MONASH University

MONASH BUSINESS SCHOOL

Department of Economics
ISSN number 1441-5429

Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech

Discussion Paper no. 2023-09

Mallory Avery, Andreas Leibbrandt and Joseph Vecci

Abstract:

The use of Artificial Intelligence (AI) in recruitment is rapidly increasing and drastically changing how people apply to jobs and how applications are reviewed. In this paper, we use two field experiments to study how AI in recruitment impacts gender diversity in the male-dominated technology sector, both overall and separately for labor supply and demand. We find that the use of AI in recruitment changes the gender distribution of potential hires, in some cases more than doubling the fraction of top applicants that are women. This change is generated by better outcomes for women in both supply and demand. On the supply side, we observe that the use of AI reduces the gender gap in application completion rates. Complementary survey evidence suggests that this is driven by female jobseekers believing that there is less bias in recruitment when assessed by AI instead of human evaluators. On the demand side, we find that providing evaluators with applicants' AI scores closes the gender gap in assessments that otherwise disadvantage female applicants. Finally, we show that the AI tool would have to be substantially biased against women to result in a lower level of gender diversity than found without AI.

Keywords: Artificial Intelligence, Gender, Diversity, Field Experiment

JEL Classification: C93

Mallory Avery: Department of Economics, Monash Business School, Monash University (email: mallory.avery@monash.edu); Andreas Leibbrandt: Department of Economics, Monash Business School, Monash University (email: andreas.leibbrandt@monash.edu); Joseph Vecci: Gothenburg University, Vasagatan, Gothenburg, Sweden (email: joseph.vecci@gu.se).

Does Artificial Intelligence Help or Hurt Gender Diversity?
Evidence from Two Field Experiments on Recruitment in Tech

Mallory Avery (Monash University)*
Andreas Leibbrandt (Monash University)†
Joseph Vecci (University of Gothenburg)‡§

This version: 7 March 2023

Abstract

The use of Artificial Intelligence (AI) in recruitment is rapidly increasing and drastically changing how people apply to jobs and how applications are reviewed. In this paper, we use two field experiments to study how AI in recruitment impacts gender diversity in the male-dominated technology sector, both overall and separately for labor supply and demand. We find that the use of AI in recruitment changes the gender distribution of potential hires, in some cases more than doubling the fraction of top applicants that are women. This change is generated by better outcomes for women in both supply and demand. On the supply side, we observe that the use of AI reduces the gender gap in application completion rates. Complementary survey evidence suggests that this is driven by female jobseekers believing that there is less bias in recruitment when assessed by AI instead of human evaluators. On the demand side, we find that providing evaluators with applicants' AI scores closes the gender gap in assessments that otherwise disadvantage female applicants. Finally, we show that the AI tool would have to be substantially biased against women to result in a lower level of gender diversity than found without AI.

Key Words: Artificial Intelligence, Gender, Diversity, Field Experiment

JEL codes: C93, J23, J71, J78

---

* Avery, Monash University, Wellington Road, Clayton VIC 3800, Australia. Email: mallory.avery@monash.edu
† Leibbrandt. Monash University, Wellington Road, Clayton VIC 3800, Australia. Email: andreas.leibbrandt@monash.edu
‡ Vecci, Gothenburg University, Vasagatan, Gothenburg, Sweden. Email: Joseph.vecci@gu.se
§ Leibbrandt acknowledges support from the Australian Research Council. We thank Alan Benson, Claire Cullen, Jan Feld, Ingrid Haegele and Edwin Ip for their helpful feedback. We thank participants at conferences including ASSA 2023, ESA Bologna 2022, AFE 2022 and seminars at Monash University, the University of Exeter, LMU, University of Zurich and the University of Gothenburg.

1. Introduction

There are substantial and persistent gender disparities in many labor markets.[1] A case in point is the continued underrepresentation of women in STEM, which has not been resolved despite substantial and often costly efforts. Such disparities are problematic for society and are difficult to overcome as underrepresented groups often face bias (Bertrand and Duflo, 2017; Neumark, 2018; Hsieh et al., 2019) and suffer from various forms of discrimination already at the initial recruitment stage (Becker, 1957; Bartos et al, 2016; Sarsons, 2017; Bohren et al. 2019; Bohren et al., 2022; Feld et al, 2022; Kessler et al., 2022; Campus-Mercade and Mengel, 2023; Bohren et al., 2023). Additionally, regardless of whether there is discrimination, the anticipation of discrimination alone can prevent labor market investments and participation by underrepresented groups (Phelps, 1972; Arrow, 1973; Anderson and Haupert, 1999; Fryer et al., 2005; Glover et al., 2017; Delfino, 2021).

There is hope that the expanded use of modern technologies such as artificial intelligence (AI) in the recruitment process will mitigate the impact of human biases and anticipation of discrimination on job application decisions, application assessments, and thus labor market outcomes (Bai et al., 2022; Bao and Huang, 2022; Li et al, 2022; Agan et al., 2023). However, it is unclear how this technological disruption in recruitment will impact the biases, real and anticipated, that underrepresented groups experience in the labor market. Some believe that AI has the power to mitigate all human biases, while others are concerned that AI will increase biases, which may then be misconstrued as impartial due to their technological origin (Cohen, 2019; Houser, 2019; Mirowska & Mesnet, 2021; Shrestha et al., 2019; Tambe et al., 2019; Vassilopoulou et al., 2022). While this debate has grown, not only among AI developers, practitioners, and academics but also in popular media, there is a dearth of evidence about how the use of AI actually affects bias and resulting employment outcomes, especially relative to human recruiters.[2]

[1] See Cain (1986), Altonji and Blank (1999), Rodgers (2009), and Giusta et al. (2020) for a sample of overviews of this literature across time.

[2] Recent articles in the popular press have raised concerns that AI-based recruitment systems may perpetuate racial bias, as they are only as unbiased as the data they are trained on, and that uncareful use of data can easily create racially biased outcomes, leading to the "hard-coding" of bias into AI systems: for example "Amazon scraps secret AI recruiting tool that showed bias against women" (Dastin 2018); "AI and hiring bias: Why you need to teach your robots well" (Kulp, 2021); "AI could be the key to ending discrimination in hiring, but experts warn it can be just as biased as humans" (Holmes, 2019); "How to stop Hiring Bias: Don't Let AI take Over HR" (Galer, 2019); and "Who is making sure the A.I. machines Aren't Racist?" (Metz, 2021).

In this paper, we experimentally study how the use of AI in recruitment affects diversity in the important tech labor market, both at the level of supply and demand and also overall. To do this, we conduct two interconnected field experiments in a real hiring environment. This allows us to measure the response of both job-seekers and employers when AI is added to this recruitment environment and subsequently how the diversity of the applicant pool, particularly the portion of the applicant pool who are considered for a position, changes.

For both labor supply and labor demand it is ex ante unclear how AI will impact behavior. For demand, i.e. in the evaluation and recruitment of diverse candidates, AI can process vast amounts of information which it can then summarize and provide to recruiters, potentially limiting the scope for human bias. However, as AI is trained on human decisions it may be biased as well, exacerbating and entrenching the biases directed towards minority candidates. The impact of AI on the supply, or application behavior, of minority candidates is also unclear and depends on whether AI impacts the perceptions of bias in the recruitment process. The main contribution of our study is to provide a comprehensive experimental design and the first casual evidence on how AI tools affect labor demand for and labor supply of minority job-seekers.

Our study takes place in the context of gender diversity in a labor market for STEM workers in the tech sector. The tech sector is fast-growing and lucrative, and there are reports of bias against women (Fry et al., 2021; Murciano-Goroff, 2022), making it both an important sector in terms of potential outcomes for applicants, employers, and society, as well as one in which there is substantial room to reduce biases. Furthermore, because many women leave the STEM-to-tech pipeline at least in part due to these biases (Beasley and Fischer, 2012), there is room for disruptive technologies to not only redistribute women across already-existing tech firms, but to also retain and draw high-skilled women back into the tech sector.

This study contains two field experiments. In the first experiment, we study whether informing applicants that they are assessed by AI instead of a human recruiter attracts or deters women from completing their application for a tech position. More precisely, we post an actual job for a web designer and invite over 700 interested job seekers to complete an application, randomly varying whether they are informed that their application is evaluated by AI software or a hiring team. We then measure application completion rates and application performance by treatment and gender. We supplement this evidence on the supply side with two complementary surveys with job-seekers to close in on the mechanisms driving application behaviors.

In the second experiment, we study the assessment of these applications by using over 500 people within tech to act as our hiring team. We randomize whether these professional assessors have access to the applicants' evaluation scores provided by the AI software, as well as whether they can infer the applicants' gender. We have these evaluations for both the applicants who applied under the hiring team and the AI software. This allows us to evaluate how supply and demand merge to generate an overall change in the diversity of the pool of applicants most likely to be considered for a job, those at the upper end of the evaluation distribution.

Our experimental results indicate substantial gains in diversity from the use of AI in recruitment, both when isolating supply and demand effects and when integrating these effects together. On the supply side, we see that the use of AI in recruitment increases the proportion of women completing the application by about 30 percentage points relative to men. This causes the closure of the gender gap in application completion rates by 36% relative to recruitment without AI, resulting from both an increase in the completion probability of women and a decrease in completion probability for men. This increase in the diversity of the applicant pool does not come at a cost to the quantity or quality of the completed applications, and complementary evidence from two surveys suggests that the gender treatment effect is driven by applicants' perceptions of the relative bias they experience from AI vs. human evaluators.

On the demand side, we find that evaluators are biased against women in this environment, with women being scored substantially lower than men when names revealing gender are shown but equal to men when names, and thus gender, are hidden. Importantly, the provision of AI scores removes this gap even though evaluators are shown names from which they can infer gender. When merging the labor supply and demand sides of the market and considering the right-tail of the distribution of evaluations, we find that adding AI to recruitment increases the representation of women at the 50th percentile of evaluated applicants by 30% and the 90th percentile of evaluated applicants by 160%.

This paper continues as follows: Section 2 covers essential background and presents a literature review; Section 3 provides an overview of the experimental design. Sections 4 and 5 outline our supply and demand experiments, respectively, starting with the experimental design, continuing with a conceptual framework, and finishing with the results; Section 6 combines the supply and demand results to evaluate market-level outcomes; Section 7 discusses what we can interpret about how our results would change if we were to use AI algorithms with varying levels of bias; and Section 8 concludes.

