17% Asian, and 26% URM for the web designer position, and 71% female, 5% Asian, and 28% URM for the content creator position.

To study applicants' behavioral responses to the use of asynchronous interviews, we randomly assign applicants to one of three treatments: an asynchronous interview, a live online interview, and a control group, with the asynchronous interview treatment consisting of four sub-treatments representing different interview formats.[7] We focus on these formats because they capture the most common forms of interview screening currently used by employers. The randomization was conducted on a rolling basis based on when the candidates applied to the position and was stratified on the basis of gender. Candidates were sent an invitation email that informed them that they had passed the first screening stage. The content of this email and the subsequent outcome variable of interest varied depending on the randomized assignment to treatment. Specifically, applicants are randomly assigned to one of the following conditions:

[7] Applicants were also cross-randomized into either a reminder (1563) or no reminder (1727) treatment. In the reminder treatment candidates were sent a reminder between 5-7 days after the initial invitation email conditional on not having completed the assessment. The goal of this was to evaluate the effect of reminders, a common feature in application processes and a potential tool for increasing completion rates. However, we find no evidence that reminders significantly influence completion rates. Thus, we do not focus on this in our main analysis below.

Control (n=667): in the control condition, the email informed candidates that they had passed the first screening stage. They were then asked to indicate if they were still available and interested in the position by clicking a link. Clicking this link and thus indicating their continued interest is the completion measure in the control condition. This allowed us to assess the proportion of people who may drop out because they had accepted another position or did not read the email. Comparison with the control allows us to measure the effect of having a screening interview stage, either asynchronous or live, with all of the additional costs and changes in beliefs associated with those interviews.

Asynchronous Interview (n=2,535): in the Asynchronous Interview treatment, candidates were invited to participate in a recorded audio or video interview as part of our recruitment process. The email template was based on the invitation email from the Asynchronous Interview provider; as such, our protocol follows standard industry practice. Candidates were then given some details on what to expect, including what the interview involved, the length of time needed and how the interview would be assessed. Candidates were informed that their interview would either be evaluated by AI or by our hiring team.[8] Finally, the email contained a link to access the interview. Thus, within the asynchronous interview treatment we have 4 sub-treatments, a 2×2 design with audio or video recording and human or AI evaluation:

- Recorded video interview evaluated by AI (n=641)
- Recorded video interview evaluated by human reviewers (n=598)
- Recorded audio interview evaluated by AI (n=643)
- Recorded audio interview evaluated by human reviewers (n=653).

Live Online Interview (n=94): Candidates assigned to this treatment were invited to a Zoom interview. To participate the candidate had to schedule a time with the interviewer using a popular scheduling tool. The interview was conducted and recorded via Zoom. The Zoom Interviews were conducted by an interviewer with prior interviewing experience who was blind to the study purpose. Comparison between the asynchronous and live interview treatments allow us to study the effect of moving from the current status quo to asynchronous interviewing, including the

[8] We informed candidates of the form of evaluation as it is increasingly mandatory to inform job applicants if AI is used in the hiring process, such as the introduction of the European Parliaments Digital Rights Act.

differences that may arise in time and effort to prepare and beliefs about the recruitment process and the firm.

We conduct this experiment using one of the world's largest automated interview platforms, which is utilized by over 60% of Fortune 100 companies and has hosted more than 70 million virtual interviews at the time. The platform provides a structured inventory of interview questions tailored to the advertised job, with each question designed to assess key job-relevant competencies such as communication and drive. We use the platform's standard set of five interview questions for programmers, web designers, and content creators.[9] Applicants are given one minute to practice each question before recording a response of up to two minutes. In the live online interview treatment, we use the same set of questions, read aloud by a human interviewer to ensure consistency across conditions. The live interviewer was instructed not to deviate from these questions.

The interview platform allows assessments by an employer's human assessors and/or the platform's proprietary AI algorithm, which evaluates only the transcripts of the interviews (i.e. no facial analysis, voice analysis, etc.). Following standard practice, the platform informs all applicants which type of assessor (AI or human) is used in the interview evaluation depending on the treatment (AI or human) to which they are assigned. Screenshots of these two pages may be found in the Appendix A. Separately, to understand the applicants' behaviors and responses to the use of asynchronous interviews, we conduct a randomized vignette experiment (n=606) using Prolific participants from specific professions who indicate that they would apply for the jobs that we advertised. We describe this in detail in Section 3.2.

[9] The following competencies are assessed for all jobs: *communication:* [assessed generally for all questions below]; *drive for results and initiatives*: "Please describe a time you set goals for a project that were hard to achieve and the steps you took to reach these goals. Describe the situation, what actions you took, and the result"; *adaptability:* "Please describe a time you had to change your course of action while working on an assignment. Describe the situation, what actions you took, and the result"; *dependability: "*Tell us about a time when you did what was right, even though there were easier options. Please describe the situation, your actions, and the outcome"; *composure: "*Please tell us about a time when you were overwhelmed or stressed. Describe the situation, your actions, and the outcome". The following competency was assessed for Programmers and Web Designers only: *problem solving: "*Tell us about a time you had to solve a problem at school or at work, but you were given confusing information. Please describe the situation, what you did, and the outcome". The following competency was assessed for Content Creators only: *relationship building: "*Please describe a time you built a great relationship with someone. Describe the situation, what actions you took, and the result. How was the relationship helpful?".

2.2 Labor Demand: Applicant Evaluations

Each interview response was independently assessed by both an AI-based evaluation tool and professional human evaluators. In addition, all resumes are evaluated by an AI-powered screening tool. We describe each of these in turn:

AI evaluation of interviews

The AI tool used to evaluate candidates is widely used by employers who use this major online interview platform; it is thus one of the most influential interview assessment algorithms in the world. All candidates who completed a screening interview were evaluated using the interview platform's AI evaluation tool. The platform produced the AI score independent of the researchers, using the same algorithms that they normally use to evaluate millions of interviews. For each interview question, a score is generated based on a predefined set of competencies. Specifically, each of the five questions are scored on two criteria: the candidate's communication skills and one additional competency aligned with the question. The additional evaluated competencies are adaptability, problem-solving, dependability, composure, drive, and relationship-building. Ratings are assigned on a scale from 0 to 5, with 5 being the highest score. To determine a final score, the scores for each competency across all questions are averaged, resulting in a final score ranging from 0 to 5, where 5 represents the highest possible score. This follows the standard practice of the AI scoring tool. The AI tool was not explicitly aware of a candidate's gender or other demographics and has no access to the candidate's resumes. Per the platform's protocol, only transcripts from the interviews are analyzed and no facial recognition or voice analysis tools are used. A summary of the AI scoring process which involved three stages are reported below (see Appendix B for more details):

1. **Transcription**: Video and audio responses were transcribed into text. This means that the AI evaluation focused solely on the content of responses, without considering tone, body language or aspects of the video such as video background. This is per the platform's protocol and is the process used in all of the platforms' AI evaluations.
2. **Natural Language Processing (NLP)**: To interpret candidates' responses, the transcribed text was analyzed using the platform's natural language processing (NLP) model, built on the RoBERTa architecture. This model extracts a range of features based on the content of each response.

3. **Scoring**: The output from the NLP model was fed into a machine learning model trained on over 125,000 interviews. Each training interview had been rated by 2–3 expert evaluators for specific competencies. Through this model, a final score, ranging from 0 to 5, is generated which represents the average competency ratings.

Human evaluation of interviews

We hired a professional international recruitment firm to evaluate all completed interviews. Specifically, four highly experienced professional evaluators were engaged to assess the candidates starting in June 2024. Evaluations were performed by these professionals using the same criteria, which reflects how real human recruiters are asked to score interviews on the platform and ensure direct comparability across evaluation methods. Depending on the interview format, the evaluators either watched recorded video responses or listened to audio responses. Each evaluator rated candidates using the interview platform's standard scoring rubric, which mirrors the exact evaluation criteria used by the AI system (e.g., communication and one additional competency was evaluated for each question).

For every interview, evaluators scored each individual question based on predefined competencies. An overall score was then generated as the average score across all questions.[10] Each candidate was evaluated by at least one randomly assigned evaluator. To assess the consistency of human judgments, a random sample of 100 candidates were independently rated by all four evaluators. The resulting inter-rater correlation was 0.66, indicating a high level of agreement in human assessments across evaluators.

Human evaluators were not informed that candidates were also assessed by an AI tool, nor were they aware that they were in an experiment or that their evaluations would be used for research. Moreover, evaluators did not have the ability to use AI tools themselves during the evaluation process, as the video and audio recordings were accessible only through a secure, non-downloadable platform interface. As is standard, the evaluators were paid a fixed wage for completing the evaluation.

[10] Following standard practice, the platform provided details on how to define each competency and evaluators were trained on these definitions.

AI evaluation of resumes

In addition to human and AI evaluations of interviews, candidates submitted their resumes as part of the initial application process, which were evaluated using a widely adopted commercial resume screening tool commonly used by employers to assess candidate suitability. This tool, used by some of the world's largest companies, compares the content of each resume against the specific job description and generates a score ranging from 0 to 100, where 100 is the highest score. The AI holistically determines whether a candidate has the relevant qualifications, required skills, and experience on their resumes. For example, in the case of programmer, the tool will look for evidence of proficiency in programming languages like Python, a range of relevant qualifications and work experience to determine the applicant's fit for the job, much like a human assessor would when assessing resumes. Resume screening of this kind has become standard practice in modern recruitment.[11] All candidates who have applied and who have submitted a readable resume are scored. No candidates were screened out based on their resume score; this avoids selection and allows us to compare the resume score to the interview scores.

2.3 Career outcomes

To assess the predictiveness of evaluation scores by professional evaluators and hiring algorithms for real-world labor market outcomes, we link candidates to their publicly available LinkedIn profiles approximately one year after their initial job application. Using a combination of full name, location, and occupation, we successfully matched around 80% of applicants to a LinkedIn profile. In Appendix Table A3, we show that there is very little difference in observables, including AI or human evaluation score, between those who have a found profile and those that do not.

We pre-registered two primary outcome measures to capture labour market success.[12] The first, *New Job or Role*, is an indicator equal to one if a candidate's job title and/or employer had changed relative to their status at the time of application. Because all candidates were actively seeking a new position when screened, a job or employer change within the following year is

[11] The resume screener first parses the resume of each candidate. The resume data is then compared against the job description. For instance, if the job description lists python coding as a suggested skill, the algorithm will give additional points if the resume lists python coding as a skill.

[12] We pre-registered a third variable- promotion, but less than 2% of candidates appear to have experienced a promotion during this period. As such we do not report this variable.

interpreted as a proxy for positive labour market success. We use this measure as a signal of candidate quality, under the assumption that stronger candidates are more likely to obtain new or improved job opportunities. The second outcome, *Employed*, is defined for the subset of applicants who were unemployed at the time of application. The variable equals one if the individual became employed within the 12-month follow-up period. This transition from unemployment to employment serves as a meaningful proxy for candidate quality: individuals who are more capable or better suited to the labor market are more likely to secure a job. As such, this outcome allows us to assess how well evaluation scores identify candidates with higher employability.

We also examine an additional, non-pre-registered outcomes: *Seniority*. For *Seniority*, we construct a measure based on each candidate's most recent job title, assigning a score on a scale from 1 (intern) to 11 (CEO), using a standardized classification system.[13] This metric serves as a proxy for candidate quality, under the assumption that more senior roles – such as 'senior programmer' versus 'junior developer' – indicate greater competence or experience (recall the aim of the interviews are to assess candidate quality and not fit). We then construct a binary indicator equal to one for candidates classified as holding a senior position, defined as roles ranked 7 or higher on the classification scale.

In summary, we construct three main outcome measures: (1) whether the candidate obtained a new job or role, (2) whether candidates who were previously unemployed became employed, and (3) the seniority level of their most recent position. These outcomes provide complementary proxies for labor market success, enabling us to evaluate the predictive validity of both AI and human evaluation scores.

[13] A script extracted key words from job titles and assigned an order based on a standard classification of seniority. Seniority classification ranged from from entry-level to executive. For instance, 'intern', 'junior', 'associate', 'analyst', 'engineer', 'senior', 'lead', 'manager', 'director', 'ceo'. As a robustness check, we also created a binary indicator for job titles containing "senior" (or related terms), and the results remain highly robust.

3. Applicant Behavior

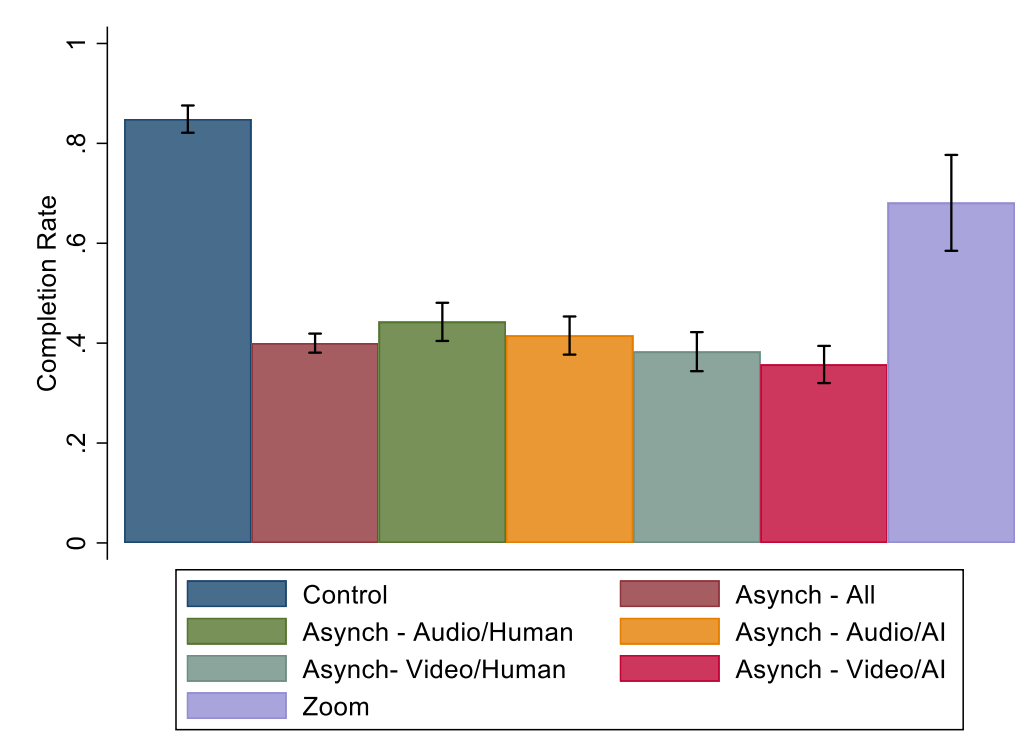
3.1 Results

3.1.1 General Applicant Pool Effects

We first consider the impact of introducing asynchronous interviews of any type to the recruitment process on the size of the applicant pool. In Figure 2 we observe that 85% of applicants in the control indicate their continued interest in the position. In contrast, only 40% of applicants complete the asynchronous evaluations. This means that 53% of the applicant pool who would have continued to be interested and engaged in the recruitment process are deterred by the introduction of asynchronous interviews (45pp, p=0.000).

We next examine the live online interview treatment (Zoom), which serves as an alternative control. 68% of invited candidates participate in the live online interview, which, while lower than in the control (17pp, p=0.000), is still higher than for the asynchronous interviews (0.28pp, p=0.000), despite the asynchronous interviews requiring less time to plan and providing more time and flexibility to prepare. In Column 1 of Table 1 we present the findings in regression form, controlling for the job to which the applicants were applying. The results are robust: the asynchronous interview treatment has 45pp fewer applicants continue than the control (p=0.000), significantly more than the live online interview treatment with only 17pp fewer than control (p=0.001 for zoom vs. control; p=0.000 for zoom vs. asynchronous). We investigate potential explanations for this difference in completion in Section 3.2.

Figure 2: Average Completion Rate by Treatment



Note: Mean rates of completing the next step of the application, as defined by their treatment, by (sub-)treatment with 95% CI bars.

To examine the influence of the different asynchronous interviewing methods, we analyze the results by their respective sub-treatments. The findings, presented in Figure 2 and columns 2-4 of Table 1, reveal some differences in completion rates across the various sub-interview types. We find 6pp lower completion in the video-recording sub-treatments compared to audio-only (p=0.003) and an insignificant 3pp lower completion in the AI-evaluation sub-treatments (p=0.170). However, these sub-treatment differences make up only a small fraction of the drop in completion between the control and the sub-treatment with the highest dropout rate: of the 49 pp (p=0.000) drop in completion between control and Video-AI, only 17% (8.5pp, p=0.002) is between Audio-Human and Video-AI.

Table 1: Impact of Asynchronous Interviews on Application Completion Rates

	(1) Completion	(2) Completion	(3) Completion	(4) Completion
Asynch.	-0.45***			
	(0.02)			
Audio-Human		-0.41***		
		(0.02)		
Audio-AI		-0.43***		
		(0.02)		
Video-Human		-0.46***		
		(0.02)		
Video-AI		-0.49***		
		(0.02)		
Video (vs. audio)			-0.06***	-0.06**
			(0.02)	(0.03)
AI (vs. human)			-0.03	-0.03
			(0.02)	(0.03)
Video#AI				0.00
				(0.04)
Zoom	-0.17***	-0.17***		
	(0.05)	(0.05)		
Job controls	Y	Y	Y	Y
Omitted Category	Control	Control	Audio-Human	Audio-Human
Omitted Category Mean	0.85	0.85	0.44	0.44
N	3296	3296	2535	2535

Note: OLS regressions with robust standard errors in parentheses. The outcome variable is a binary variable that equals 1 if they completed the next step of the application as defined by their treatment and 0 if not. Columns 1-2 include the full sample of applicants, while columns 3-4 include only those applicants in the asynchronous interview treatment. Job control is a dummy variable for the job. Significant at * p<.1, ** p<.05, *** p<.01.

3.1.2 Effects on Applicants by Qualifications

One may argue that deterring a substantial number of applicants is not necessarily a problem. Applicants often submit many applications, including to positions for which they may be weakly qualified or only marginally interested (e.g., Horton and Vasserman, 2021). From an employer's perspective, such deterrence could be beneficial if it disproportionately screens out lower-quality applicants. What would be suboptimal, however, is if deterrence also affects highly qualified candidates, thereby reducing the quality of the applicant pool.

In this subsection we study whether the substantial deterrence of applicants includes those applicants who are most qualified. Measuring the quality or productivity of applicants is not

straightforward and often complex. As such, we consider four separate proxies of applicant qualification to gain a holistic picture of our results by qualification: education, experience, CV score, and future employment outcomes. Table 2 reports the completion rate of the interview process (or continuation in the case of the Control treatment) by treatment, with the samples restricted only to the top applicants according to each proxy of qualifications. This is informative as it tells us the proportion left of each quality subgroup in the applicant pool. For example, Column 1 and 3 shows that 87.39% of applicants with a college degree or higher remain in the control applicant pool, whereas only 40.99% of applicants with a college degree remain in the asynchronous applicant pool. The results are consistent across proxies and show that asynchronous interviews deter top applicants. This finding also holds when comparing asynchronous interviews with the live online interview. This means that employers lose a significant number as well as proportion of their top applicants when choosing asynchronous interviews as compared to traditional live interviews.

3.1.3 Effects by Applicant Gender and Race

Finally, asynchronous interviews may discourage certain groups of candidates which can impact workplace composition. We consider the impact of asynchronous interviews on application completion rates by gender and race and the resulting demographics of the applicant pool.[14] In Figure 3, we present the completion rates of men and women (panel A) and of White/Asian and URM applicants (panel B) by treatment.[15] For gender, we find that while men and women are equally likely to complete in the control treatment, men complete 5.1 percentage points more in the asynchronous treatment than women (p=0.01). As can be seen in Columns 1-2 of Table 3, this result comes from men's application completion rate dropping by 7pp less than women's with the introduction of asynchronous interviews (p=0.055).[16] This results in an applicant pool that is 5pp

[14] Moving forward, we stop using the live online interview treatment – with 94 subjects, we do not have the power for these heterogeneity analyses.

[15] It is useful to note that while Asians are a minority in the US (they make up around 6% of the population), they make up around 10% of the STEM workforce, as such, they are not a minority in this field. For this reason, we merge White and Asian.

[16] We consider whether there are systematic differences in the gender gap in completion or the effect of asynchronous interviews on this gender gap by the gender composition of the position. We find no systematic patterns consistent with the gender composition of the position affecting these outcomes.

more male than in the control treatment (Columns 3-4 of Table 3, p=0.054). We find no similar pattern for race.[17]

Table 2: Completion rate by treatment with restricted sample by applicant qualification proxies

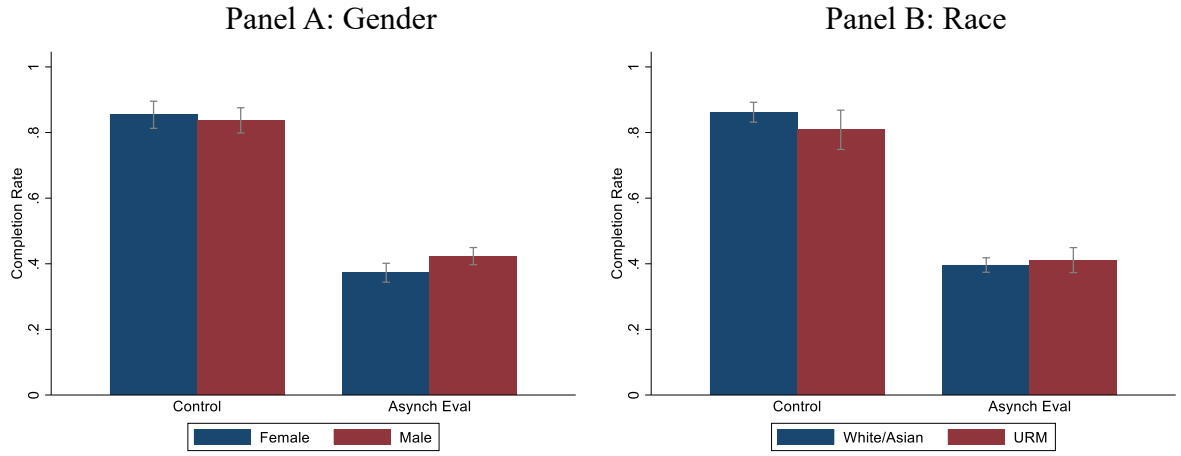
Sample Restrictions:	(1) Control	(2) Live Interview	(3) Asynchronous Interview	(3) - (1)	p-value	(3) - (2)	p-value
College degree or above	87.39%	66.67%	40.99%	-46.40%	0.000	-25.67%	0.000
Postgraduate degree	87.62%	76.92%	41.08%	-46.54%	0.000	-35.84%	0.010
Top 25% experience	87.56%	77.27%	38.92%	-48.64%	0.000	-38.35%	0.000
Top 10% experience	89.04%	77.78%	36.46%	-52.58%	0.000	-41.32%	0.012
Top 25% CV score	88.81%	70.00%	40.43%	-48.38%	0.000	-29.57%	0.008
Top 10% CV score	91.67%	71.43%	44.44%	-47.23%	0.000	-22.23%	0.159
Senior job role	79.41%	75.86%	38.84%	-40.57%	0.000	-37.02%	0.000
Leadership job role	76.82%	85.00%	38.45%	-38.37%	0.000	-46.55%	0.000

Note: Top 25% experience include applicants who have at least 8 years of relevant work experience. Top 10% experience includes applicants who have at least 14 years of relevant work experience. Top 25% and Top 10% CV score restrict samples to applicants whose CV scores are among the top 25% and top 10%, respectively. Senior job role includes applicants whose LinkedIn profile reports either a senior, lead, managerial job title or above, Leadership job role includes managerial job title or above. The interpretation is as follows: take, the top 25% of CV scores. Among applicants whose CV scores place them in the top 25% of the full sample, 88.81% remain in the control applicant pool, implying that 11.19% do not complete the control, while only 66.67% of those with a top-25% CV score remain in the live interview applicant pool.

Result 1: *Asynchronous interviews strongly impact labor supply: they reduce the size of the applicant pool and deter women and the most qualified job-seekers. Audio asynchronous interviews deter fewer job-seekers than video asynchronous interviews.*

[17] We examine whether specific features of these interviews drive the gender gap in completion. Results are reported in Table A1, where we estimate an interaction between the treatments and gender. Overall, we find no gender difference across the different modes of asynchronous interview, suggesting there is no one individual mode of asynchronous interview preferred by males or females.

Figure 3: Average Completion Rate by Treatment and Applicant Gender and Race



Note: Mean rates of completing the next step of the application, as defined by their treatment, by treatment and gender (panel A) and race (panel B) with 95% CI bars.

Table 3: Gender, Treatment, and Completion

	(1)	(2)	(3)	(4)
	Complete	Complete	Frac Male	Frac Male
Asynch	-0.48***	-0.48***	0.04	0.05*
	(0.03)	(0.03)	(0.02)	(0.02)
Male	-0.03	-0.03		
	(0.03)	(0.04)		
AsynchXMale	0.07*	0.07*		
	(0.03)	(0.04)		
Job Controls	X	X	X	X
Other Controls		X		X
Sample	Full	Full	Completers Only	Completers Only
N	3,103	2,958	1,527	1,456

Note: OLS regressions with robust standard errors in parentheses, "other controls" include indicators for race, above median experience, being employed full time, and being a college grad. The dependent variable in columns 1 and 2 is an indicator variable equaling 1 if they completed the next step in the application, as defined by their treatment, and 0 otherwise. The dependent variable in columns 3 and 4 is an indicator variable if the individual is male. Regressions 1 and 2 include the full sample in the control and asynchronous treatments. Regressions 3 and 4 include those who completed the next step in the application, as defined by their treatment, in the control and asynchronous treatments. Significant at * p<.1, ** p<.05, *** p<.01.

3.2 Vignette Experiment

A key finding is that around half of the applicants do not complete an asynchronous interview when invited. We use a conceptual framework and an incentivized randomized vignette experiment to examine the role of plausible economic and behavioral factors.

3.2.1 Conceptual Framework

To understand the mechanisms for dropping out, we use a simple framework, where an applicant completes an interview if the participation constraint is satisfied. That is if the perceived net gain of doing the interview exceeds the outside option:

$$Probability\ of\ success\ \times\ Value\ of\ job\ -\ Costs \geq\ Outside\ Options$$

The use of asynchronous interviews means that many more interviews can be conducted at much lower costs for employers: unlike a live interview (such as a Zoom interview), there is no scheduling inconvenience or the need to have enough interviewers, which may constrain the number of interview invitations that can be sent out. As such, it is possible that many more applicants are invited to the interview, and that applicants may recognize this when evaluating their likelihood of successfully getting the position conditional on interviewing. As such, *probability of success* may depend on factors such as perceived competition in terms of number of applicants who reach the same stage, and perceived selection. For example, if 20% are invited to an interview, then it signals that the applicant is within the top 20% of applicants and has a higher probability of success than an interview process where more candidates are invited.