2. Background Information and Literature Review

Artificial Intelligence in Recruitment

AI's ability to greatly enhance the efficiency of the recruitment process has led to its rapid and widescale adoption (von Krogh, 2018; Malik et al., 2020; Opitz et al., 2022; Vrontis et al., 2022).[3] According to a survey in 2018 by LinkedIn, 67 percent of hiring managers and recruiters were using some form of AI in the recruitment process (LinkedIn, 2018, with similar results found in a survey of HR professionals (Heilmann, 2018). AI usage is expected to grow, with a leading industry group estimating that around 80% of HR professionals expect AI to have a moderate to significant impact on HR and recruitment in 2023 (Wall Street Journal, 2022). More and more organizations use AI because they believe that it can identify suitable candidates through comprehensive and objective evaluation criteria, substantially reducing human time, effort, and subjectivity in the hiring process (Gee, 2017; Lengnick-Hall et al., 2018; Upadhyay & Khandelwal, 2018; Meister, 2019; Malik et al., 2020; Vrontis et al., 2022). Early field experiments show that candidates selected by AI are more likely to progress through the interviews and receive a job offer, more likely to accept a job offer conditional on the offer being made, less likely to indicate having competing offers, and are more productive once hired (Cowgill, 2018a). There is also early evidence that AI can improve hiring in such a way that could benefit society through the promotion of better teachers and the recruitment of police officers who are less likely to use violent force (Chalfin et al., 2016).

However, significant controversy has arisen about whether the growing use of AI will improve or worsen bias against minority groups in the labor market. In the context of recruitment, AI tools rely on training data to make predictions about which new applicants are a good fit for a particular job. This can involve training an algorithm using data on prior successful hires and

[3] AI can be broadly defined as a system of computer-aided solutions for performing tasks using text, data, numbers, images, or sound as inputs for tasks using complex mathematical algorithms to deliver task outputs in the form of decision aids or problems solved (e.g., von Krogh, 2018). The most common AI tools in recruitment are: a) natural language processing to analyse a candidate's personality and values, b) social media analysis, c) analysis of video interviews such as analysis of face expression, sound of voice, and emotions, d) video resumes, e) software that screens resumes for keywords, f) shortlisting candidates based on keywords, g) anonymized job applications, h) skills-based assessment, i) game-based recruitment, j) chat bots to communicate and provide feedback to candidates, k) reference checking systems, and l) embedded candidate management systems.

asking it to predict outcomes for new applicants. By analyzing patterns in the training data, the algorithm learns to identify the most important factors and applies that knowledge to assess new candidates. Even though algorithms can be created that are gender-blind, i.e. they do not use gender as a determining factor in the algorithm, biases can still leak through as different groups tend to also differ on the other dimensions used by the algorithm (see Miller (2019, pg. 30-31) for an example using gender and hobbies listed on a CV, and Sharkey (2018) for further examples). While there is some nascent literature discussing the conditions under which AI tools can be developed to be fair (Cowgill, 2018a; Cowgill et al., 2020; Li et al., 2020; Benson et al, 2022), there is as yet no literature evaluating diversity outcomes when integrating AI tools into recruitment.[4] Our paper provides the first evidence as to whether the use of AI tools in recruitment will impact diversity in labor markets.

It is also crucial to not only evaluate the potential impact of AI without human involvement on applicant evaluations and bias against minority groups, but also to understand how employers integrate AI-generated information into their hiring decisions. Typically, outputs from AI tools are used as inputs for human decision makers, who then ultimately decide, rather than as the ultimate deciding factor. Providing human decision makers with information from AI, rather than just using the output of the AI itself, reduces the efficiency benefits found in earlier papers – ranging from small reductions (Cowgill, 2018a; Stevenson and Doleac, 2022) to an almost complete elimination (Glaeser et al., 2019) of the gains from AI when humans are making AI-assisted decisions.[5] Our paper contributes to this literature by providing the first empirical evidence on whether human recruiters incorporate AI assessments into their decision-making processes when evaluating real job candidates.

In addition, AI in recruitment does not only affect the assessment of employers but can also affect the application choices of job-seekers. In particular, if minority applicants believe that there is greater bias against themselves from AI than from human evaluators, then they may

[4] There is some literature evaluating the impact of the use of algorithmic tools on judges' bail granting decisions, which has a racial component (Cowgill, 2018b; Stevenson and Doleac, 2022); however, the distributive results from these papers are ambiguous, without a clear indication of improved or worsened outcomes for minority defendants.

[5] This may be, in part, due to algorithm aversion, or the tendency of humans to discount information produced by AI when informed that the AI is imperfect in some way (Dietvorst et al., 2015; Burton et al., 2020; Jussupow et al., 2020) or when provided with AI-generated feedback that conflicts with their own already-formed evaluations (Serra-Garcia and Gneezy, 2023). Despite these issues with human-AI decision making, it is very unlikely that humans will be taken out of important decisions such as hiring (Dietvorst et al., 2018; Logg et al., 2019; Chugunova and Sele, 2020; Dargnies et al., 2022).

decrease their application rates. On the other hand, if they believe AI will have less bias than humans, the use of AI may increase their application rates. It is not clear, a priori, what applicants believe or even should believe, given the mixed messages by experts and the media (Wachter-Boettcher, 2017; Galer, 2019; Zielinski 2020). In line with this ambiguity, the public also appears to be ambiguous in their beliefs about the relative fairness of AI and humans, with some studies finding that people believe AI is fairer than humans and others finding that people believe humans to be more fair than AI (Lee and Baykal, 2017; Lee, 2018; Wang, 2018; Acikgoz et al., 2020; Harrison et al., 2020; Marcinkowski et al., 2020; Lee and Rich, 2021; Newman et al., 2020; Zhang and Yencha, 2022). Importantly, we do not know whether such beliefs depend on minority status (Starke et al., 2022) and whether these beliefs impact behavior, such as job application decisions. Our study fills this gap and provides new evidence on the beliefs that job applicants have about the relative fairness of AI and human evaluation but also how those beliefs translate into real application behavior.[6]

Diversity, Bias, and Discrimination in the Labor Market

An extensive set of interventions have been considered and tested to reduce discrimination and bias in labor markets and recruitment, with mixed success. For example, blinding resumes to gender, group interviewing, and balancing the gender composition of evaluation panels have had mixed results for gender diversity (Goldin and Rouse, 2000; Bagues and Esteve-Volart, 2010; Paola and Scoppa, 2015; Bagues et al., 2017; Deschamps, 2018; Cook et al., 2019; Domínguez, 2021; Benson et al, 2022; Mocanu, 2023). While affirmative action has been found to improve the outcomes of intended beneficiaries in the short-run (e.g. Niederle et al., 2013), there is mixed evidence on its impact on employer biases or beneficiaries' long-run outcomes (e.g. Miller, 2017; Dianat et al, 2022). Literature on ban-the-box legislation, which prevents employers from seeing applicants' criminal records until late in the recruitment process, actually find substantially

[6] Lee (2018), Newman et al. (2020), Acikgoz et al. (2020), and Zhang and Yencha (2022) ask survey respondents to report how fair or unfair various hypothetical hiring, promotion, and firing decisions were when made by a human or an AI. Lee (2018), Newman et al. (2020) and Acikgoz et al. (2020) find that respondents view decisions made by AI to be less fair than those same decisions made by humans, but do not consider how minority status impacts these beliefs. Zhang and Yencha (2022) find that women, along with those with less education and less income, more strongly interpret the use of AI in hiring to be unfair than other groups but does not provide evidence on how they perceive the relative fairness of AI vs. humans. Dargnies et al. (2022) provides the most related literature on this topic. In an online experiment, they show that subjects in the role of workers prefer to have the decision of whether they or another worker is hired made by a human rather than an algorithm, but they do choose the algorithm more often when they are informed gender will not factor into the algorithm's decision.

negative effects for minority groups with higher crime rates, even for applicants without a criminal record (Agan and Starr, 2018; Doleac and Hansen, 2020). In contrast to these interventions, we test whether AI assessment tools, though not necessarily used with the intention of promoting diversity, can reduce discrimination in assessments.

Employer bias is not the only cause of disparate outcomes in the labor market – if applicants or workers anticipate or perceive there to be bias, that can generate substantial changes in their behavior potentially harming the pipeline of diverse candidates (Fernandez-Mateo and Fernandez, 2016; Haegele, 2023). This can trigger stereotype threat, or the perception that others have biased beliefs about your ability or characteristics based on your demographic group, negatively impacting performance (Roberson and Kulik, 2007; Spencer et al., 2016), potentially decreasing the quality or quantity applications from a person anticipating bias. Substantial evidence exists that information provided in the recruitment process, either about the job, the application process, or the evaluation process, including information related to bias and diversity, can substantially impact application behavior and change the gender or racial composition of applicants, though not always with the intended effect (Flory et al., 2015; Leibbrandt and List, 2018; Gee, 2019; Banerjee et al., 2021; Delfino, 2021; Flory et al., 2021 & forthcoming). Our study contributes to this literature by providing first evidence on how job-seekers respond to being told that job assessments are made by AI, and whether this changes overall application behavior and behavior by minority status.

A key feature of our field experiment is that we study the impact of an intervention (in our case AI) on the supply and demand of a labor market both independently and jointly, which allows us to understand the effect of the intervention at the market level. For the most part, this is difficult to accomplish, with the existing literature usually only analyzing one of these three things. Specifically, natural experiments using real world data generally capture the impact on an entire system without being able to disentangle supply from demand, whereas field experiments typically study either supply or demand behavior, while holding the rest of the market as given. Only a few related experimental paradigms seek to understand the entire market, such as the literature on gift exchange lab experiments (e.g. Fehr et al., 1997; Gächter and Fehr, 2008) and some studies examining statistical discrimination (Anderson and Haupert, 1999; Fryer et al., 2005; Dianat et al., 2022). In the field experimental space, List (2004) provides an example of an experiment that considers both supply and demand and market outcomes on discrimination. Our study contributes to this literature by studying the impact to both the supply and demand sides of a market, as well as the market level response, of a technological shock in a field setting. This design can also be adapted to study other questions where understanding responses at both the individual supply and demand sides and the entire market are of interest.

3. Overview of Experimental Design

The experimental design consists of two novel field experiments. In experiment 1 (Section 4), we study the impact of introducing AI to the recruitment process on the diversity of labor supply by advertising a real job for a web developer and measuring real application behavior by applicant gender. In the second experiment (Section 5), we study the demand side by measuring how employers' evaluation of candidates, based on the candidates' response in the interview questions and their CV changes when also provided with the AI-produced evaluation scores. We pre-registered the experiment at the AEA registry (AEARCTR-0008296) and received ethics approval.

The Tech Sector as a Test-Case Gender Biased Labor Market

In our study, we use the tech sector as a test-case labor market, studying the impact of using AI in recruitment on gender diversity. The tech sector and STEM workforce is expected to grow by 9.2% by 2029, compared to a 3% increase for the non-STEM workforce, and STEM workers earn on average 65% more than non-STEM workers (Fry et al., 2021). Furthermore, demand for STEM workers far outstrips supply, particularly in industry and government (Xue and Larson, 2015). However, while women make up 47% of the United States labor force and 50% of the United States STEM labor force, they make up only 25% of tech workers (Fry et al., 2021; US Bureau of Labor Statistics, 2021). Women both face and anticipate bias against them in the tech sector and drop out throughout the STEM-to-tech pipeline (Beasley and Fischer, 2012; Makarova et al., 2016; Fouad and Santana, 2017; Sassler et al., 2017; Van Veelen et al., 2019; Bloodhart et al., 2020) at least in part because they cannot break into the male-dominated, higher paying tech sector (Aguirre et al., 2022). As such, there is room for interruptive technologies to not only redistribute women across already-existing tech firms, but to also retain and draw women back into the tech sector.

Our study is embedded in the recruitment of a web developer. Web developers are tech workers who specialize in the development of websites. Web developers alone make up 5% of the tech workforce, with that percentage set to rise in the next 10 years (Bureau of Labor Statistics, 2022a). Web developers typically require a Bachelor's degree, the modal education requirement in the tech field, and are paid on average $78,300 per year, solidly in the standard range for tech positions ($57,910 to $131,490) (Bureau of Labor Statistics, 2022b).

The AI Tool

In this experiment, we use a popular AI-assisted recruitment tool from a leading international company that provides applicant screening software used by a growing number of firms. The software uses Machine Learning and Natural Language Processing to read candidates' interview answers for fit to the position, further, translating answers into scores for personality traits, work-based traits and communication skills. It also provides an overall score out of 100 for each candidate, where 100 is the highest possible score. This tool simulates an interview through a chat-box format, in which the chat-box asks standard interview questions and applicants are invited to type in their responses.

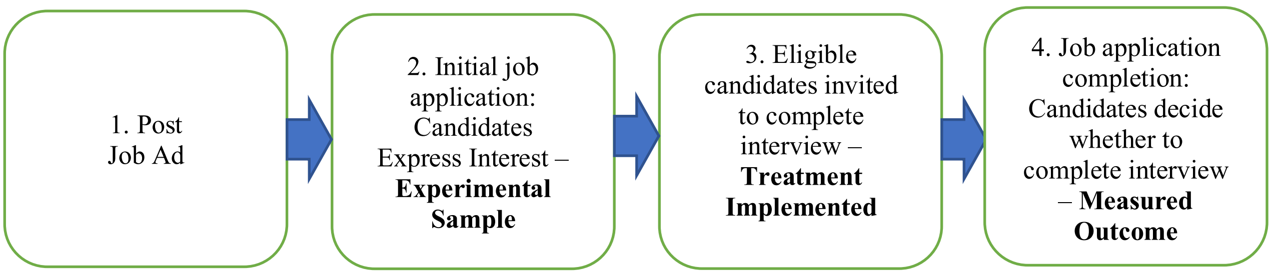
The AI tool used in this paper, like other popular AI-assisted recruitment tools such as HireVue, Humanly, HireScored and Paradox.ai, is marketed to recruiters as being unbiased. The AI-tool provider argues that the training data and resulting tools have gone through extensive testing to check for bias among protected attributes. If bias is identified the machine learning model is updated. However, like the other AI tools, the process through which bias is "avoided" is not transparent (if it even is avoided), as it falls within the black box of the AI and machine learning process and is considered part of the proprietary intellectual property of the AI tool provider. We therefore argue that the AI tool used here is similar to existing AI tools on the market and behavior towards this tool should be no different from behavior towards AI recruitment tools more generally. Further, despite claims of unbiasedness by these and similar tools, it remains unclear how they are perceived by users, as there is limited research on the topic. In this paper we comprehensively study behavior towards AI recruitment tools by studying not only behavior towards this general tool but we also conduct a survey of the wider tech population on perceptions of bias (see Section 4.4 for further details).

4. Field Experiment on the Supply Side: Job-applications in the Presence of AI Assessment

4.1 Experimental Design (Supply Side)

Our supply side experiment (Experiment 1) is embedded in a real recruitment drive for a web developer. Figure 1 provides an overview where we define our experimental sample, when our treatment is implemented, and what we measure as our experimental outcome.

Figure 1: Overview of Experiment 1



In stage 1 of the experiment, we advertised a web developer position across major job sites in the United States, including general job sites (e.g. indeed.com) and specialized tech job sites (e.g. Dice). The job was open to anyone who was based in the United States. Appendix B contains the job advertisement in full. It was posted for 1 month. As is common practice, individuals were invited to express interest in the job by filling out a short set of questions including their demographic information, contact information, education and experience, how they learned web development, and whether they are currently employed (stage 2). Applicants also had to upload their CV. This is sometimes referred to as the initial job application stage. We received a total of 726 unique applications from candidates residing in the United States. Appendix Table A.1 describes this experimental sample. 76.1% of the interested candidate sample are male which is consistent with the job being in a male-dominated industry.

We invited all 726 candidates to complete the online interview in which we implemented the treatment (stage 3). They were sent an email informing them that they had proceeded to the next round of the application process which requires answering a set of online interview questions. The content of this email varied depending on the randomized assignment to treatment. In the **AI-Supply treatment** (henceforth *AI-Supply*), candidates were told that their responses would be evaluated by AI and in particular that "*(…) the questions will be evaluated by Artificial Intelligence (AI) software provided by {redacted}. The AI will read your answers for fit to the position,* *personality, work-based traits and inter-personal skills*." Candidates assigned to the **Human-Supply treatment** (henceforth *Human-Supply*) were sent an identical email except that the email stated *"(...) the questions will be evaluated by our in-person hiring team, who will read your answers for fit to the position, personality, work-based traits and inter-personal skills.*" Candidates were assigned to only one treatment.

The presentation of the online interview was identical between treatments. In all emails candidates were given a link to the assessment. The assessment interface was the same irrespective of treatment (see Appendix Figure A.1). The assessment began by eliciting the candidate's name and email followed by a brief set of instructions including a reminder that the assessment will be evaluated by AI (in the *AI-Supply* treatment) or our in-person hiring team (in the *Human-Supply* treatment). The latter information was the only difference at the assessment stage. All candidates were sequentially asked the same five common interview style questions and were instructed to write between 50-150 words per question.[7] Upon completing the assessment, candidates answered a brief consumer-experience survey eliciting their attitudes towards the assessment such as their perception of bias in and satisfaction with the assessment process. We label this survey the *applicant survey*. In this experiment, our key outcome is the completion of the interview which constitutes the completion of the job application (stage 4). A candidate is defined to have completed the job application if they submit the last interview question. No candidate dropped out between completing the interview and responding to the post-interview survey.

4.2 Hypotheses

For AI recruitment tools to impact behavior, applicants must believe that AI-assisted recruitment signals something about the recruitment process, and that this changes their application behavior in some way. The main area that we hypothesize to be important is beliefs about the gender bias of the AI assessment tool.[8] As discussed in section 2, it is unclear ex ante whether

[7] The questions are: "What do you find most motivates you to achieve results?"; "Where has commitment set you apart from others in your peer group?"; "Please share a favorite experience of working with a team and your contribution to it?"; "What steps do you follow to evaluate a problem before making a decision? Why?"; "Can you give me an example of when your determination has set you apart from others in your peer group?"

[8] In addition to bias, it is also possible that the AI assessment signals to applicants' information about the value of the position in the organization, the number of anticipated applicants, or that the candidate is at an earlier stage of the application process (relative to being interviewed by a HR team). Thus, the use of AI could indicate a lower expected likelihood and value of obtaining the job, and thus lead to a decrease in applications. However, we have no reason to believe, ex ante, that there will be gender differences in the interpretation of this signal or response to that information.

applicants believe or should believe that AI affects gender bias in recruitment. There is a body of evidence showing that women differentially experience worse outcomes in the traditional hiring process, i.e. when evaluated by humans, including in the tech sector (e.g. Feld et al, 2022). However, it is unclear to which extent AI will replicate the biases of humans. There is some evidence suggesting AI could reduce bias (Cowgill, 2018a; Cowgill et al., 2020; Li et al., 2020) with another body suggesting it will entrench biases further, making them not only worse but also more difficult to identify (e.g., Wachter-Boettcher, 2017; Galer, 2019; Yarger et al., 2019; Köchling and Wehner, 2020; Zielinski 2020; Kordzadeh and Ghasemaghaei, 2022; Patty and Penn, 2022). Furthermore, the literature directly testing perceptions of bias in AI is sparse and varied (Chugunova and Sele, 2020), and there is no research about how these perceptions affect labor supply.[9] This suggests two possibilities. The first, which motivates hypothesis 1a, is that women believe that AI will reduce gender bias. It is plausible that women have experience being discriminated by human evaluators while the actual evidence on bias perpetuated by AI is less systematic and generally based on anecdotal evidence (Dastin 2018). If women respond to changes in anticipated bias by changing their application behavior, for example because less bias provides them a better chance at getting the job, conditional on completing the application, we should then see an increase in women's applications with AI rather than human evaluation. Given that we anticipate that more men than women will express interest in this tech position and enter the application stage, the use of AI in recruitment should close the gender gap in completed applications by counteracting some of the gender imbalance in the initial application pool.

Hypothesis 1a: *Compared to men, women are relatively more likely to complete their job application when they are assessed by AI than when they are assessed by humans. This shrinks the gender gap in the pool of completed job applications.*

Second, as argued above it is also possible that the gender effect could be in the opposite direction. The competing hypothesis is also motivated by gender differences in beliefs about the probability of bias in the AI vs. the human treatments, but is instead predicated on the possibility

[9] Marcinkowski et al. (2020) show that believing AI to be more fair than human decision making in university admissions is negatively correlated with students stating they would avoid applying to universities using AI in admissions after hypothetically getting a negative outcome from an AI-assisted university admissions decision.

that women believe that AI poses a greater risk of bias than traditional human recruitment, possibly based on recent media coverage highlighting these concerns.