The type of interview may also indicate something about the *value of the job*. Different interviews may signal certain attributes about the employer, such as work conditions, salary, among other factors. It is possible that applicants think that employers who automate their interview process care more (or less) about their applicants because they are perceived to be spending less or putting less effort into recruitment, which may signal how they may treat their employees. *Costs* may encompass costs such as the effort and time required to schedule and (in our case virtually) attend the interview. *Outside options* also matter if an applicant perceives that they have more and better options.

Finally, we also consider *behavioral factors* that may influence our framework. In particular, we measure the perceived unfairness of the process, which may affect applicants' perceptions of chances of success and may also be seen as behavioral costs, and aversion to new technology (e.g., technophobia).

Taking into consideration these factors, we design a vignette experiment where we elicit beliefs and perceptions based on the above framework. We study whether being invited to an

asynchronous interview instead of a live online interview changes any of the factors, as well as whether these factors are correlated with the stated likelihood of completing an interview.

3.2.2 Method

For our vignette experiment, we employ a sample of "would-be" applicants for our three jobs (n=606). To do so, we recruit US-based Prolific participants filtered by job functions and skills which match our three jobs (content creator, web designer, and programmer).[18] Each participant is shown the corresponding job advertisement which matches their skill profile. Only those who indicated that they would apply if they were looking for a job can participate in the survey. We have 201 participants for content creators, 201 participants for web designers and 204 participants for programmers. We pre-registered the data collection at the AEA registry (AEARCTR- 0013356).

We employ a between-subject design. 304 participants are randomized into receiving a hypothetical asynchronous video interview invitation, and 302 participants are randomized into receiving a live online interview invitation. Both invitations used wording based on the interview invitation email from the original study. Subsequently, participants report (i) their likelihood of completing the interview (0–10), (ii) incentivized beliefs regarding the number of applicants, proportion of applicants invited (0-100), proportion of interview completed (0-100), and the wage offered (elicited using a binarized scoring rule), and (iii) perceptions of the employer and of the recruitment process, specifically the signal about perceived work conditions, the time and effort to complete the interview, and fairness of the recruitment process (each measured on a 0–10 scale).[19] Finally, participants complete a survey where we elicit information on the following additional factors. As a measurement of outside options, we ask participants if they applied for a job today, how many interviews they would expect to receive within a month and what wage they would likely be offered. To help understand if technophobic factors explain our results, we also elicit the widely used Technology Readiness Index (TRI 2.0) (0-5) (Parasuraman and Colby, 2015).

[18] For programmers, we recruit participants who are proficient in python; for web designers, we recruit participants who are proficient in CSS; for content creators, we recruit participants who were employed in journalist, marketing, copywriting, communications, and creative writing roles.

[19] In addition, we re-elicit their likelihood to complete the interview after providing information on how many applicants were invited to the interview (high 75% or low 3% - in random order). These numbers are chosen since 75% of our applicants were randomized into asynchronous interviews and 3% of our applicants were randomized into the live online interview.

3.2.3 Results

First, and consistent with our natural field experiment, we observe that participants randomized into the live online interview treatment report a higher likelihood of completing the interview than those randomized into the Automated treatment (p=0.000).[20] Next, Table 5 summarizes the differences in beliefs about the various economic and behavioral factors based on the two treatments. We observe that the use of asynchronous interviews leads to changes in beliefs for two factors. First, in the live online interview treatment, participants believe that 21% of applicants were invited to the live online interview, whereas in the Asynchronous treatment, participants believe that 35% of applicants were invited to the asynchronous interview (p=0.00). This indicates that applicants believe that the asynchronous interviews are less selective, and the remaining competition is greater.[21] Second, we find that participants think that asynchronous interviews are less fair than live online interviews. As such, we can conclude that the use of asynchronous interviews changes beliefs about an economic factor relating to competitiveness (more applicants invited to the interview) and a behavioral factor concerning costs (a more unfair interview process).[22]

---

[20] The average likelihood of completion is 8.77 for live online interview and 8.01 for asynchronous interviews (p=0.000). If we assume that those who self-report their likelihood to complete at 9 or 10 to be the ones who are most likely to complete the interviews in reality, we find that 69% of participants in the live online interview treatment would complete the interview compared to 47% of those in the Asynchronous treatment (p=0.0305). These numbers are roughly comparable to our main results in the original study, where completion rates were approximately 79% and 47% for live online interview and Asynchronous (pooled), respectively.

[21] We have additional evidence that the economic factor of competitiveness alone could determine whether an applicant completes their interview. Following the main task in the experiment, we provide information on how many applicants were invited to the interview (High: 75%/800 applicants; Low: 3%/30 applicants) and re-ask the participants their likelihood of completing the interview in a random order. In a between-subject analysis, we focus on the first scenario (high or low) that they are randomized into within their treatment (live online interview or asynchronous). First, we find that those who were told that 75% of applicants were invited report lower likelihood to complete the interview compared to those told that 3% were invited (6.49 vs 8.48, p=0.00). The results remain the same whether a participant was in the live online interview or asynchronous treatment (Zoom: 6.45 vs 8.76, p=0.00; Asynchronous: 6.54 vs 8.19, p=0.00). These results indicate that not only does this economic factor significantly impact a participant's likelihood of completing an interview, but this significance of this factor does also not depend on whether the interview is asynchronous or live online.

[22] In our framework, outside options is a factor which we cannot study using the different treatments. Instead, we look at whether participants who believe that they have better outside options are less likely to complete the interviews in the asynchronous treatment. We divide our sample into 1) below and above median number of interviews participants believe that they would receive within a month if they applied for a job (3) and 2) below and above median wage participants believe they would be offered if they applied for a job ($35). We do not find statistically significant differences in likelihood of completing an asynchronous interview in either case, with the self-reported likelihood being 8.37 vs 8.40 (p=0.83) and 8.26 vs 8.51 (p=0.13) for below and above median number of interviews and wage, respectively.

Table 5: Possible Mechanisms for Job-seeker Dropout

	Factors	Zoom	Asynchronous	Difference	p-value
Competition	Incentivized Belief: number of applicants	892	875	-17	0.69
	Incentivized Belief: proportion of applicants invited	21%	35%	14 p.p.	0.00***
	Incentivized Belief: proportion of invited applicants completed	53%	54%	1% p.p.	0.47
	Imputed: number of completions	99	165	66	0.00***
Value of Job	Incentivized Belief: wage	$36.72	$36.86	$0.14	0.91
	Perceived work environment (non-salary factors) [0-10]	6.99	6.83	-0.16	0.27
Costs	Perceived time and effort required to complete the interview [0-10]	6.61	6.39	-0.22	0.19
	Perceived fairness of interview method [0-10]	7.84	6.64	-1.2	0.00***

Note: This table reports the average value of each variable described in 3.2.2 by treatment (Zoom and Asynchronous). The number of completions is imputed by multiplying each participant's belief about number of applicants, proportion of applicants invited and proportion of invited applicants completed together. We note the difference in the values between treatments and the p-value of the t-tests of these differences in the last two columns.

Next, we regress participants' likelihood of completing the interview against the factors as well as their outside options and their Technology Readiness Index score. The results are reported in column 1 in Table 6 for the full sample. We find that both factors that we identify above (belief about proportion of applicants invited to interview and perceived fairness) significantly predicts a participant's likelihood to complete the interview.[23] The impact of being invited to a live online interview instead of an asynchronous interview also decreases once these other factors are controlled for and is no longer significant. We observe the similar results when we restrict the sample to male and female, respectively (see columns 2 and 3 of Table 6).

[23] We note that beliefs about work conditions (excluding salary) also significantly predict completion rates, but there are no statistical differences between treatments for this factor. Technology readiness is positively correlated but not statistically significant (p=0.124), suggesting completion is unlikely to be driven by technophobia.

Table 6: Predictors of Interview Completion (OLS)

	(1) Full Sample	(2) Male only	(3) Female only
	Likelihood to complete	*Likelihood to complete*	*Likelihood to complete*
Number of applicants	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Proportion Invited	−0.011***	-0.013***	-0.011**
	(0.003)	(0.005)	(0.005)
Proportion Completed	0.004	0.004	0.005
	(0.002)	(0.003)	(0.004)
Wage	−0.002	-0.002	0.000
	(0.005)	(0.007)	(0.007)
Work environment (non-salary factors)	0.257***	0.321***	0.195***
	(0.044)	(0.064)	(0.064)
Effort	0.005	-0.042	0.048
	(0.033)	(0.048)	(0.047)
Fairness	0.380***	0.326***	0.448***
	(0.035)	(0.051)	(0.050)
Outside Option: no of interviews	−0.001	0.011	-0.008
	(0.015)	(0.023)	(0.021)
Outside Option: wage	0.001	0.006	-0.006
	(0.004)	(0.005)	(0.006)
Technology Readiness Index	0.198	0.282	0.146
	(0.129)	(0.179)	(0.198)
Asynchronous	−0.104	-0.045	-0.171
	(0.146)	(0.206)	(0.214)
Constant	3.589*	3.220***	3.668***
	(0.613)	(0.858)	(0.907)
Observations	596	333	252
R-squared	0.374	0.324	0.465

Note: This table reports an OLS regression whether the dependent variable is the likelihood of completing the interview on a scale 0-10 where 10 is most likely to complete. Column 1 reports the OLS for the full sample, column 2 restricting the sample to males and column 3 restricting the sample to females. Significant at *** p<0.01, ** p<0.05, * p<0.1.

Result 2: *Participants in asynchronous interviews believe that there is more competition and that the interview process is less fair. These two factors are significant predictors of a self-reported lower likelihood of interview completion.*

4. Interview Evaluations and Labor Market Outcomes

This section consists of two main parts: the interview evaluation and the assessment of labor market outcomes that were conducted between March 2024 and June 2025. This section was pre-registered at the AEA registry (AEARCTR- 0013356), see Appendix 3.

4.1 Summary Statistics

Our main sample comprises the job applicants who completed the asynchronous interview assessment (N=1019).[24] Table A2 shows that 59% of the applicants are men, 32% are classified as an underrepresented minorities (URM, which comprises Hispanic, Black or Native American/Pacific Islander), 47% are White and 18% Asian.[25] On average, participants have 6.3 years of work experience. Around 30% are employed full-time, while 38% are unemployed. In terms of education, nearly half (49%) have a 4-year college degree.

4.1.1 Who is Selected by the Different forms of Evaluation?

Because evaluation scores determine which applicants advance, they shape the kinds of individuals who remain in the candidate pool. We therefore analyze how AI and human evaluators score candidates across key characteristics, with particular attention to gender and ethnicity. Table 8 tests for differences between the standardized AI and human evaluation scores using t-tests in Panel A and between gender and ethnicity in Panel B, with corresponding p-values reported in square brackets.[26] We standardize the scores by evaluation method and job. This allows the analysis to focus on relative rankings of candidates, which is the relevant margin for selection decisions. We find that, on average, women receive higher scores from the AI than human evaluators, yielding a statistically significant difference of 0.193 SDs ($p = 0.0001$). By contrast, human evaluators score male applicants higher than AI evaluators ($p < 0.001$), whereas URM applicants receive 0.208 standard deviations lower scores from human evaluators than from AI

[24] There are 25 candidates who completed the interview, but a score was not generated. This is because their answers were too short for AI to generate a score. 27 applicants completed their assessment and indicated that they were neither male nor female. We exclude these individuals from the analysis.

[25] In practice, fewer than 1.5% of the sample identify as Native American/Pacific Islander, so the URM category predominantly consists of Black and Hispanic candidates. We find no difference across a number of characteristics between Asian and White applicants. Our question eliciting ethnicity allowed for applicants to select multiple ethnicities. Over 80% of applicants selected a single ethnicity. For those who selected more than one, we classify individuals as White only if they selected White and no other group, and as URM (underrepresented minority) if they identified as Black, Hispanic, or a multiracial combination including either (e.g., White and Black).

[26] The correlation coefficient between the standardized human and AI scores is 0.43, suggesting a moderate relationship.

evaluators (p = 0.001). These patterns indicate systematic differences in how AI and human evaluators assess candidates across key demographic groups.

Table 8: Comparing Between AI and Human Characteristics

	Std AI Score	Std. Human Score	Diff [p-values]
Panel A: Comparing Between AI and Human			
Male	-0.096	0.053	-0.150 [0.000]
Female	0.142	-0.051	0.193 [0.001]
URM	0.119	-0.090	0.208 [0.001]
Non URM	-0.060	0.057	-0.117 [0.003]
Panel B: Comparing between Male and Female And URM and Non URM			
	Female	Male	Diff [p-values]
Std. AI Score	0.145	-0.105	0.250 [0.000]
Std. Human Score	-0.070	0.043	-0.113 [0.078]
	Non URM	URM	Diff [p-values]
Std. AI Score	-0.068	0.122	-0.191 [0.005]
Std. Human Score	0.043	-0.097	0.140 [0.037]

Note: This table reports standardized AI scores and human scores by characteristic. The last column reports the difference between the two scores with p-values in square brackets. Panel A reports differences between the AI score and the Human score, while Panel B compares between charasteristics. For Panel A, since the same interview is assessed by both AI and a human, comparisons are made within individual so p-values are calculated using a pairwise t-test. For Panel B, comparisons are made across individuals using a ttest. URM refers to under-represented minority which is equal to one for those that indicate they are African American, Hispanic or Pacific Islander. Non URM refers to those non-African American, Hispanic or Pacific Islander, which comprises mostly of White and Asian candidates.

To test the robustness of these findings, Table 9 presents a regression in which each individual candidate has two observations, one for the human evaluator score and another for the AI evaluator, equivalent to an individual fixed effects model. We include interactions between a human evaluator dummy (equal to one for human evaluator and zero for an AI evaluator) and candidate characteristics (e.g., male) to identify whether evaluation differences vary systematically by group. The dependent variable is the standardized evaluation score. The results confirm our previous patterns: human evaluators score men significantly higher than the AI, while URM candidates receive significantly lower scores from humans relative to the AI. The magnitude of these differences is substantial – human evaluators rate male candidates 0.34 standard deviations higher than the AI evaluator and URM candidates 0.33 standard deviations lower than the AI does.

Table 9: Estimating Whether Human Evaluators Score Candidates Differently than the AI Evaluator (OLS)

	(1) Eval Score	(2) Eval Score
Human Evaluator	-0.193***	0.117***
	(0.056)	(0.039)
Male*Human Evaluator	0.343***	
	(0.070)	
URM*Human Evaluator		-0.325***
		(0.075)
Constant	-0.007	-0.012
	(0.017)	(0.017)
Observations	1,975	1,992
R-squared	0.026	0.021
Number of individuals	1,019	1,027

Note: The dependent variable is the standardized evaluation score. Standard errors are clustered at the individual level. We include individual fixed effects. Column 1 reports the difference in difference model between males and the human evaluator type relative to the AI evaluator type. Column 2 reports the difference in difference model between URM and the evaluator type (Human or AI). URM refers to under-represented minority which is equal to one for those that indicate they are African American, Hispanic or Pacific Islander. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

While our earlier analysis focuses on mean differences, hiring decisions are determined primarily by candidates in the upper tail of the distribution. These high-scoring applicants are the ones most likely to advance and ultimately receive job offers. To assess whether AI and human

evaluators select different types of candidates, we examine the short list based on the highest-scoring groups in each interview method. We also report the shortlist generated by the CV screening tool, which includes applicants who did not complete the interview assessment. As a result, comparisons with the CV screener reflect a different selection process (CV vs interview). We construct parallel shortlists by identifying the top 10% and top 25% of candidates according to AI-based scores and human-based scores. In Table 10, we report the proportion of the top 10% and 25% short lists that are male or URM. The composition of top-ranked candidates differs depending on whether selection is based on human or AI evaluations. Among the top 25% of candidates ranked by human evaluators, 62.9% are male, compared to 48.1% when ranked by the AI with the CV-only evaluation falling in between at 55%. This gender gap widens further in the top 10%, where 62.9% of human-selected candidates are male versus just 42.2% in the AI-selected group while the CV-only evaluation again sits at 55%. A similar pattern emerges for URM candidates: they make up only 29.2% of the top 25% and top 10% in human rankings – close to the 28% under CV-only evaluation – but rise to 36.7% and 42.0%, respectively, under AI evaluation.

Table 10: Who is shortlisted by evaluator type and size of shortlist.

	CV top 25%	Human top 25%	AI top 25%	CV top 10%	Human top 10%	AI top 10%
Male	55.7%	62.9%	48.1%	55.1%	62.9%	42.2%
URM	28.9%	30.3%	37.4%	28.1%	30.0%	42.0%

Note: This Table reports the proportion of each characteristic in the respective shortlist. The Human score at the top 25% and top 10% cutoffs are the same, hence the compositions are the same. The CV shortlist includes the full sample so individuals who have not completed the assessment. URM refers to under-represented minority which is equal to one for those that indicate they are African American, Hispanic or Pacific Islander.

We further examine these differences in shortlist composition in Table 11. As in previous analyses, we create two observations per candidate – one for their AI score and one for their human score. We then estimate a regression model where the dependent variable equals one if the candidate is ranked in the top 25% (columns 1-3) or the top 10% (columns 4-6) of evaluation scores. To test for differential evaluator effects, we interact a dummy variable indicating human evaluation with key candidate characteristics (e.g., gender). We find that human evaluators are 15 percentage points more likely than AI to include male candidates in the top 25% of the evaluation distribution, and 10.8 percentage points more likely in the top 10%. Conversely, human evaluators

are 7.6 percentage points less likely than AI to shortlist candidates from URM groups at both thresholds. This confirms that AI is more likely to short list females and URMs. This brings us to our first study 2 result.

Table 11: Proportion in the top 25% and 10% by characteristic and evaluator type

	(1) Top 25%	(2) Top 25%	(3) Top 10%	(4) Top 10%
Human Evaluator	-0.067**	0.049**	0.095***	0.186***
	(0.027)	(0.019)	(0.025)	(0.018)
Male*Human Evaluator	0.147***		0.108***	
	(0.034)		(0.032)	
URM*Human Evaluator		-0.088**		-0.076**
		(0.036)		(0.033)
Constant	0.235***	0.233***	0.095***	0.092***
	(0.008)	(0.008)	(0.008)	(0.008)
Observations	2,038	2,054	2,038	2,054
R-squared	0.020	0.008	0.105	0.103
Number of Individuals	1,019	1,027	1,019	1,027

Note: The dependent variable in column 1-2 is equal to one if the candidate is in the top 25%, in column 3-4 it is equal to one if the candidate is in the top 10%. This is equivalent to a fixed effects model. Column 1 and 3 reports the difference in difference model between males and the human evaluator type relative to the AI evaluator type. Column 2 and 4 reports the difference in difference model between URM and the evaluator type (human or AI). URM refers to under-represented minority which is equal to one for those that indicate they are African American or Hispanic. Standard errors are clustered at the applicant level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Result 3: *The AI evaluator rates women and underrepresented racial minorities more favorably than human evaluators and produces shortlists with a higher share of these groups.*

4.2 Labour Market Outcomes using LinkedIn Data

It is difficult to determine whether one evaluator type performs better than another. However, since all candidates in our sample were actively seeking work, those who performed better on the labour market within a year can reasonably be considered more successful, potentially signalling that the candidate is higher-quality. As discussed in Section 2, we focus on three labor market outcomes taken from candidates LinkedIn profiles: (i) whether the candidate secured a new job or role, (ii) whether the candidate experienced a transition from being unemployed to

employed, and (iii) job seniority. We successfully match approximately 80% of candidates to a LinkedIn profile.[27]

We estimate an OLS model with the LinkedIn labour market variables as the dependent variable and the standardised evaluation scores as explanatory variables: Table 12 presents the results for the 'New Job or Role' outcome; Table 13 reports estimates for the transition from unemployed to employed outcome; and Table 14 focuses on seniority. For each outcome, we estimate models using AI, human, and CV standardised evaluation scores separately, as well as jointly, to assess their relative predictive power. Comparison with the CV score allows us to compare the evaluations of the interviews to an alternative screening tool. We find that AI evaluation scores are a strong and consistent predictor of whether a candidate obtains a new job or role within 12 months. Across all specifications, the coefficient on the AI score is positive and statistically significant at the 1% level (e.g., column 1: 0.095, $p < 0.01$), even after including human scores and the CV scores (columns 5-7). In contrast, human scores are only weakly predictive of labour market success. Its useful to also highlight that the AI evaluation score coefficient is more than double the size of the other evaluation types. The CV scores, taken from traditional resume information, also show a significant and positive association with job outcomes independent of the interview evaluations (columns 3–6). We find consistent results when examining the employed and seniority outcomes: the AI evaluation score strongly predicts whether a previously unemployed candidate becomes employed (Table 13) and assigns higher scores to those who hold more senior roles (Table 14). Results are highly robust to a standard set of demographic controls including education, years of experience, employment status and job applied (see Appendix Table A5-A7). We also test whether the predictive power of evaluator scores varies by gender or URM status and find no difference in predictiveness by gender or ethnicity (see Appendix Table A8 and A9).

One concern is that not all individuals update their LinkedIn profile. While we were unable to observe when users last logged into LinkedIn, we addressed this concern by collecting data on the most recent instance in which individuals reacted to a post – an indicator of recent platform

[27] LinkedIn data have been widely to study career trajectories, job mobility, and labor market outcomes, as it contains key information on job changes, titles, and employers (Kahn et al., 2021; Xu, 2022). Importantly. prior work shows that LinkedIn profiles capture employment transitions and occupational upgrading, particularly for early-career and professional workers (Altonji et al., 2020).

Table 12: Relationship between Evaluator Score and Obtaining a New Job or Role.

	(1) New Job/Role	(2) New Job/Role	(3) New Job/Role	(4) New Job/Role	(5) New Job/Role	(6) New Job/Role	(7) New Job/Role
AI Score	0.072***				0.068***	0.068***	0.064***
	(0.018)				(0.018)	(0.020)	(0.020)
Human Score		0.039**		0.033*		0.009	0.008
		(0.017)		(0.018)		(0.020)	(0.020)
CV Score			0.043***	0.051***	0.051***		0.051***
			(0.010)	(0.018)	(0.018)		(0.018)
Constant	0.438***	0.439***	0.431***	0.438***	0.436***	0.438***	0.436***
	(0.017)	(0.017)	(0.010)	(0.018)	(0.018)	(0.017)	(0.018)
Testing Across Coefficients using t-test							
Human Score - AI Score						-0.058*	-0.055*
						(0.033)	(0.034)
Observations	799	835	2,437	778	749	792	744
R-squared	0.020	0.006	0.007	0.016	0.030	0.020	0.030

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant obtained a new job or new role. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the standardised evaluation score given by the AI tool. The human score is the score of the standardised human evaluator, and the CV score the standardised score of the CV tool. Column 3 contains the full sample of applicants. The coefficient is 0.054, with a p=0.00, when restricting the sample to those who complete an asynchronous interview. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table 13: Relationship between Evaluator Score and Moving from Unemployed to Employed.

	(1) Emp	(2) Emp	(3) Emp	(4) Emp	(5) Emp	(6) Emp	(7) Emp
AI Score	0.079***				0.069**	0.088***	0.080**
	(0.029)				(0.029)	(0.033)	(0.034)
Human Score		0.012		0.008		-0.020	-0.021
		(0.027)		(0.028)		(0.032)	(0.033)
CV Score			0.054***	0.049	0.059*		0.058*
			(0.017)	(0.031)	(0.031)		(0.032)
Constant	0.469***	0.452***	0.450***	0.445***	0.459***	0.469***	0.458***
	(0.029)	(0.028)	(0.017)	(0.029)	(0.030)	(0.029)	(0.030)
Testing Across Coefficients using t-test							
Human Score - AI Score						-0.108*	-0.100*
						(0.056)	(0.057)
Observations	292	310	900	296	279	290	278
R-squared	0.025	0.001	0.011	0.009	0.034	0.026	0.035

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant moved from being unemployed to employed. The sample is restricted to those who were unemployed at the time of the job application. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the standardised evaluation score given by the AI tool. The human score is the standardised score of the human evaluator, and the CV score contains the standardised score of the CV tool. Column 3 contains the full sample of applicants. The coefficient is 0.049, with a p=0.11, when restricting the sample to those who complete an asynchronous interview. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table 14: Relationship between Evaluator and Score Seniority.

	(1) Seniority	(2) Seniority	(3) Seniority	(4) Seniority	(5) Seniority	(6) Seniority	(7) Seniority
AI Score	0.061***				0.066***	0.077***	0.080***
	(0.021)				(0.022)	(0.023)	(0.024)
Human Score		-0.012		-0.007		-0.040*	-0.039*
		(0.020)		(0.021)		(0.023)	(0.023)
CV Score			-0.003	-0.000	-0.004		-0.004
			(0.012)	(0.021)	(0.021)		(0.021)
Constant	0.423***	0.423***	0.432***	0.420***	0.419***	0.424***	0.422***
	(0.020)	(0.020)	(0.012)	(0.021)	(0.021)	(0.020)	(0.021)
Testing Across Coefficients using t-test							
Human Score - AI Score						-0.118***	-0.120***
						(0.038)	(0.039)
Observations	582	603	1,734	562	546	577	542
R-squared	0.014	0.001	0.000	0.000	0.017	0.019	0.021

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant has a senior job title. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the standardised evaluation score given by the AI tool. The human score is the standardised score of the human evaluator, and the CV score the standardised score of the CV tool. Column 3 contains the full sample of applicants. The coefficient is -0.001, with a p=0.96, when restricting the sample to those who complete an asynchronous interview. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

activity.[1] In Appendix Tables A10-11, we re-estimate Table 12-14 but restrict the sample to those who have reacted to a post in the last 12 months. Our results are highly robust to this restriction with the interpertation of the predictiveness of the evaluator scores not changing.