Hypothesis 1b: *Compared to men, women are relatively less likely to complete their job application when they are assessed by AI than when they are assessed by humans. This increases the gender gap in the pool of job applications.*

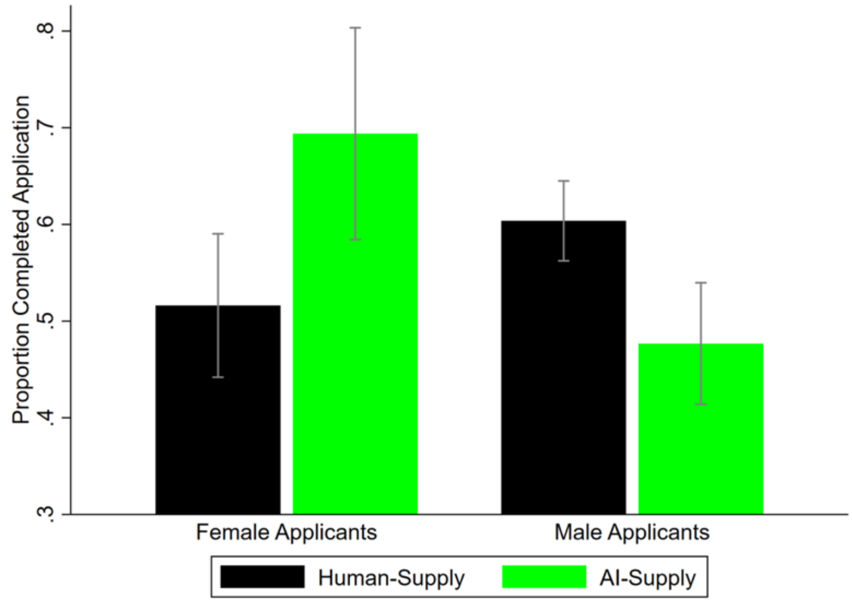
4.3 Supply Side Experimental Findings

The Application Decision

We find strong support for Hypothesis 1a. In Figure 2, we show that the probability of a woman completing the interview stage of the application increases by about 18 percentage points (p=.03), or 35%, when we announce that the evaluation is conducted by AI instead of a human recruiter team, whereas the probability of a man completing decreases by 13 percentage points (p=.01), or 21.5%. This is relative to the *Human-Supply* treatment, in which we find that women are marginally less likely than men to complete (51.6% vs. 60.4%, p=.09).[10]

[10] Overall, 56.5% (410 out of 726) of invited candidates complete the application. The application likelihood is in line with other recent papers using a similar design, which tend to find application completion rates of candidates who showed initial interest at 34.3%-67.8% (Leibbrandt and List, 2018; Flory et al., 2021; Feld et al., 2022). There are no significant differences in application completion rates across treatments (58.3% in *Human-Supply* vs. 52.5% in *AI-Supply*; t-test, p=.15).

Figure 2: Proportion of Candidates Completing Interview by Gender and Treatment



Notes: The figure represents the proportion of candidates of a given gender and treatment that complete the assessment. The left two columns illustrate the application behavior of female candidates and the next two columns represent application behavior of male candidates. 90% confidence intervals are shown.

Table 1 shows that these patterns are robust in a regression framework. Our key variable of interest is the interaction between the treatment and the gender of the applicant. The first column of the model is estimated without controls; we include controls in the second column to disentangle the result for gender from other things that may be correlated with gender and also possibly affected by the treatment. For example, if the use of AI attracts less qualified candidates, who believe they have a better chance with AI evaluation than human evaluation, and women are also less qualified, this may drive our finding.[11] We thus control for the applicant's self-reported type of web design training (University courses, non-university courses, and/or self-taught), years of

[11] We capture candidate quality in two ways: i) the AI-generated interview scores applicants receive across treatments (Figure A.2); ii) and the qualifications of the applicants that complete the interview portion of the application (Table A.2). We do not find that the use of AI systematically changes applicant quality. While we do see some gender differences in qualifications overall, with women being more likely to have university-level web design training and less experience with certain programming languages, we find no change in either the qualifications of men or women when moving to AI evaluation. Furthermore, in our gender-blind human evaluation treatment, described in section 5, we find that evaluators do not judge men's and women's applications differently given these different qualifications, suggesting further that these differences in qualifications do not reflect a difference in quality. In Appendix Table A.3, we show that this lack of differences in qualifications is also true for non-completers, indicating that there is no difference in selection across the two treatments either overall or by gender.

experience in web design, education, and programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular). Additionally, we control for the race of the applicant, as racial minority status may also interact with application behavior in response to AI, as well as controls for the time between when applicants completed their initial expression of interest and received the invitation to the next stage by email. Appendix Table A.4 adds each set of controls individually, showing that one set of controls does not substantially matter.

Table 1: Regression Results, Application Completion by Gender and Treatment

Models	Dependent Variable (1) Application Complete	(2) Application Complete
AI Supply	-0.127***	-0.117**
	(0.046)	(0.046)
Female Applicant	-0.088*	-0.090*
	(0.052)	(0.053)
AI Supply × Female Applicant	0.305***	0.265***
	(0.092)	(0.096)
Controls included	N	Y
Constant	0.604***	0.780***
	(0.0251)	(0.136)
N	726	726
Comparison across coefficients:		
AI Supply + AI Supply × Female Applicant	0.178**	0.148*
	(0.080)	(0.083)

Notes: We use an OLS to estimate the models. The first column reports estimate without controls and controls are added in the second column. The dependent variable is an indicator variable whether the applicant completed the interview assessment. The variable AI Supply is equal to one if the applicant was randomly assigned to the *AI Supply Treatment*. The comparison "AI Evaluation + AI Evaluation * Female Applicant" is the sum of the effect for the coefficient *AI Evaluation* and AI *Evaluation X Female Applicant.* Controls include indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of experience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), time between providing initial information and receiving the email with the interview invitation, and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander). Data are from the experiment 1. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

We see consistent results across both models, indicating that women increase their completion by about 19 percentage points (AI Evaluation + AI Evaluation * Female Applicant, p=.03) whereas men complete the interview portion of the application about 12 percentage points less when AI was used in the evaluation (p=.01). Finally, the difference in difference (which can be interpreted as the impact of AI compared to humans for females relative to males) indicates that using AI increases the proportion of women completing their application relative to men by between 27-30 percentage points. These shifts in application completion generate changes in the demographics of the final applicant pool. 24% of those who initially express interest in the position is female; this drops to 22% for the final applicant pool in *Human-Supply*. On the other hand, the final applicant pool in *AI-Supply* is 29% female (p=.11).

Result 1: *In contrast to the standard hiring evaluation procedure, AI evaluation increases women's application completion rates and decreases men's completion rates. This leads to women completing the application at a rate 27-30 percentage points higher than men when AI is used.*

4.4 Mechanisms behind Application Decision

To study the mechanisms behind application decisions we use both the *applicant survey* - elicited after job applicants complete the assessment – and a separate *general survey* with 129 adults belonging to the US tech labor force.[12] In the general survey, we introduced the survey respondents to the advertised position and the evaluation measure and asked them a series of questions about how they would feel about being evaluated in this way by a hiring team and by AI, in random order. By combining these two surveys, we are able to study both what the relevant population believes about AI-assisted recruitment relative to traditional human-only recruitment, as well as to evaluate how these beliefs might impact selection into these different application types. For example, suppose women typically anticipate greater bias from human evaluation than from AI. In that case, we should see that women report anticipating greater bias from humans than

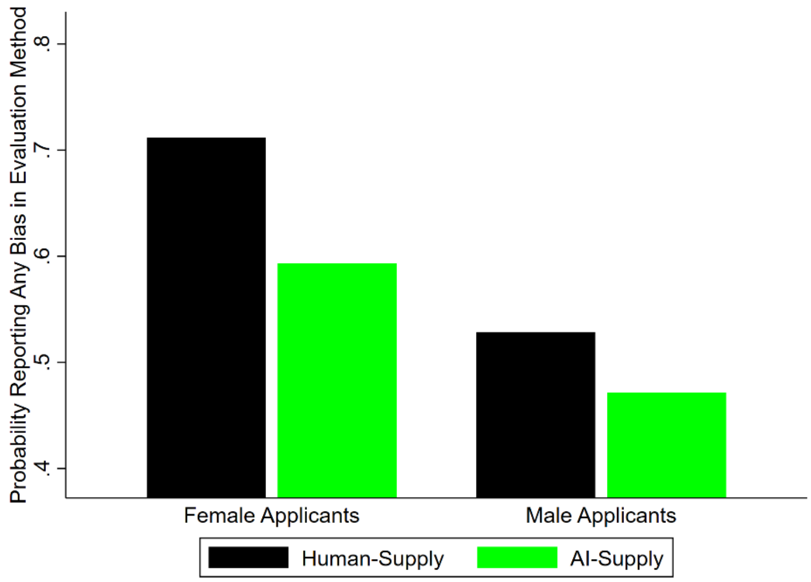
[12] This general survey was approved by the university ethics committee but was not pre-registered. We additionally surveyed 124 adults in the US labor force who are not employed in tech positions. We focus on the tech sample as this is the most relevant sample to our job applicants. Information on the responses of adults not employed in tech positions can be found in Appendix C. While there are some differences between adults employed and not-employed in tech, we find that women still believe that they are more likely to face bias from human evaluation than AI evaluation (t-test, diff=0.19, p=.00), suggesting that women in non-tech fields may increase their application completion rates in response to the shift from human to AI evaluation similarly to women in tech.

AI in the general survey. However, what we then observe in the applicant survey depends on whether they change their application completion behavior based on this anticipated bias. If they do not incorporate anticipated bias into their application completion decision, we should continue to see a gap in anticipated bias between human and AI evaluation similar in magnitude to what is found in the general survey. On the other hand, if they incorporate the anticipated bias into their application decision by completing less when anticipated bias is higher, we should observe a smaller or non-existent gap between anticipated bias from human vs. AI compared to the general survey. This is because those who most anticipate bias from humans should be the ones not applying in the human treatment, and thus they should not be captured in our applicant survey with its selection on only completed applicants.