Result 4: *AI evaluation is more predictive of labour market outcomes than the human evaluators.*

4.3 Mechanisms: Understanding the Differences between AI and Human Evaluators

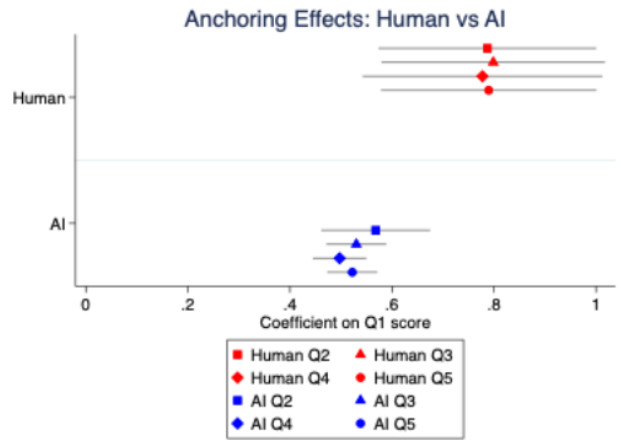
In this subsection, we examine possible mechanisms that may explain why the AI and human evaluators differ in their evaluation performance. We focus on three prominent human cognitive limitations that have been shown to affect behavior (e.g., Tversky & Kahneman, 1974; Kahneman, 2011): (1) anchoring; (2) inconsistency and time-of-day effects; and (3) rating compression.

4.3.1 Anchoring

The first mechanism we test is anchoring – the tendency for initial impressions to disproportionately influence subsequent evaluations (Tversky & Kahneman, 1974; Simmons et al., 2011; Furnham & Boo, 2011). Because human evaluators must score five interview questions sequentially, the first response may serve as a reference point, shaping scores on subsequent items. Under cognitive overload, evaluators may lean even more heavily on this initial impression when faced with complex or lengthy responses. By contrast, AI is not bound by the sequential format: it evaluates each question independently, which should mitigate this form of bias. We find evidence for anchoring in our data (see Figure 4). Human scores on Q2–Q5 are highly correlated with Q1 (all correlations >0.80), and later questions are similarly correlated with one another (>0.83). In contrast, the AI questions show only moderate correlations (0.49–0.56), more consistent with independent evaluation of each question.

[1] A reaction on LinkedIn is a quick, emoji-based response that users can give to posts, articles, or updates. It is a common way for users to interact on the platform. By definition, one must be active on LinkedIn to have reacted. Note, it is still possible that one has been active on LinkedIn without reacting to any posts. Further, we argue that for any potential bias from outdated LinkedIn profiles to threaten our conclusions, it would need to be systematically correlated with the type of evaluation a candidate received—specifically, it would require that candidates evaluated more favorably by the AI (or by humans) are also more or less likely to update their profiles or react to posts. This seems unlikely, as LinkedIn activity is not determined by the evaluation treatment itself and is plausibly independent of how candidates were scored at the screening stage.

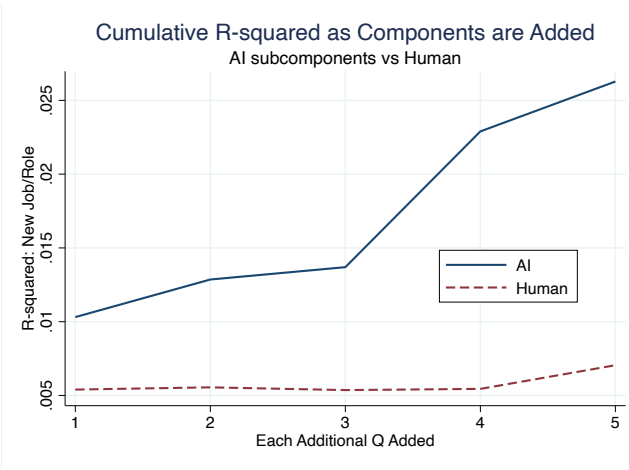
Figure 4: Correlations between rating of first question and subsequent questions.



Note: This figure reports the coefficient from a regression where the dependent variable is the score in Q1 and the right hand side variable is the score from Q2 to Q5, where each coefficient is a separate estimate. The figure includes 95% confidence internal.

If human judgments are influenced by anchoring effects, later evaluations may provide little new information in predicting labor market outcomes, and the predictive value of additional questions will be minimal. To test this, we examine whether the explanatory power (R^2) of the model in predicting later job outcomes evolves as evaluation questions are sequentially added from Q1 to Q5. Figure 5 plots the cumulative R^2 for AI scores (solid blue) and human scores (dashed red), where the horizontal axis indicates the number of questions included in the model and the vertical axis shows the model's predictive power for whether candidates secured a new job/role (our main labor market outcome). If evaluators extract new information from each question, R^2 should rise meaningfully as additional questions are included in the model. For AI, this is what we observe: the cumulative R^2 increases steadily from Q1 to Q5, indicating that each subcomponent captures distinct, predictive information. For humans, however, the cumulative R^2 is nearly flat, rising only marginally across questions. The near-zero incremental value of later questions is consistent with anchoring effects. This pattern suggests that human evaluators may not be treating questions independently but instead forming an impression early and carrying it forward.

Figure 5: Cumulative R^2 as Questions are Added to Predict Obtaining a New Job or Role



Note: This figure reports the R^2 from a regression where the dependent variable is the score in Q1 and we sequentially add the other questions (in order). The horizontal axis shows the R^2 as each question is added.

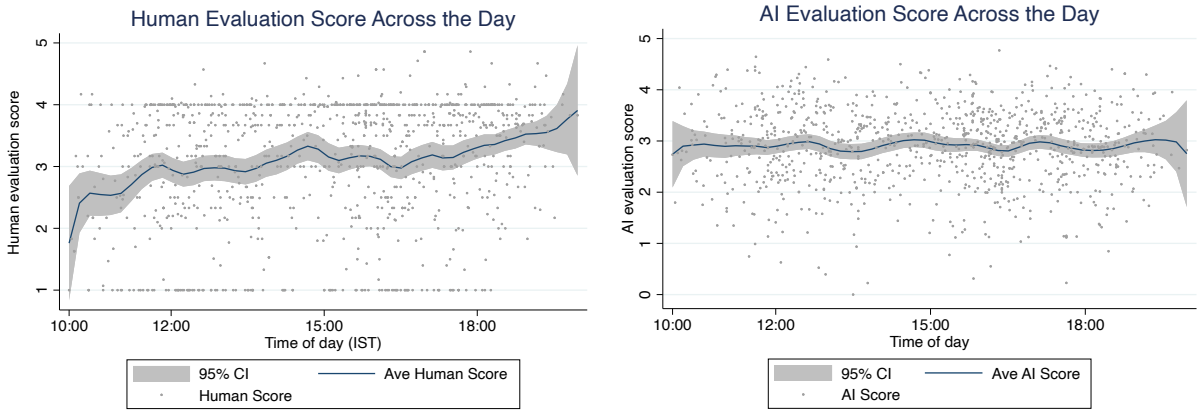
4.3.2 Humans are inconsistent and suffer from time-of-day effects

Unlike algorithms, human evaluators are subject to fluctuations in energy, mood and attention. These factors may not only lower the quality of decisions but also increase susceptibility to systematic errors and biases (Mobius et al., 2022). Consistent with this, research documents time-of-day effects across a range of professional settings: judges grant parole less frequently later in the day, teachers assign systematically different grades, and physicians alter prescribing patterns as fatigue sets in (Danziger et al., 2011; Kouchaki & Smith, 2014, Gabaix, 2019; Hirshleifer et al., 2019). Such temporal inconsistencies highlight potential biases in human evaluation, in contrast to AI that applies criteria consistently and without degradation over time.

To empirically assess whether human evaluation assessment may suffer from time-of-day effects, we exploit the fact that each human evaluator scored approximately 250 interviews across 15 days and they were assigned candidates in a random order. For each evaluator we order their interviews by the exact completion timestamp and construct a variable that indicates the time of day the evaluation took place. Figure 6 plots scatter points of individual evaluation scores against the time of day at which interviews were assessed, including lines of best fit and 95% confidence intervals. For human evaluators (left panel), the figure illustrates a systematic drift across the day:

scores begin relatively low in the late morning, rise around midday, flatten in the afternoon, and increase again in the evening. By contrast, AI evaluation scores (right panel) display an almost flat line indicating consistency across the day. This illustrates that while human assessments are subject to temporal variability, AI evaluations remain steady throughout the day. These results are highly robustness to a number of estimation models including with evaluator fixed effects.

Figure 6: Evaluation score by time of assessment.



Note: This figure plot the scatter plot of the evaluation score by time of human evaluation. The line reports the correlation between time of day and the evaluation score with 95% confidence intervals. The left panel reports the human evaluation score and the right panel the AI evaluation score at the time the score was evaluated by humans.

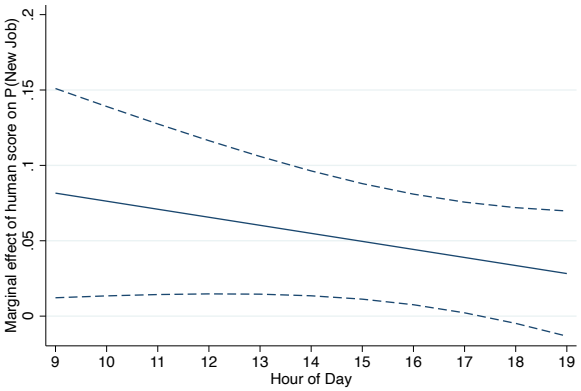
To assess whether this temporal variability also weakens predictive validity, that is, the extent to which an evaluation score predicts whether a candidate later obtains a new job, we estimate a regression with new job/role as the dependent variable and the main variable of interest as the interact between human scores and the time of day. Figure 7 shows the marginal effect of the interaction by time of day, showing that the predictiveness of the human evaluation score falls as the day progresses. This suggests that late-day evaluations may contain more noise and convey less information about true candidate quality.

4.3.3 Ratings Compression

A third possible mechanism is rating compression: the tendency for human evaluators to avoid using the full scoring scale and instead cluster scores toward the middle or upper-middle range (also related to central tendency bias). This pattern is well documented in organizational and personnel psychology. This research finds that evaluators frequently avoid extreme scores due to social concerns, fear of appearing unfair, or to minimize complaints (Levy & Williams 2004; Bol

2011; Prendergast, 1999; Golman, R., and Bhatia, 2012). Compressed ratings have also been found to correlate only weakly with objective performance outcomes (Levy & Williams, 2004).

Figure 7: Effect of Human Evaluation Score on Labor Market Outcomes by Time of Assessment

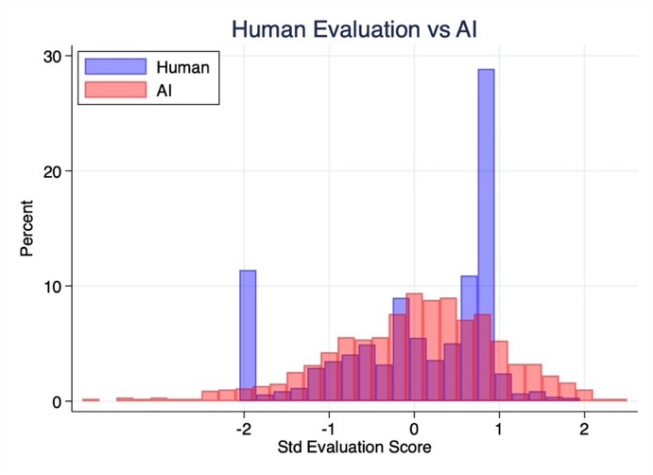


Note: This figure reports the interaction coefficient taken from the estimate of the new job/role and the interaction between time of day and the evaluation score. We include 95% confidence intervals as dotted lines.

To study this mechanism, we examine the standardized distributions of evaluation scores produced by the AI and human assessors.[34] We standardize scores by evaluation type to ensure the scores are comparable. Consistent with existing research, we find that human evaluators compress their ratings (see Figure 8). The human distribution exhibits pronounced spikes, indicating that evaluators frequently bunch candidates into a limited set of preferred categories. Nearly half of all human-assigned scores fall between 3 and 4 on a 5-point scale. Human evaluators' lack of dispersion may reduce evaluators' ability to differentiate candidates. By contrast, AI scores exhibit a smoother and wider distribution that closely resembles a bell-shaped curve, suggesting finer differentiation across candidates. To examine whether rating compression among human evaluators affects their ability to predict labor market outcomes, we exploit variation in score dispersion across evaluators. Specifically, we test whether evaluators who use a wider range of the rating scale – reflected in higher within-evaluator score variance – produce evaluations that are more predictive of subsequent employment outcomes. We find that evaluators with greater within-evaluator variance tend to predict labor market outcomes more accurately, although results must be interpreted very cautiously given the limited number of evaluators.

[34] This may also be a proxy for how much information each evaluator extracts from candidate responses and how finely they distinguish across applicants. For instance, if human evaluators are less able to process diverse signals and thus may be less able to distinguish between candidates, their score distributions may exhibit compression.

Figure 8: Distribution of Standardized Human and AI scores



Result 5: *Human evaluators exhibit several cognitive biases that influence their assessments and reduce their predictive validity of labour market outcomes. In contrast, the AI evaluator appears less affected by these biases.*

5. Conclusion and Discussion

Hiring has entered a new era, with technology revolutionizing one of the oldest recruitment tools: interviews. This study advances our understanding of what this shift means for employers, job applicants, and the functioning of the labour market. First, we provide some of the earliest field-experimental evidence on how asynchronous interviews influence job applicant behaviours. By enabling a clean comparison between asynchronous interviews, traditional online interviews, and a no-interview control group, our design isolates how asynchronous interviewing platforms affect participation at the interview stage. The results show that asynchronous interviews substantially deter applicants, revealing that recruitment technologies operate not only through screening, but also through selection into the applicant pool. A complementary vignette experiment shows that this deterrence is not driven by unfamiliarity with the technology, but rather by concerns about fairness and competitiveness. This highlights that recruitment platforms are not neutral intermediaries: how they structure and frame the hiring process has first-order effects on applicant behaviour.

Second, we provide one of the earliest direct comparisons between professional human evaluations and AI-generated evaluations of the exact same interviews in a real hiring environment. Observing both systems side by side allows us to identify systematic differences in evaluation patterns, including demographic consequences that affect who is shortlisted. This comparison underscores that the introduction of AI into hiring changes not only the scale and cost of evaluation, but also the nature of judgment itself.

Third, we link both human and algorithmic interview assessments to subsequent labour-market outcomes. This allows us to evaluate the predictive validity of different screening technologies. AI-evaluated interviews outperform professional human evaluations in predicting later job success, suggesting that algorithmic assessments extract economically relevant signals that are difficult for even experienced evaluators to identify consistently.

Fourth, we examine mechanisms that drive divergence between human and AI evaluations. Professional human evaluators exhibit systematic patterns – including time-of-day effects, anchoring, and ratings compression – that are absent in algorithmic assessments. These findings point to inherent limits of human judgment under cognitive and organizational constraints and help open the "black box" of how emerging screening technologies differ from traditional evaluation.

Taken together, these contributions provide new evidence on how technological innovations reshape access to employment and inform ongoing debates about efficiency and fairness in recruitment. Importantly, they show that hiring technologies cannot be evaluated solely by their ability to rank candidates conditional on participation. By altering who enters the interview stage, these tools directly affect the composition of the applicant pool, with implications for matching efficiency that precede any formal screening decision.

The transition from human to AI recruitment raises fundamental questions about job applicant behavior, assessment, and mobility that can be more easily examined during periods when both systems operate side by side. Our results have significant implications for employers and policymakers. The high deterrence associated with asynchronous interviews – particularly among high-performing applicants and women – suggests that employers should deploy such tools with caution. However, our vignette evidence indicates that deterrence is driven by perceived fairness and competitiveness rather than the technology itself. This implies that employers may mitigate crowding-out by being more selective prior to the interview stage and by clearly signalling that interviews are competitive, structured, and designed to enhance fairness.

AI-based evaluation predicts subsequent labour market success more accurately than professional human evaluators, selects more women and underrepresented minorities. Furthermore, it does so using interview transcripts alone, i.e. without access to resumes, background information, voice, facial features, or direct demographic indicators. This suggests that algorithms can identify soft skills that are valuable in the labour market but difficult to assess reliably through human judgment. With further AI development, our findings likely represent a lower bound on the potential of such technologies.

At the same time, the limitations of human evaluation highlighted in this study suggest that professional judgment should not be treated as an unbiased benchmark. Even experienced evaluators exhibit systematic patterns – such as time-of-day effects, anchoring, and ratings compression – that can distort their evaluations. These findings imply that differences between human and algorithmic assessments partly reflect constraints inherent to human judgment. Recognizing these limits is important when interpreting human evaluations in hiring decisions and when assessing claims about the relative fairness or accuracy of alternative screening tools.

Finally, our findings point to important dynamic and welfare implications. Early participation shapes the data on which AI screening tools are trained, and systematic deterrence at the interview stage may influence how future hiring technologies evolve and whom they ultimately serve. From a market-design perspective, screening technologies that improve ranking conditional on entry may nevertheless reduce welfare if they discourage participation by high-quality applicants or by certain demographic groups. By capturing a rare transition period in which human and algorithmic systems operate side by side, this study shows that the economic consequences of AI in hiring depend critically on *where* in the hiring pipeline these technologies are introduced. Designing recruitment systems that are both efficient and inclusive therefore requires accounting not only for predictive accuracy, but also for behavioural responses and selection effects that shape access to employment.

References

Aka, Ada, Emil Palikot, Ali Ansari, and Nima Yazdani. "Better Together: Quantifying the Benefits of AI-Assisted Recruitment." *arXiv preprint arXiv:2507.08029* (2025).

Amer, Abdelrahman, Ashley C. Craig, and Clémentine Van Effenterre. "Decoding Gender Bias: The Role of Personal Interaction." IZA Discussion Paper No. 17077 (2024).

Autor, David H. "Wiring the labor market." *Journal of Economic Perspectives* 15, no. 1 (2001): 25–40.

Autor, David H., and David Scarborough. "Will Job Testing Harm Minority Workers?" NBER Working Paper 10763 (2004). https://doi.org/10.3386/w10763

Avery, Mallory, Andreas Leibbrandt, and Joseph Vecci. "Does artificial intelligence help or hurt gender diversity? Evidence from two field experiments on recruitment in tech." (2024). CESifo Working Paper No. 10996, Available at SSRN: https://ssrn.com/abstract=4764343

Awad, Edmond, Loukas Balafoutas, Li Chen, Edwin Ip, and Joe Vecci. "Artificial intelligence and debiasing in hiring: Impact on applicant quality and gender diversity." *Available at SSRN* (2023). https://doi.org/10.2139/ssrn.4626059

Bishop, John H. "The economics of employment testing." (1988).

Bol, Jasmijn C. "The determinants and performance effects of managers' performance evaluation biases." *The Accounting Review* 86, no. 5 (2011): 1549–1575.

Brenner, Falko S., Tuulia M. Ortner, and Doris Fay. "Asynchronous video interviewing as a new technology in personnel selection: The applicant's point of view." *Frontiers in psychology* 7 (2016): 863.

Brookings. (2025). Gender, race, and intersectional bias in AI resume screening via language model retrieval. Brookings Institution. https://www.brookings.edu/articles/gender-race-and-intersectional-bias-in-ai-resume-screening-via-language-model-retrieval

Carvajal, Daniel and Franco, Catalina and Isaksson, Siri, Will Artificial Intelligence Get in the Way of Achieving Gender Equality? (2025). Available at SSRN: https://ssrn.com/abstract=4759218

Cowgill, B. "Bias and productivity in humans and algorithms: Theory and evidence from resume screening." *Columbia Business School, Columbia University* (2020).

Cowgill, B. "The impact of algorithms on judicial discretion: Evidence from regression discontinuities." *Unpublished Manuscript, Columbia Business School* (2018).

Cowgill, B, Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. "Biased programmers? or biased data? a field experiment in operationalizing ai ethics" In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 679-681. (2020).

Dargnies, Marie-Pierre, Rustamdjan Hakimov, and Dorothea Kübler. "Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence." *Management Science* (2024).

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences* **108**, no. 17 (2011): 6889–6892

de la Croix, David, Matthias Doepke, and Joel Mokyr. "Clans, guilds, and markets: Apprenticeship institutions and growth in the preindustrial economy." *Quarterly Journal of Economics* 133, no. 1 (2018): 1–70.

Denzer, Manuel, Thorsten Schank, and Richard Upward. "Does the internet increase the job finding rate? Evidence from a period of expansion in internet use." *Information Economics and Policy* 55 (2021): 100900.

Dettling, Lisa J. "Broadband in the labor market: The impact of residential high-speed internet on married women's labor force participation." *ILR Review* 70: 2 (2017): 451-482.

Falls, B., Willis, C., & Liff, J. (2025). The Impact of Explanations on Applicant Reactions to Automated Asynchronous Video Interviews. *International Journal of Selection and Assessment*, 33(2), e70009.

Feld, Jan, Edwin Ip, Andreas Leibbrandt, and Joseph Vecci. "Identifying and overcoming gender barriers in tech: A field experiment on inaccurate statistical discrimination." CESifo Working Paper No. 9970 (2022).

Feld, Jan, Edwin Ip, Andreas Leibbrandt, and Joseph Vecci. "Do Financial Incentives Encourage Women to Apply for a Tech Job? Evidence from a Natural Field Experiment." *AEA Papers and Proceedings* 113 (2023): 432-435.

Fountain, Christine. "Finding a job in the internet age." *Social Forces* 83: 3 (2005): 1235-1262.

Fuller, Joseph B., Manjari Raman, Eva Sage-Gavin, and Kristen Hines. "Hidden Workers: Untapped Talent." *Harvard Business School*, September 2021. https://www.hbs.edu/managing-the-future-of-work/Documents/research/hiddenworkers09032021.pdf.

Furnham, Adrian, and Hua Chu Boo. 2011. "A Literature Review of the Anchoring Effect." *The Journal of Socio-Economics* 40 (1): 35–42.

Gabaix, Xavier. “Behavioral inattention.” In *Handbook of Behavioral Economics: Applications and Foundations 1*, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, 261–343. Amsterdam: Elsevier, 2019

Golman, R., & Bhatia, S. (2012). Performance evaluation inflation and compression. *Accounting, Organizations and Society*, *37*(8), 534-543.

Hiemstra, Annemarie MF, Janneke K. Oostrom, Eva Derous, Alec W. Serlie, and Marise Ph Born. "Applicant perceptions of initial job candidate screening with asynchronous job interviews: Does personality matter?." *Journal of Personnel Psychology* 18, no. 3 (2019): 138.

Hirshleifer, David, Sonya S. Lim, and Siew Hong Teoh. “Driven to distraction: Extraneous events and underreaction to earnings news.” *Journal of Finance* 64, no. 5 (2009): 2289–2325

Mitchell Hoffman, Lisa B Kahn, Danielle Li, Discretion in Hiring, *The Quarterly Journal of Economics*, 133 (2) 2018, https://doi.org/10.1093/qje/qjx042

Horton, John, and Shoshana Vasserman. "Job-seekers send too many applications: Experimental evidence and a partial solution." *Manuscript, MIT* (2021).

Ip, E. Fair AI in hiring: Experimental evidence on how biased hiring algorithms and different debiasing methods affect the quality and diversity of applicants. *Behavioral Science & Policy*, 11(1), 44-54 (2025).

Jabarian, Brian, and Luca Henkel. “Voice AI in Firms: A Natural Field Experiment on Automated Job Interviews.” *SSRN Working Paper* (2025). https://doi.org/10.2139/ssrn.5395709

Jaser, Zahira, Dimitra Petrakaki, Rachel Starr, and Ernesto Oyarbide-Magaña. “Where Automated Job Interviews Fall Short.” *Harvard Business Review* (2022).

Kahneman, Daniel. Thinking, Fast and Slow. New York :Farrar, Straus and Giroux, (2011).

Kausel, Edgar E., Samuel A. Culbertson, and David A. Madrid. “Overconfidence in personnel selection: When and why unstructured interview information can hurt hiring decisions.” *Organizational Behavior and Human Decision Processes* 137 (2016): 27–44.

Kouchaki, Maryam, and Ian H. Smith. “The Morning Morality Effect: The Influence of Time of Day on Unethical Behavior.” *Psychological Science* **25**, no. 1 (2014): 95–102.

Kuhn, Peter, and Hani Mansour. "Is internet job search still ineffective?." *The Economic Journal* 124, no. 581 (2014): 1213-1233.

Li, Danielle, Lindsey Raymond, and Peter Bergman. “Hiring as exploration.” *The Review of Economic Studies*, 2025;https://doi.org/10.1093/restud/rdaf040

Levy, Paul E., and Jane R. Williams. “The social context of performance appraisal: A review and framework for the future.” *Journal of Management* 30, no. 6 (2004): 881–905.

McDaniel, Michael A., Debra L. Whetzel, Frank L. Schmidt, and Steven D. Maurer. “The validity of employment interviews: A comprehensive review and meta-analysis.” *Journal of Applied Psychology* 79, no. 4 (1994): 599–616.

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya Rosenblat. “Managing Self-Confidence: Theory and Experimental Evidence.” *Management Science* 68 (2022): 7793–7817.

Navarra, Katie. The real costs of recruitment, April 11, 2022. https://www.shrm.org/topics-tools/news/talent-acquisition/real-costs-recruitment.

Nawrat, Andrew. “Inside HireVue’s Acquisition of Modern Hire.” *Verdict*, 2023

Neckerman, Kathryn M., and Joleen Kirschenman. "Hiring strategies, racial bias, and inner-city workers." *Social problems* 38, no. 4 (1991): 433-447.

Nugent, Sarah E., and Sarah Scott-Parker. “Recruitment AI Has a Disability Problem: Anticipating and Mitigating Unfair Automated Hiring Decisions.” *Intelligent Systems, Control and Automation: Science and Engineering* 102 (2022): 85–96.

Parasurama, Prasanna, and João Sedoc**.** “Degendering Resumes for Fair Algorithmic Resume Screening.” *arXiv preprint* (2021). https://arxiv.org/abs/2112.08910

Parasuraman, Ananthanarayanan, and Charles L. Colby. "An updated and streamlined technology readiness index: TRI 2.0." *Journal of service research* 18, no. 1 (2015): 59-74.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of economic literature*, *37*(1), 7-63

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). *Mitigating bias in algorithmic hiring: Practices and limitations. Communications of the ACM*, 63(10), 56–63.

Simmons, Joseph P., Leif D. Nelson, Jeff Galak, and Shane Frederick. 2011. “Intuitive Biases in Choice versus Estimation: Implications for the Wisdom of Crowds.” *Journal of Consumer Research* 37 (1): 1–15.

Suvankulov, Farrukh, Marco Chi Keung Lau, and Frankie Ho Chi Chau. "Job search on the internet and its outcome." *Internet Research* 22, no. 3 (2012): 298-317.

Tversky, Amos, and Daniel Kahneman. 1974. “Judgment under Uncertainty: Heuristics and Biases.” *Science* 185 (4157): 1124–1131

Wang, Z., Wu, Z., Guan, X., Thaler, M., Koshiyama, A., Lu, S., Beepath, S., Ertekin, E. and Perez-Ortiz, M., 2024, November. Jobfair: A framework for benchmarking gender hiring bias in large language models. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 3227-3246).