Figure 3 presents the rates at which male and female respondents to the general survey report anticipating any bias against someone of their own gender from evaluation by either a hiring team or AI. Overall, women in the tech sector report a higher perception of bias compared to their male counterparts, regardless of the type of evaluation performed (p=0.01). We see that women are 12 percentage points (paired t-test p=.05) more likely to express concern about bias from human evaluation than from AI evaluation, indicating that, in the tech population, women anticipate experiencing more bias against themselves from traditional human hiring than from AI evaluation. However, when we ask respondents in the *applicant* survey whether they have any concern about bias from the evaluation method they experienced in their application, we find that the women who have completed the application under human and AI evaluation are equally likely to report being concerned about bias (t-test, diff=-0.07, p=.49). Combined with the finding that women are more likely to complete the application with AI evaluation than with human evaluation, these results indicate that women are discouraged from completing the application in the human treatment due to anticipated bias against them from that evaluation type, and that the lowered anticipated bias from the AI evaluation allows for more women to complete the application when evaluated by AI. Men, on the other hand, show no difference in their anticipated bias from these two evaluation measures.[13]

[13] We also examine perceptions about the value and status of the position depending on the treatment. Specifically, we ask the experimental sample, selected on having completed the application, their anticipated value and status in the organization if they got the position. We asked the survey respondents the same questions about the evaluation method they were asked to imagine experiencing as an applicant. Finally, we also asked the survey respondents to guess the hourly wage offered for the position. In general, we find no consistent evidence that perceptions of value

Figure 3: Perceptions of Bias depending on Evaluation Method and Candidate Gender



Notes: This figure presents the means of respondents to the general survey reporting that they anticipate any bias from either evaluation mechanism.

Result 2: *Women in the tech sector express more concern about bias from human evaluation than from AI evaluation. However, overall, women in the tech sector are more likely to report concerns about bias, regardless of the evaluation type, compared to men.*

5. Field Experiment on the Demand Side: Assessing Job-applications in the Presence of AI

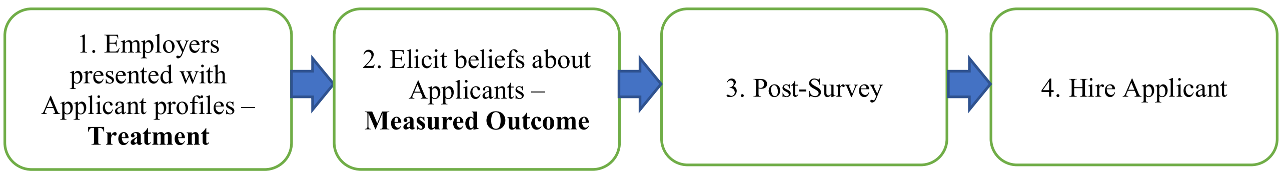
5.1 Experimental Design (Demand Side)

Our demand side experiment (Experiment 2) continues the process of our real recruitment of a web developer. In this section, we turn to the evaluation of the applicants from Experiment 1. Figure 4 provides an overview of Experiment 2 including when our treatment is implemented and what we measure as our experimental outcome. We recruited individuals who work in the tech sector, henceforth called “evaluators”, to evaluate our applicants. Such outsourcing of recruitment decisions is common: as of 2015, nearly two thirds of companies in the US outsourced at least part of their recruitment activities (SHRM 2015). Further, freelancing is increasingly used as a method

or status of the position varies across the treatments, could drive our main result, or could conflict with our findings on bias (Appendix Table A.5).

of filling important business needs, including recruitment (Dua et al., 2022), and recruiting services were listed as a fast-growing industry for freelancers by Forbes in 2021 (Stahl, 2021). Evaluators were recruited using a panel service provided by Qualtrics. We paid each evaluator US$ 20 to complete the 20-minute evaluation task. In stage 1, the evaluators were given a brief description of the context and their task. To ensure they understood the instructions, they could not proceed until they responded correctly to several comprehension questions.

Figure 4: Overview of Experiment 2



In stage 2, evaluators evaluated 4 profiles, 2 female and 2 male, taken from a random subset of applicants and appearing in a random order. For each applicant they were shown the responses to the assessment questions and information taken from the applicant's CV (education, years of web development experience etc.). Evaluators then had to rate each candidate based on how well they thought they would perform if hired as a web developer, on a scale between 0-100 (where 100 is high). To incentivize evaluators, they were truthfully told that their evaluation score will be used to select who will be hired. More precisely, they were told that "*We will use the answers you provide to decide whether the person will move on to the next step of the hiring process. The decisions you provide here will have real outcomes both for us hiring a good web developer and for the individuals who have applied for this job. We want you to help us pick the best web developer for our project. In order to decide whether or not we should ask an individual for an interview, we want you to rate each applicant on how well you think they would perform if hired as a web developer*."

As specified in the pre-analysis plan, to increase the number of evaluations of an individual applicant (and thus increase power), Experiment 2 used a stratified random subset of the total number of applicants completed. This subsample comprises 300 applicants, 202 male candidates and all 98 female applicants. This subsample maintains the distribution of applicant characteristics and AI scores within gender and treatment but oversamples female applicants and the *Human-*

Supply treatment (see Appendix Table A.6 for the balance test comparing the sample shown to evaluators relative to the full sample of completers).

Evaluators were randomized into one of three treatments, receiving different sets of information on which to base their evaluation of the applicants. In the control, called **Human-Demand** treatment (henceforth *Human-Demand*), for each of the four applicants, evaluators were shown responses to the assessment questions and a short applicant profile. The applicant profile included the applicant's first name, demographics and information about web development experience. See Table 2 for an example profile, in which evaluators usually observed a CV with some basic information including the name, providing a signal of gender. The **AI-Demand** treatment (henceforth *AI-Demand*) is identical to the control except evaluators were also informed of the evaluation score (score out of 100) given by the AI software. This treatment allows us to test the combined effect of AI and humans when making a judgement about candidate quality relative to humans alone. Lastly, the **no-name** treatment (henceforth *No-Name*) is identical to the control except that the profile of the applicant excluded the applicant's first name. This variant allows us to establish how different male and female applicants are evaluated when their gender is not known and thus how much any gender gaps found in the other two treatments is the result of evaluators knowing and making decision based off the applicant's gender. In each treatment, evaluators also received the applicant's responses to the interview questions.

Table 2: Example of applicant profile

Name	Andrew P
Highest Education level	Some college
Years of Web Development Experience	5
Learned Web Development from	University
What coding languages do you have experience using?	Java, CSS

Note: This table provides an example of an applicant profile in the *Human-Demand* Treatment.

After evaluating the applicants, all evaluators completed a short survey, i.e. stage 3. The survey collects additional information related to the research (e.g., whether they think women would perform worse on these kinds of jobs, job experience, demographics etc.). This survey is used to help understand why differences between AI and human evaluators may exist. Finally, in stage 4, we used the evaluations provided by the human evaluators and the AI software to determine a short-list of applicants considered for hire. Offers were extended to multiple candidates.

The total sample of evaluators is 507. We have 202 evaluators in the *Human-Demand* treatment, 145 in the *AI-Demand* treatment, and 156 in the *No-Name* treatment.[14] As each human evaluator evaluates 4 applicants, this results in a total sample of 2017 individual evaluations. Appendix Table A.7 details key characteristics of the evaluator sample. As required to participate in the experiment, 100% of the evaluators work in the technology industry in the US. Just over 66% of the sample are male, 43% of the sample have achieved at most a 4-year college while 30% have a post graduate degree. Around 96% are employed (90% full time and 5% part time), and the average age is 43. The evaluation sample is comprised of managers (25%), senior managers including directors and business owners (22%), software developers including web developers (16%) and consultants and general tech workers who work broadly in software development (22%). Importantly, 84% are currently or were involved in hiring decisions in their job.

5.2 Conceptual Framework and Hypotheses

In this subsection, we outline and explain the hypotheses for experiment 2.

Hypothesis 2: *Evaluators score women lower than men in the Human Demand Treatment. The gender gap is smaller in the no name treatment.*

In Hypothesis 2 we argue that there are gender differences in the evaluation score when gender is known and this difference will be minimized when gender is unknown. This hypothesis is motivated by existing evidence on discrimination in the selection of job applicants both in STEM and the labor market more broadly. However, any disparities in evaluations between men and women in the *Human-Demand* treatment may be due to male and female applicants having different qualifications which may be valued differently by the evaluators, or may be due to evaluators having biased beliefs about applicants' ability based on their gender. To identify

[14] As we expected a greater variance in the Human-Demand treatment, we collected more observations in this treatment to maximise power (see Czibor et al., 2019 for a detailed discussion on this topic). The differences in the number of observations were prespecified in our pre analysis plan. Further, we prespecified 500 evaluators and collected 507 observations. The additional observations are the result of our data collection partner failing to close the survey when the target sample was reached. For full transparency we always include the 7 additional evaluators. Due to a software issue 11 observations were not recorded.

whether any disparities in evaluations are due either entirely or partially to gender, we also measure whether the disparities change when gender is not easily knowable, i.e. in the *No-Name* treatment. We can attribute any difference in gender disparities between the *Human-Demand* and *No-Name* treatments to gender bias in the evaluation.[15]

Our next set of hypotheses focus on the *AI-Demand* treatment. These hypotheses are premised on the expectation that there are no gender differences in the score predicted by AI, which is shown to be true in Figure A.3(p-value=0.27 and at the mean, male=30, female=32, p=.27). In this case it is not clear what the effect of providing the AI score to evaluators will be. We argue there are two plausible scenarios. First, an existing set of literature in various contexts has shown that AI tools generate better outcomes than humans (Chalfin et al., 2016; Cowgill 2018a; Stevenson and Doleac, 2022). We argue that if a sufficient proportion of evaluators believe the AI score is useful, they could use this information to update their beliefs about the (relative) quality of female applicants. This could consequently reduce gender difference in the evaluation score (i.e., the differences discussed in Hypothesis 2).

Hypothesis 3a: *Gender differences in the evaluation score in the AI-Demand treatment will be reduced compared to the Human-Demand treatment.*

Alternatively, despite evidence that AI can perform better than humans, it is not clear whether humans actually hold this belief. There is growing evidence that people are algorithmic averse, meaning they have an aversion to using or trusting algorithms even when they have been shown to perform better than humans (Dietvorst et al., 2015; Burton et al., 2020; Jussupow et al., 2020). This is especially true if people believe the AI is imperfect in some way (Dietvorst et al., 2015). If a sufficient proportion of evaluators are algorithm averse, they may not believe the AI score will be useful and therefore potentially disregard it. This is consistent with recent evidence that practitioners are often unaware of how AI works, the specifics of the AI tool they or their organization uses, and the impact the AI tool may have on diversity and discrimination outcomes

[15] We acknowledge that the gap between the *Human-Demand* and *No-Name* treatments could also occur if gender is informative about performance in web development in the absence of objective performance metrics beyond the information already provided to them about applicant backgrounds including education and experience. While we do not have evidence on our applicants' web developer skill in order to directly test this, Feld et al. (2022) shows that men and women who apply for a job in a related field, programming, have similar skill levels as tested by an objective measure, though evaluators believe men to be better than women.

(Adamovic et al., 2022). Furthermore, evaluators are also subject to the media coverage of AI, arguing that AI is biased against minorities, including women (Wachter-Boettcher, 2017; Galer, 2019; Zielinski 2020).

In Appendix D, we also show that certain conditions about evaluators beliefs about the true positive and false positive rates of the information from the AI, plus the gender bias they anticipate coming from the AI system, are necessary to reduce the gender gap in evaluations. Specifically, we show that evaluators must have substantial confidence in the fact that a positive signal from the AI is really positive in order to have equivalent posteriors for men and women when they start with biased priors. Evidence from the literature suggests that these qualifications for the unbiased incorporation of AI information may not be met, either because employers are not that confident in the information provided by AI (Dietvorst et al., 2015; Cowgill, 2018a; Glaeser et al., 2019; Stevenson and Doleac, 2022; Burton et al., 2020; Jussupow et al., 2020) or they believe the AI to be biased in some way (Wachter-Boettcher, 2017; Galer, 2019; Zielinski 2020). Based on this and if AI aversion is common in our sample, we would expect that human evaluators will not use the AI information and thus we predict the following:

Hypothesis 3b: *Evaluators score men higher than women in the AI-Demand treatment.*

5.3 Demand Side Results

In the *Human-Demand* treatment, we find that on average evaluators score applicants 73.06 out of 100, with a median of 78 and 50% of scores falling between 64 and 85. We find gender differences corroborating H2: men are scored on average 74.51 whereas women are scored only 71.60, a difference of 0.15 standard deviations (diff=2.90, t-test, p=.03). The gender gap is more pronounced towards the right tail of the distribution (Figure 5, panel A): men are 6.8 percentage points (p =.04) more likely to be in the top 25% and 7.73 pp. (p<.001) more likely to be in the top 10% of scores while there is no difference in the likelihood of a man vs. woman scoring in the top 50% (diff=0.01, t-test, p =.81). In Table 3 we present the corresponding regression analysis for the mean, using OLS regressions (column 1), and the 50^{th}, 75^{th}, and 90^{th} percentiles using quantile regressions (columns 2-4), with the evaluator score as the dependent variable, the gender of the applicant as the main variable of interest, and we control for applicant characteristics and AI score. Both show consistent results across all formulations. Appendix Table A.8 and A.9 show these analyses with and without controlling for the AI score, indicating the results here are not dependent on controlling for the AI score. The regression analysis supports the finding, with women scoring 3 points lower than men on average and most of the difference occurring above the 50th percentile.

In the *AI-Demand* treatment, the results are substantially different. First, we find that on average evaluators provide applicants a relatively low score of 56.27 with a median of 55 (50% of scores are between 37 and 77), which is likely due to lower scores provided by the AI (the mean AI score is 31.25). Supporting H3a, we find no significant gender differences in *AI-Demand* (p=.61): men are scored only 1.07 points higher than women, equivalent to a mere 0.04 standardized difference. This result is consistent with the actual AI score, which produced no gender differences (male=30, female=32, p=.27, see also Figure A.3).

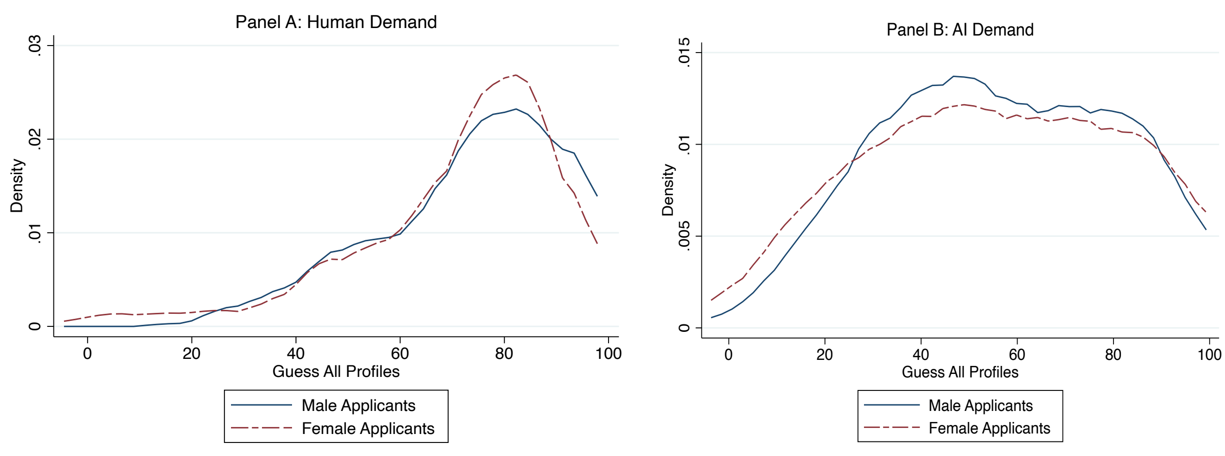
Table 3: Human Evaluators vs Artificial Intelligence

Models	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Human-Demand				AI-Demand			
	OLS	Quantile Regression			OLS	Quantile Regression		
		50th	75th	90th		50th	75th	90th
Female Applicant	-3.025**	-0.587	-3.225***	-3.020**	-0.398	0.399	2.223	4.618*
	(1.224)	-1.392	(1.049)	(1.479)	(1.719)	(2.668)	(3.074)	(2.360)
Applicant Controls Included?	Y	Y	Y	Y	Y	Y	Y	Y
AI Score Included?	Y	Y	Y	Y	Y	Y	Y	Y
Constant	57.42***	52.61***	74.59***	76.99***	28.24***	29.21***	59.13***	78.36***
	(4.983)	-5.838	(4.686)	(4.811)	(8.360)	(7.842)	(14.58)	(16.50)
N	805	805	805	805	591	591	591	591

Notes: In columns 1 and 5 we use an OLS to estimate the models with robust standard errors clustered at the evaluator level. In columns 2-4 and 6-8 we use quantile regressions at the 50th (columns 2 and 6), 75th (columns 3 and 7) and 90th (columns 4 and 8) percentiles. Applicant controls include indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of eYperience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), time between providing initial information and receiving the email with the interview invitation, and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander). The dependent variable is the score given by evaluators in the *Human-Demand* (columns 1-4) and AI-Demand (columns 5-8) treatments. N is the number of observations in each treatment. As discussed in Section 5.1, the number of observations differ between the two treatments. Data are from the experiment 2. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

We also find that the fraction of men and women in the top 50%, 25%, and 10% is almost identical (50%: diff=0.03, p =.49; 25%: diff=0.01, p =.87; 10%: diff=0.03, p =.27). This is visualized in Panel B of Figure 5 which shows that the distributions of scores for men and women are similar.[16] Similarly, these results are supported by columns 5-8 of Table 3, which shows that women are scored the same as men both at the mean and top 25% and marginally better at the top 10% of the distribution.

Figure 5: Distributions of Evaluations for Male and Female Applicants in Human and AI-Demand



Notes: This figure presents the density of evaluations in the *Human-Demand* and *AI-Demand* treatments by the gender of the applicant.

To understand whether gender gaps observed in the *Human-Demand* treatment are consistent with gender bias, we turn to our *No-Name* Treatment, where applicant gender is unknown. Supporting H3, we find that in the *No-Name* treatment, unlike in the *Human-Demand*, there are no gender differences in scores, neither in means (t-test, diff=0.18, p=.92) nor in the top 50% (diff=0.03, p=.32) or 10% (diff=0.02, p=.48) of the distribution. Columns 1-4 of Table 4 present the results in regression form either using OLS (column 1) or quantile regressions at the 50th (column 2), 25th (column 3) and 90th (column 4) percentiles. The results are the same: men and women do not receive different scores when names, and thus gender, are not provided to the evaluators. Furthermore, we can see in columns 5-6 of Table 4 that, compared to the *No-Name* variant, men are relatively more favored in the *Human-Demand* Treatment, while no more favored in the *AI-Demand* Treatment, particularly at the right tail of the distribution. At the 90th percentile, the difference in the gender gap between the *Human-Demand* and *AI-Demand* treatments is 6.6

[16] Figure A.3 provides these distributions for the no-name treatment and pure AI-generated scores.

points (t-test p=.02).[17] Appendix Tables A.10 and A.11 present these analyses with and without controlling for the AI score and shows that the results are consistent across all formulations.

Result 3: *Women are scored worse than men by human evaluators in Human-Demand, particularly at the right tail of the distribution; this gap in scores is driven by the evaluators knowing the applicants' gender. Providing AI scores to evaluators results in women not being scored worse either at the mean or right tail.*

The gender gap in evaluations in *Human-Demand* and corresponding lack of gender gap in evaluations in *AI-Demand* show that at least some evaluators are influenced by the AI score in their evaluation.[18] Interestingly, we find that the large majority of evaluations are substantially different from the AI-score: only 8% of evaluations in *AI-Demand* are within 5 points of the AI evaluation score provided to the evaluator. This shows that evaluators neither simply copy the score nor completely rely on the AI evaluation score and suggests that they find the score to be informative but not conclusive.

With regards to gender, we observe no difference in the likelihood of the evaluation equaling the AI score by the gender of the applicant (t-test, diff=0.003, p=.88) and that there is no difference in the gap between the AI score and the evaluation by gender, whether in raw (diff=-0.79, p=.69) or absolute (diff=0.11, p=.95) terms. However, there is some suggestive evidence that evaluators' disparate priors' factor into their largely equivalent posteriors: men are marginally more likely than women to be given an evaluation higher than the score provided by AI (diff=-0.06, p=.07), while women are significantly more likely than men to be given an evaluation lower than their AI score (diff=0.06, p=.02), reflecting the disparities in evaluations we see without AI information.

[17] By using *No-Name* as a baseline, we can speak to whether men are being favored or women are being harmed when gender is known in *Human-Demand*, following Feld et al. (2016). Our results show that men do not receive an increase in score when moving from *No-Name* to *Human-Demand* whereas women receive a decrease as gender is revealed. However, because we did not elicit evaluators' beliefs about the gender of the applicants in the *No-Name* treatment, we cannot differentiate between women being harmed when gender is known or evaluators just assuming applicants are overwhelmingly male in the *No-Name* treatment.

[18] The analysis and findings for the remainder of this subsection were not specified in the pre-analysis plan.