Wilson, K., & Caliskan, A. (2025). No thoughts just AI: Biased LLM recommendations limit human agency in resume screening. Working paper arXiv:2509.04404

Zuo, George W. "Wired and hired: Employment effects of subsidized broadband Internet for low-income Americans." *American Economic Journal: Economic Policy* 13, no. 3 (2021): 447-482.

Zhang, Shuo, and Peter J. Kuhn. “Measuring Bias in Job Recommender Systems: Auditing the Algorithms.” *NBER Working Paper* 32889 (2024), National Bureau of Economic Research.

APPENDIX

Appendix A: Additional Figures and Tables

Figure A1: Job Ad for Content Creator

Content creator for leading international organization in the research sector

Job Information

- Opportunity for a Content Creator
- Compensation commensurate with experience
- Telecommuting: work from anywhere you want within the US
- Contract work with flexible work hours
- Start date can be discussed to suit your needs

Job Description

We are seeking a creative and talented Content Creator to communicate our organisation's research and achievements to generate impact. As a Content Creator, you will play a pivotal role in shaping our digital presence and engaging our audience through compelling and innovative content.

This role offers a fantastic opportunity for someone with a passion for digital marketing, and creative expression.

Responsibilities

- Develop and execute engaging content for a website related to a research project
- Collaborate with the research team to generate content.
- Create visually appealing graphics and images that help explain the research.
- Write, edit, and proofread content for accuracy, consistency, and adherence to our voice.
- Monitor and analyze content performance metrics, making data-driven adjustments to optimize engagement and reach.
- Manage content calendars, ensuring timely and consistent delivery of content.
- Maintain a strong understanding of SEO principles to enhance content visibility and search rankings.

You will know you are successful in this role if you have

- Strong writing skills with the ability to adapt content to different platforms and target audiences.
- Creativity and the ability to generate unique and engaging content ideas.
- Excellent communication and collaboration skills to ensure the research information is adequately understood.
- Ability to work independently and with others.

Figure A2: Job Ad for Web Designer

Web Designer for leading international organization in the research sector

Job Information

- Opportunity for a creative Web Designer
- Compensation commensurate with experience
- Telecommuting: work from anywhere you want within the US
- Contract work with flexible work hours
- Start date can be discussed to suit your needs

Job Description

We are looking for a Web Designer to create a minimalist website that effectively communicates information about our aims, research and achievements.

In this role, you will have the opportunity to express your creativity, talent and drive and design a website that stands out.

Responsibilities

- Create wireframes, prototypes, and mockups to visualize website layouts and user interfaces.
- Implement best practices in user experience and user interface design to enhance website functionality and usability.
- Work closely with developers and content creators to ensure integration of design elements and front-end functionality.
- Design and develop responsive and visually appealing websites that provide a great user experience across various devices.
- Maximize webpage visibility
- Provide feedback and thoughts on the team.

You will know you are successful in this role if you

- Enjoy website design
- Are able to design a beautiful website front end
- Have knowledge of HTML & CSS. Knowledge of JavaScript would also be useful
- Ability to work independently and with others

Figure A3: Job Ad for Programmer

Programmer for leading international organization in the research sector

Job Information

- Opportunity for a programmer
- Compensation commensurate with experience
- Telecommuting: work from anywhere you want within the US
- Flexible work hours
- Start date can be discussed to suit your needs

Job Description

We are looking for a programmer to work on cutting-edge research projects that aim to help understand societal challenges and make a difference in society. The programmer will also help develop a website for the project in collaboration with a web designer and content creator.

Responsibilities:

- Working with researchers from around the world.
- Writing and developing software and solutions.
- Creating and modifying code to meet requirements.
- Providing feedback and thoughts on the projects.
- Collaboration with the design and content teams to bring designs, content and user interfaces to life through coding.
- Debug and troubleshoot issues, performing rigorous testing to ensure the quality and functionality of developed applications.

You will know you are successful in this role if you have

- Proficiency in a programming language such as Python.
- Experience with HTML, CSS and JavaScript is valuable.
- Understanding data structures, algorithm design, problem solving.
- Collaboration, communication and teamwork skills.
- Ability to work independently and with others.

Figure A4: Platform's default page for human evaluation

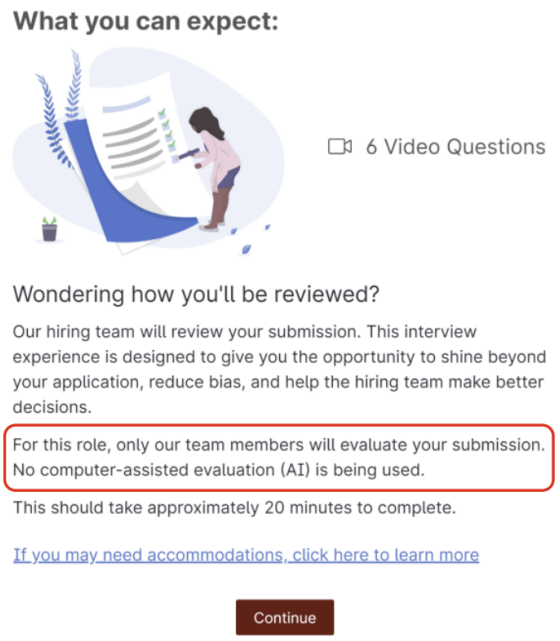


Figure A5: Platform's default page for AI evaluation

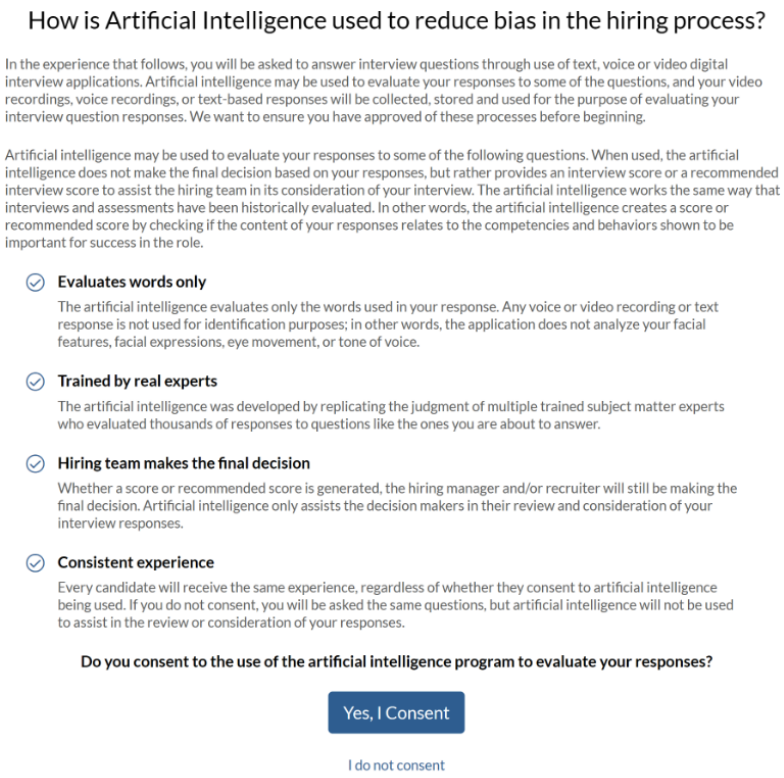


Table A1: Completion by Gender and Asynchronous Interview Type

	Complete	Complete	Complete	Complete	Complete
Male	0.06	0.05*	0.06**	0.07**	0.06
	(0.04)	(0.03)	(0.03)	(0.04)	(0.04)
Audio-AI	-0.03				
	(0.04)				
Video-Human	-0.05				
	(0.04)				
Video-AI	-0.05				
	(0.04)				
Male#Audio-AI	-0.00				
	(0.06)				
Male#Video-Human	-0.02				
	(0.06)				
Male#Video-AI	-0.06				
	(0.06)				
AI		-0.01	-0.02	-0.01	-0.03
		(0.03)	(0.02)	(0.03)	(0.04)
Video		-0.06***	-0.04	-0.04	-0.05
		(0.02)	(0.03)	(0.03)	(0.04)
AI#Male		-0.02		-0.02	-0.00
		(0.04)		(0.04)	(0.06)
Video#Male			-0.04	-0.04	-0.02
			(0.04)	(0.04)	(0.06)
AI#Video					0.03
					(0.06)
AI#Video#Male					-0.04
					(0.08)
Job controls	X	X	X	X	X
Omitted Category	Female#Audio-Human	Female-Human	Female-Audio	Female#Audio-Human	Female#Audio-Human
Omitted Category Mean	0.40	0.38	0.39	0.40	0.40
N	2466	2466	2466	2466	2466

Note: OLS regressions with robust standard errors in parentheses. The outcome variable is a binary variable that equals 1 if they completed the next step of the application as defined by their treatment and 0 if not. Column 1 interacts male with the treatment (e.g AI audio treatment). Column 2 interacts male with whether the evaluation was by AI or human, column 3 interacts male by whether the interview was video, column 4 includes both interactions and column 5 includes a triple interaction between video AI and male. Job control is a dummy variable for the job. Significant at * p<.1, ** p<.05, *** p<.01.

Table A2: Summary Statistics

	(1) N	(2) mean	(3) Sd	(4) min	(5) max
Male	1019	0.588	0.492	0	1
URM	1003	0.324	0.468	0	1
White	1003	0.474	0.500	0	1
Asian	1003	0.180	0.385	0	1
Years of experience	1018	6.322	5.644	0	43
Full time employed	1019	0.302	0.459	0	1
Unemployed	1019	0.377	0.485	0	1
Education					
Less than High school	1019	0.005	0.070	0	1
High school	1019	0.077	0.266	0	1
Some college	1019	0.155	0.362	0	1
2-year college	1019	0.119	0.325	0	1
4-year college	1019	0.490	0.500	0	1
Postgrad	1019	0.153	0.360	0	1

Notes: This Table reports the summary statistics for the sample of applicants who completed an interview evaluation. URM refers to under represented minorities comprising of those who indicate they were Black, Hispanic or Native American/Pacific Islander. Education refers to a candidates high education level completed. Less than high school indicates that the candidate did not complete a high school certificate.

Table A3: Comparison Between the Sample with LinkedIn vs Without LinkedIn Accounts

	(1) LinkedIn Sample	(2) No-LinkedIn	(3) Diff
URM	0.31	0.38	0.07**
Male	0.60	0.53	0.07
Human Score	3.07	3.02	0.05
AI Score	2.88	2.89	0.01
4-year college degree or higher	0.65	0.61	0.04
Full time employed	0.31	0.28	0.03
Unemployed	0.37	0.41	0.04
Content Creator	0.27	0.31	0.04
Programmer	0.33	0.25	0.08**
Web Developer	0.39	0.43	0.04
Learn Uni	0.64	0.61	0.03
Learn Self	0.69	0.72	0.03
F-Stat		1.00 p-value=0.445	

Note: This Table reports the difference in characteristics between the sample with LinkedIn data and the sample without LinkedIn data. The first column reports the characteristics for the LinkedIn sample of applicants and column 2 reports the characteristics for the non LinkedIn subsample. The sample is restricted to those who complete the assessment. *** p<0.01, ** p<0.05

Table A4: Examining Gender and Ethnicity as Predictors of Evaluation Score

	(1) AI Eval	(2) AI Eval	(3) Hum Eval	(4) Hum Eval	(5) AI Eval	(6) AI Eval	(7) Hum Eval	(8) Hum Eval	(9) AI Eval	(10) AI Eval	(11) Hum Eval	(12) Hum Eval
Male	-0.190***	-0.095*	0.114*	0.085								
	(0.049)	(0.053)	(0.065)	(0.074)								
White					-0.094*	-0.102**	0.068	0.069				
					(0.048)	(0.047)	(0.063)	(0.063)				
URM									0.135**	0.118**	-0.143**	-0.142**
									(0.053)	(0.052)	(0.069)	(0.070)
Constant	2.997***	3.112***	3.001***	2.890***	2.925***	3.118***	3.038***	2.888***	2.845***	3.032***	3.117***	2.975***
	(0.037)	(0.074)	(0.051)	(0.095)	(0.034)	(0.076)	(0.045)	(0.098)	(0.029)	(0.078)	(0.037)	(0.097)
Controls	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Observations	965	964	1,010	1,009	973	972	1,019	1,018	990	988	1,036	1,034
R-squared	0.015	0.090	0.003	0.020	0.004	0.090	0.001	0.016	0.007	0.093	0.004	0.021

Note: This Table reports the relationship between gender and ethnicity and the human and AI evaluation score. AI Eval refers to the evaluation score of the AI and Hum Eval refers to the evaluation score of humans. We include a standard set of demographic controls in each even column. URM refers to under-represented minority which is equal to one for those that indicate they are African American or Hispanic. Standard errors are clustered at the applicant level. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table A5: Relationship between Evaluator Score and Obtaining a New Job or Role Controlling for Individual Charasteristics