Table 4: No-Name Evaluation and Comparison to Human-Demand and AI-Demand

Models	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	No-Name Treatment				Evaluation			
	OLS	Quantile Regression			OLS	Quantile Regression		
		50^{th}	75^{th}	90^{th}		50^{th}	75^{th}	90^{th}
Female Applicant	2.077	1.803	-0.436	-0.377				
	(1.875)	(1.785)	(1.038)	(1.485)				
Female Applicant					1.206	-0.144	-0.620	-0.0403
					(1.547)	(1.710)	(0.847)	(1.345)
Human-Demand					1.138	-2.700	-1.904**	0.00476
					(1.779)	(1.661)	(0.808)	(1.131)
AI-Demand					-16.42***	-22.94***	-12.97***	-6.492***
					(2.258)	(1.938)	(1.204)	(1.931)
Female Applicant × Human-Demand					-3.879**	-0.131	-2.550**	-3.595*
					(1.808)	(2.088)	(1.201)	(1.932)
Female Applicant × AI-Demand					-1.792	0.590	-0.241	3.027
					(1.959)	(3.249)	(2.681)	(2.845)
Gender Gap – Human-Demand vs. AI-Demand					p=.22	p=.81	p=.39	p=.02**
Applicant Controls	Y	Y	Y	Y	Y	Y	Y	Y
AI Score	Y	Y	Y	Y	Y	Y	Y	Y
Constant	58.87***	70.97***	81.99***	86.30***	53.87***	63.02***	74.47***	83.53***
	(8.547)	(8.121)	(5.588)	(5.231)	(4.254)	(4.253)	(2.327)	(3.819)
N	621	621	621	621	2017	2017	2017	2017

Notes: In columns 1 and 5 we use an OLS to estimate the models with robust standard errors clustered at the evaluator level. In columns 2-4 and 6-8 we use quantile regressions at the 50^{th} (columns 2 and 6), 75^{th} (columns 3 and 7) and 90^{th} (columns 4 and 8) percentiles. Applicant controls include indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of experience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), time between providing initial information and receiving the email with the interview invitation, and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander). The dependent variable is the score given by evaluators in the no-name treatment (columns 1-4) and all treatments (columns 5-8) treatments. Gender Gap – *Human-Demand* vs. *AI-Demand* provides the t-test comparing the gender gap in *Human-Demand* vs. *AI-Demand*. N is the number of observations in each regression. Data are from the experiment 2. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

A closer look at key evaluator characteristics reveals interesting evaluation patterns. We focus on the following characteristics: (i) evaluator beliefs about web development skills in the general population, (ii) evaluator gender, (iii) evaluator age, and (iv) experience in hiring web developers. We argue these are key characteristics for the following reasons. Feld et al. (2022) show that beliefs about skill in the general population potentially explains gender disparities in evaluations for a similar tech role, programming. Evaluator gender and age can also affect evaluation by indicating experience, gender attitudes, and acceptance of AI technology, which may then impact how AI is integrated into the decision-making process. Finally, we consider whether the evaluators have any prior experience in hiring web developers, as a more experienced evaluator may have more accurate beliefs about the relative skills of men and women in web development.

Table 5 presents regressions of the evaluator score on applicant gender and whether the evaluation is from the *Human-* or *AI-Demand* treatment, split by the above evaluator characteristics. In columns 1-2 we restrict the same to those who believe that men are more skilled than women, while columns 3-4 restricts the sample to those that believe there is little or no difference. [23] Column 1 and 2 show that evaluator beliefs about the web development skill of the general population are correlated with gender differences in the evaluator score in the *Human-Demand* treatment, but we find no such relationship for the *AI-Demand* treatment. In particular, evaluators who believe that men are more skilled at web development than women evaluate women 5.62 points worse than men in *Human-Demand* (t-test: p=.00). This is equivalent to a 7.2% or 0.31 s.d. decrease from the mean evaluation men receive. In contrast, when they are provided with AI scores, these evaluators do not evaluate men and women differently (t-test: diff=-0.57, p=.76), indicating that this group responds to the AI information. On the other hand, the evaluators that report believing that men and women are equally capable of web development in the population evaluate men and women equally in both *Human-Demand* (t-test: diff=-1.58, p=.33) and *AI-Demand* (t-test: diff=-1.31, p=.52) (see column 3 and 4). The patterns for the other evaluator characteristics are less pronounced.

[23] We define an evaluator as believing men are better than women in web design if they answer 60 or above to the question "Among all people living in the United States (regardless of their profession), do you think women or men are, on average, more skilled at web development? Please answer on a scale that ranges from 0 'women are more skilled' to 100 'men are more skilled'". The cut-off of 60 indicates a clear indication that the individual believes men are better than women, allowing for some wiggle-room around 50 which would indicate anticipating equal skills between the two groups. About 40% of individuals fall into this category. We get similar results if we make the cut-off at 51.

Finally, we briefly report the extent to which evaluators vary in their deviations from the AI score, which is suggestive of the extent to which they rely on the AI score. We observe that evaluators who believe that men are more skilled than women in web development and evaluators with prior experience in hiring web developers both systematically deviate more from the provided AI score in their evaluations (Appendix Table A.12, columns 1-2) and that the deviations do not differ across applicant gender (Appendix Table A.12, columns 3-6). This finding complements prior work indicating that those with greater experience in a decision task are more averse to relying on AI tools, even if it can improve their decision-making (Burton et al., 2020).

Result 4: *The closing of the gender gap in evaluations in AI-Demand is more pronounced for evaluators who hold gendered beliefs about relevant job skills. While these evaluators' assessments deviate more from the AI score than other evaluators' assessments, their deviations are not different for female and male applicants.*

6. Bringing the Supply and Demand Side Together: Labor Market Analysis

In the prior two sections, we show that the use of AI in recruitment in a male-type environment increases the proportion of applicants that are female and increases the evaluation score of female applicants relative to male applicants. In this exploratory section, we show how these two forces combine to generate shifts in the diversity of applicants in this small-scale labor market.[24] In particular, we show the impact that these two separate forces have on the gender composition of what we consider the "short-listed" applicants – those that have been evaluated as being in the top of the distribution and would most likely be considered for a job offer. Identifying the gender composition of this part of the distribution is important to understand the impact that AI will have on the gender composition of hired workers.

Our design benefits from having applicants from both supply treatments evaluated by evaluators in both demand treatments, allowing us to identify not only how these supply and demand elements combine to change the gender composition across the distribution, but also to decompose those changes into supply- and demand-shifts. To do this, we construct a sample of applicants with evaluations for all four categories (*Human-Supply/Human-Demand*, *Human-Supply/AI-Demand, AI-Supply*/*Human-Demand*, and *AI-Supply/AI-Demand*) that maintains the distribution of applicants from the supply side treatment while assigning the evaluations from the

[24] The analysis and findings in this section were not specified in the pre-analysis plan.

appropriate demand side treatment. Thus, for example, the *AI-Supply*/*Human-Demand* group has the distribution of applicants by gender and qualifications matching what was found in the *AI-Supply* treatment with the evaluations by gender and qualifications for applicants in the *Human-Demand* treatment. Appendix E describes the construction of this sample. With this new sample, we estimate the fraction of the applicant pool that is female across the different evaluation distributions.

Figure 6 shows that the fraction of female applicants decreases in all applicant-evaluation pairings as the evaluation quantile increases past the 50th quantile. However, we find that there are substantial differences across applicant-evaluation pairings. By comparing the *Human-Supply/Human-Demand* group (solid black line) with the *AI-Supply/AI-Demand* group (dashed green line), we find that the fraction of applicants that are female decreases with the evaluation quantile much quicker in a world without AI in recruitment compared to a world with AI in recruitment. Specifically, in a world with AI, applicants at the 50th percentile are 8.6 pp. more female (p=.00), applicants at the 75th percentile are 6.9 pp. more female (p=.00) and applicants at the 90th percentile are 7.7 pp. more female (p=.00) than in a world without AI. These changes range from an increase in the fraction of women by 30% at the 50th percentile to 160% at the 90th percentile over no-AI levels.

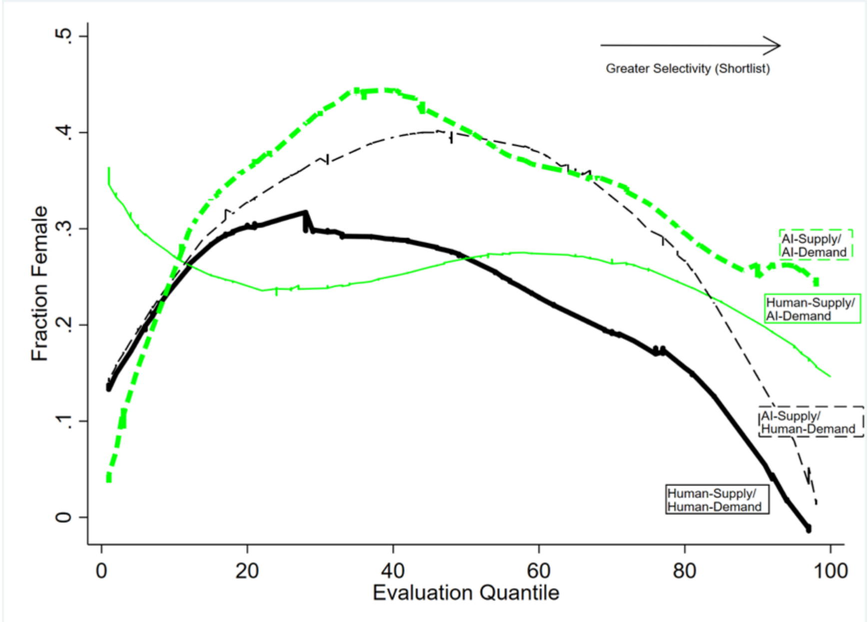
We can also evaluate how much this increase in gender diversity in the top nth quantile is driven by changes in application behavior when told the AI will evaluate them and how much is driven by changes in evaluations when evaluators are provided with the AI scores. In Figure 6 by comparing the solid line and dashed line of a particular color, we can evaluate, within an evaluation type, how much of the gender differences are driven by applicant behavior. Here, we see differences in the impact of applicant behavior across the distribution of evaluations – at the 50th percentile, the entire difference between the world with and without AI is driven by applicant behavior, whereas at the 90th percentile, applicant behavior is a much less important driver of differences in outcomes.