	(1) New Job/Role	(2) New Job/Role	(3) New Job/Role	(4) New Job/Role	(5) New Job/Role	(6) New Job/Role	(7) New Job/Role
AI Score	0.102***				0.093***	0.099***	0.084***
	(0.025)				(0.025)	(0.028)	(0.027)
Human Score		0.037**		0.032*		0.004	0.008
		(0.017)		(0.018)		(0.020)	(0.020)
CV Score			0.043***	0.043**	0.043**		0.051***
			(0.010)	(0.019)	(0.019)		(0.018)
Applicant Controls	Y	Y	Y	Y	Y	Y	Y
Constant	-0.007	0.206***	0.400***	0.251***	0.048	-0.006	0.169**
	(0.093)	(0.072)	(0.030)	(0.076)	(0.096)	(0.096)	(0.077)
Observations	798	834	2,436	777	748	791	744
R-squared	0.034	0.020	0.010	0.026	0.039	0.033	0.030

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant obtained a new job or new role. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the evaluation score given by the AI tool. The human score is the score of the human evaluator, and the CV score the standardised score of the CV tool. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Applicant controls are: years of experience, whether the applicant is employed or unemployed; a dummy equal to one if the applicant has a graduate degree and job fixed effects. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A6: Relationship between Evaluator Score and Moving from Unemployed to Employed controlling for Individual Charasteristics

	(1) Emp	(2) Emp	(3) Emp	(4) Emp	(5) Emp	(6) Emp	(7) Emp
AI Score	0.114***				0.099**	0.135***	0.118**
	(0.039)				(0.039)	(0.045)	(0.046)
Human Score		0.009		0.008		-0.030	-0.027
		(0.027)		(0.028)		(0.033)	(0.034)
CV Score			0.048***	0.037	0.047		0.044
			(0.018)	(0.032)	(0.032)		(0.032)
Applicant Controls	Y	Y	Y	Y	Y	Y	Y
Constant	0.007	0.312***	0.387***	0.332***	0.071	0.028	0.079
	(0.138)	(0.105)	(0.044)	(0.109)	(0.141)	(0.140)	(0.143)
Observations	291	309	899	295	278	289	277
R-squared	0.049	0.020	0.015	0.029	0.054	0.052	0.056

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant moved from being unemployed to employed. The sample is restricted to those who were unemployed at the time of the job application. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the evaluation score given by the AI tool. The human score is the score of the human evaluator, and the CV score the standardised score of the CV tool. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Applicant controls are: years of experience, whether the applicant is employed or unemployed; a dummy equal to one if the applicant has a graduate degree and job fixed effects. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A7: Relationship between Evaluator Score and Seniority controlling for Individual Charasteristics

	(1) Seniority	(2) Seniority	(3) Seniority	(4) Seniority	(5) Seniority	(6) Seniority	(7) Seniority
AI Score	0.048*				0.049*	0.057*	0.058*
	(0.028)				(0.028)	(0.031)	(0.032)
Human Score		-0.010		-0.008		-0.025	-0.025
		(0.019)		(0.020)		(0.022)	(0.022)
CV Score			-0.016	-0.012	-0.011		-0.012
			(0.012)	(0.020)	(0.021)		(0.021)
Applicant Controls	Y	Y	Y	Y	Y	Y	Y
Constant	0.502***	0.684***	0.632***	0.713***	0.523***	0.554***	0.586***
	(0.105)	(0.080)	(0.034)	(0.084)	(0.108)	(0.108)	(0.112)
Observations	582	603	1,734	562	546	577	542
R-squared	0.120	0.123	0.108	0.143	0.140	0.125	0.147

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant has a senior job title. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the evaluation score given by the AI tool. The human score is the score of the human evaluator, and the CV score the standardised score of the CV tool. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Applicant controls are: years of experience, whether the applicant is employed or unemployed; a dummy equal to one if the applicant has a graduate degree and job fixed effects. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A8: Relationship between Evaluator Score and New Job/Role by Gender

	(1) New Job/Role	(2) New Job/Role	(3) New Job/Role	(4) New Job/Role
AI Score	0.097**			0.096**
	(0.040)			(0.044)
Male	0.052	0.057	-0.003	0.117
	(0.152)	(0.113)	(0.021)	(0.168)
Male*AI Score	0.002			-0.013
	(0.050)			(0.057)
Human Score		0.043		0.014
		(0.027)		(0.031)
Male*Human Score		-0.009		-0.013
		(0.035)		(0.041)
CV Score			0.043***	0.067**
			(0.015)	(0.030)
Male*CV Score			-0.005	-0.035
			(0.020)	(0.038)
Constant	0.120	0.289***	0.434***	0.087
	(0.125)	(0.086)	(0.016)	(0.138)
Observations	779	814	2,367	724
R-squared	0.023	0.007	0.007	0.030

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant obtained a new job or new role. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the evaluation score given by the AI tool. The human score is the score of the human evaluator, and the CV score the standardised score of the CV tool. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A9: Relationship between Evaluator Score and New Job/Role by URM

	(1) New Job/Role	(2) New Job/Role	(3) New Job/Role	(4) New Job/Role
AI Score	0.115***			0.112***
	(0.029)			(0.033)
URM	0.126	-0.001	-0.021	-0.001
	(0.152)	(0.116)	(0.022)	(0.168)
URM*AI Score	-0.054			-0.076
	(0.050)			(0.057)
Human Score		0.039*		-0.017
		(0.021)		(0.025)
URM*Human Score		-0.005		0.068
		(0.036)		(0.043)
CV Score			0.043***	0.050**
			(0.012)	(0.022)
URM*CV Score			0.003	0.001
			(0.022)	(0.041)
Constant	0.117	0.324***	0.439***	0.171*
	(0.085)	(0.070)	(0.012)	(0.094)
Observations	788	824	2,402	734
R-squared	0.023	0.006	0.008	0.034

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant obtained a new job or new role. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the evaluation score given by the AI tool. The human score is the score of the human evaluator, and the CV score the standardised score of the CV tool. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A10: Restricting New Job/Role to active users

	(1) New Job/Role	(2) New Job/Role	(3) New Job/Role	(4) New Job/Role	(5) New Job/Role	(6) New Job/Role	(7) New Job/Role
AI Score	0.110***				0.108***	0.097***	0.096***
	(0.028)				(0.029)	(0.031)	(0.032)
Human Score		0.047**		0.042*		0.018	0.018
		(0.020)		(0.022)		(0.023)	(0.024)
CV Score			0.027**	0.032	0.034		0.035
			(0.012)	(0.022)	(0.022)		(0.022)
Testing Across Coefficients using t-test							
Human Score - AI Score						-0.079*	-0.079*
						(0.045)	(0.047)
Constant	0.202**	0.378***	0.513***	0.391***	0.205**	0.184**	0.185*
	(0.084)	(0.066)	(0.012)	(0.070)	(0.086)	(0.092)	(0.095)
Observations	574	604	1,740	559	534	569	530
R-squared	0.027	0.009	0.003	0.012	0.031	0.026	0.031

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant obtained a new job or new role. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the evaluation score given by the AI tool. The human score is the score of the human evaluator, and the CV score the standardised score of the CV tool. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A11: Restricting Employment to active users

	(1) Emp	(2) Emp	(3) Emp	(4) Emp	(5) Emp	(6) Emp	(7) Emp
AI Score	0.100***				0.094***	0.089**	0.082**
	(0.033)				(0.034)	(0.036)	(0.037)
Human Score		0.046*		0.041*		0.019	0.021
		(0.024)		(0.025)		(0.027)	(0.028)
CV Score			0.028*	0.025	0.035		0.032
			(0.015)	(0.026)	(0.026)		(0.026)
Testing Across Coefficients using t-test							
Human Score - AI Score						-0.070	-0.061
						(0.054)	(0.055)
Constant	0.214**	0.358***	0.513***	0.367***	0.228**	0.188*	0.195*
	(0.097)	(0.076)	(0.015)	(0.080)	(0.099)	(0.105)	(0.108)
Observations	401	425	1,173	397	377	399	376
R-squared	0.023	0.009	0.003	0.010	0.026	0.024	0.027

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant moved from being unemployed to employed. The sample is restricted to those who were unemployed at the time of the job application. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the evaluation score given by the AI tool. The human score is the score of the human evaluator, and the CV score the standardised score of the CV tool. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A12: Restricting Seniority to active users

	(1) Seniority	(2) Seniority	(3) Seniority	(4) Seniority	(5) Seniority	(6) Seniority	(7) Seniority
AI Score	0.075**				0.084***	0.097***	0.104***
	(0.031)				(0.032)	(0.034)	(0.035)
Human Score		-0.020		-0.014		-0.046*	-0.044*
		(0.023)		(0.024)		(0.025)	(0.027)
CV Score			0.003	-0.003	-0.004		-0.003
			(0.014)	(0.024)	(0.024)		(0.024)
Testing Across Coefficients using t-test							
Human Score - AI Score						-0.144***	-0.148***
						(0.050)	(0.052)
Constant	0.221**	0.505***	0.442***	0.478***	0.190**	0.300***	0.272**
	(0.093)	(0.075)	(0.014)	(0.078)	(0.095)	(0.102)	(0.105)
Observations	459	476	1,346	441	427	454	423
R-squared	0.013	0.002	0.000	0.001	0.016	0.019	0.021

Notes: We use an OLS to estimate the models. The dependent variable is an indicator variable equal to one if the applicant has a senior job title. This data comes from candidates LinkedIn profile approx. 12 months after initial application. The AI score is the evaluation score given by the AI tool. The human score is the score of the human evaluator, and the CV score the standardised score of the CV tool. AI Score – Human Score reports a t-test that compares the differences between the AI score and the Human Score coefficient. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Appendix B: Details on the AI Tool

B.1: The NLP Models

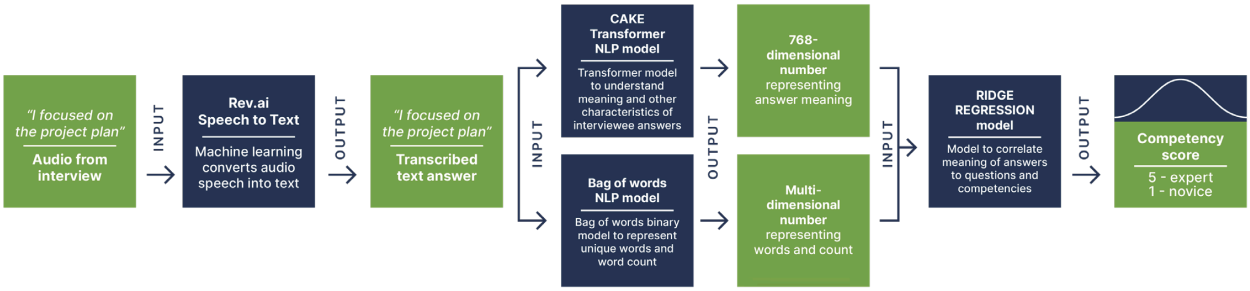
To make sense of the responses the AI tool firm uses a NLP model which is based on a Robustly Optimized Bidirectional Encoder Representations from Transformers model, or 'RoBERTa'. The text analyzed by the NLP model is processed by a 'deep neural network', a technology which comprises a collection of connected nodes or 'neurons' which can attribute a particular weight or significance to various features of the language presented to it. Specifically, the system training happens in two steps: (1) by predicting masked words in a large number of documents, the system learns about the English language, and (2) further refining this model with interview transcripts to understand the nuances of language which might be expressed in job interview scenarios. The output of the neural network is a numerical value – known as a 'vector' – which the model has attributed to the particular answer to an interview question that has passed through the neural network.

In addition to their deep neural network' NLP model, they also used an older and simpler NLP method, called "binary bag of words." This looks at all the words in the answer, with no consideration of the grammar and order of the words. These features add to their ability to explain sentences since they can look at the relative weights of different words in the model. For example, they can identify things like saying the word "team" contributes positively to a candidate's score for the teamwork model.

B.2: Assessing and scoring each candidate

Once the NLP model has understood and assigned numerical values to the candidate's response to an interview question, this numerical value is fed into a 'ridge regression model' (a machine learning system) along with the "binary bag of words" analysis. The ridge regression model has been trained to identify responses of a similar nature and then score those responses against the customer's chosen competencies.

The AI system scores each of the candidate's responses according to a Behavioural Anchored Rating Scale (BARS) for each competency. The BARS guides or content are based on data from thousands of real-life interviews, covering a diverse range of interviewees and job types and the scoring uses five rating levels, from 'novice' to 'expert'. The models they use to assess candidates through interviews have been trained using expert human evaluations of structured interview responses. To create the training assessment scores for each BARS, the AI firm collects thousands of expert human rater evaluations of standardized interviews and uses these ratings to train the models to score candidate interview responses. Their assessment development work and rater studies have drawn upon 125,000 interview evaluations, which include over 500,000 applicant interviews scored. They collected scoring data from interviews for different levels of roles, type of companies, and geographic locations. The expert raters then manually scored each response in the interviews against each competency, with 2-3 separate evaluators scoring each candidate's answer. During the training process, they held regular calibration discussions to ensure consistency in scores from each rater. Based on the above training, their ridge regression model is able to score candidates' responses, by comparing them to the manually scored responses during the training exercise.


"I focused on the project plan"
Audio from interview
INPUT
Rev.ai Speech to Text
Machine learning converts audio speech into text
OUTPUT
"I focused on the project plan"
Transcribed text answer
INPUT
CAKE Transformer NLP model
Transformer model to understand meaning and other characteristics of interviewee answers
Bag of words NLP model
Bag of words binary model to represent unique words and word count
OUTPUT
768-dimensional number representing answer meaning
Multi-dimensional number representing words and count
INPUT
RIDGE REGRESSION model
Model to correlate meaning of answers to questions and competencies
OUTPUT
Competency score
5 - expert
1 - novice