Table 5: Evaluations in Human- and AI-Demand by Evaluator Characteristics

Models	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
	Believe Male>Female		Believe Male≤Female		Female Evaluator		Male Evaluator		Below Median Age		Above Median Age		No Web Developer Hiring Experience		Web Developer Hiring Experience	
Female Applicant	-5.42***	-5.619***	-0.236	-1.576	-0.105	-0.990	-2.503	-3.189**	-0.601	-1.887	-3.517*	-3.723**	-1.193	-2.351	-2.127	-2.661
	(1.900)	(1.889)	(1.631)	(1.608)	(2.200)	(2.115)	(1.545)	(1.535)	(1.719)	(1.708)	(1.810)	(1.766)	(1.676)	(1.611)	(1.798)	(1.774)
AI-Demand	-13.38***	-13.11***	-19.50***	-19.66***	-18.45***	-19.00***	-15.77***	-15.65***	-17.92***	-18.32***	-16.62***	-16.25***	-22.43***	-22.40***	-12.32***	-12.24***
	(3.213)	(3.189)	(2.512)	(2.449)	(3.547)	(3.482)	(2.410)	(2.415)	(2.638)	(2.618)	(2.923)	(2.901)	(2.443)	(2.438)	(2.994)	(2.961)
Female Applicant × AI-Demand	5.44**	5.04**	-0.716	0.267	-0.931	-0.766	2.728	3.090	1.927	2.809	1.469	1.225	0.176	0.395	3.110	3.486
	(2.392)	(2.339)	(2.595)	(2.441)	(3.533)	(3.247)	(2.083)	(2.023)	(2.449)	(2.321)	(2.780)	(2.692)	(2.594)	(2.423)	(2.576)	(2.469)
Applicant Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Evaluator Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
AI Score	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Constant	66.13***	60.86***	54.58***	45.71***	58.40***	49.03***	65.34***	58.68***	52.37***	45.68***	69.27***	62.33***	58.97***	49.09***	68.90***	63.87***
	(7.443)	(7.727)	(6.111)	(6.210)	(7.942)	(8.048)	(5.544)	(5.746)	(6.729)	(6.803)	(6.420)	(6.485)	(6.302)	(6.539)	(7.287)	(7.246)
N	545	545	843	843	470	470	918	918	716	716	672	672	657	657	731	731

Notes: We use an OLS to estimate the models with robust standard errors clustered at the evaluator level. The odd columns include applicant and evaluator controls; the even columns report estimates with controls for the AI score added. Applicant controls include indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of experience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), time between providing initial information and receiving the email with the interview invitation, and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander). The evaluator controls include an indicator for the evaluator believing men are better than women at web development (except in columns 1-4), an indicator for evaluator gender (except in columns 5-8), an indicator for the evaluator being above the median age (i.e. born before 1981) (except in columns 9-12), and an indicator for having prior web development experience (except in columns 13-16). The dependent variable is the score given by evaluators in the *Human-Demand* and *AI-Demand* treatments. The omitted category is evaluations given to male applicants in the *Human-Demand* treatment. The sample in columns 1-2 is evaluators identified as believing men are better at web development than women. The sample in columns 3-4 is evaluators identified as not believing men are better at web development than women. The sample in columns 5-6 are female evaluators. The sample in columns 7-8 are male evaluators. The sample in columns 9-10 are evaluators born after the median birth year of 1981. The sample in columns 11-12 are evaluators born before the median birth year of 1981. The sample in columns 13-14 are evaluators who do not report having prior experience hiring web developers. The sample in columns 15-16 are evaluators who do report having prior experience hiring web developers. Data are from the experiment 2. Significance levels are *** p<0.01, ** p<0.05, * p<.1

Figure 6: Fraction of Females at Each Evaluation Quantile



Notes: This figure shows the fraction of a particular quantile (x-axis) that is female (y-axis) for 4 simulated samples based on results from Human-Supply and *Human-Demand* (solid black), *AI-Supply* and *Human Demand* (dashed black), *Human-Supply* and *AI-Demand* (solid green), and AI-Supply and AI-Demand (dashed green). The distribution of evaluations for each applicant-evaluation treatment pair is rescaled into quantiles, and then the gender composition of each 1-percentage quantile is calculated. Then the distributions are estimated using a Lowess estimation.

Alternatively, we can consider the impact that providing AI scores has on evaluations by applicant gender, holding constant applicant behavior, by comparing the black and green lines within a line type. Specifically, by comparing the black and green solid lines we consider the impact of different evaluation types within the *Human-Demand* applicant pool. We find that at the 50^{th} percentile there is no difference in the gender distribution, suggesting that at the middle of the distribution providing evaluators with applicants AI scores does not change beliefs. However, when moving towards the right of the distribution this gap grows, indicating that providing AI scores has a greater effect on the gender diversity of the top n^{th} of applicants as the quantile increases past 50. This pattern is replicated in the AI treatment applicant pool (comparing black and green dashed lines) and indicates that the effect of using AI in recruitment increases in importance with the selectivity of the recruitment process, whereas the effect on applicant behavior decreases in importance with selectivity.

Result 5: *Shifting from traditional human-only assessment to an AI-assisted assessment more than doubles the fraction of women at the top of the distribution. Both applicant behavior and evaluator behavior significantly contribute to this shift.*

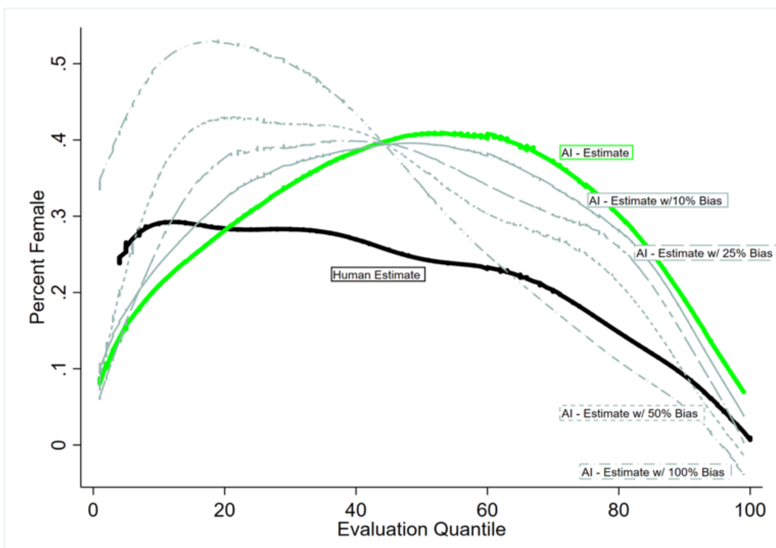
7. Discussion – How Much Bias is Too Much Bias?

We find no gender gaps in the distribution of scores generated by the AI tool we use (see for example, Figure A.3). Thus, any changes across treatments in evaluations come from human, rather than algorithmic, sources. However, there is great concern that some AI tools are biased against minority groups such as women. In this section, we estimate how our results change if we use AI that has increasing levels of bias against women, in terms of the overall percentage of the top applicants that would be female. This allows us to estimate how the use of biased AI tools could impact diversity. Appendix F describes the construction of these estimates and Figure A.4 compares the actual sample, as constructed in section 6, to the estimated sample.

To do this back-of-the-envelope calculation, we assume that neither the applicants nor the evaluators change their behavior in response to the AI tools as the bias changes. For applicants, this is a weak assumption as the applicants are not informed during the assessment about the (lack of) bias in the AI tool used, so changing the level of its bias should not change their behavior. These applicants brought to the application whatever perceptions of bias in AI vs. human evaluations they had already formed, and we did not give them information that would have changed these beliefs. For evaluators, the assumption is stronger as it is positing that evaluators, upon seeing AI scores that differ between men and women, will not place less weight on the AI score.

Figure 7 shows outcomes for biases against women of 10%, 25%, 50%, and 100%. While the black and green solid lines provide the baseline estimates for the *Human-Supply*/*Human-Demand* and *AI-Supply/AI-Demand* cases, the dashed grey lines indicate the impact of a 10%, 25%, 50%, and 100% bias against women in the AI scores on the fraction female in the *AI-Supply/AI-Demand* case. We can see that it requires a substantial bias against women, at 25%, for outcomes for women to be worse with AI than without, at any quantile above 50%, (i.e., the group of applicants most likely to be considered for jobs). Furthermore, AI with a 25% bias against women only does worse than no AI at the most extreme right tail of the distribution (i.e., for the most selective jobs). This follows from our results in Section 5, which finds that towards the center of the distribution almost all of the impact of AI on diversity comes through applicant behavior, which we reasonably assume to be unchanged by increased bias in this exercise, while towards the right tail of the distribution more of the impact comes from changes in the evaluator behavior, which is where our estimation strategy allows for bias to filter through. Even a bias against women of 100% in the AI score would still generate greater diversity in applicants scored at the 50th percentile than was attainable without AI, specifically because there is such a strong impact of AI on application behavior. This suggests that substantial bias against women would have to be generated by the AI software to worsen diversity outcomes relative to recruitment with no AI, and there the impact is primarily on the applicant pools for the most selective jobs.

Figure 7: Estimated Fraction Female Across Evaluation Quantiles, by Evaluation Type and Bias of AI



Notes: This figure shows the simulated fraction of a particular quantile (x-axis) that is female (y-axis) for 5 simulated samples based on results from Human-Supply and *Human-Demand* (solid black), AI-Supply and AI-Demand (dashed green), and AI-Supply and AI-Demand with 10%, 25%, 50%, and 100% bias against women in the AI scores used to calculate AI-Deman (dashed grey lines).

8. Conclusion

The last 50 years have been marked by radical advancements in Information Technology (IT) including the widespread adoption of the internet and the development of increasingly sophisticated software. These advancements have transformed labor markets often with differing impacts on minorities. For example, online job boards and other digital platforms can make it easier for minorities to find and apply for jobs, while e-learning and online training programs can reduce barriers to develop skills needed to succeed in the modern workforce. On the other hand, there are concerns that IT may exacerbate existing inequalities in the labor market. For example, automation and other forms of IT-driven productivity growth may displace certain types of jobs traditionally held by minorities, such as those in manufacturing and other low-skilled occupations. There is also the potential for AI and machine learning to perpetuate and even accentuate the bias that exists in society, especially when the data used to train these models reflect the biases of the past. This can result in unfair treatment of minorities in the job market and other areas. As AI tools such as those used in the hiring process become increasingly prevalent it is vital to understand their impact on the labour market. We present the first field experimental study that assesses the impact of such AI tools on both the demand and supply of minority job candidates.

It is important to consider the multiple impacts these technologies can have on labor markets. In our study, we examine the impacts on supply (applicants) and demand (evaluators/employers). This renders it possible to comprehensively estimate how the diversity of the candidate pool changes and from which side of the market drives the impact. Our study shows significant effects on both the supply and demand sides of the labor market, with the greatest impact observed among the most qualified applicants.

Importantly, there is still a significant human element in even the most radical technological advancements and thus it is crucial to understand the interaction between human and machines. We incorporate this interaction and present a design that specifically studies the impact on candidates and evaluators when AI assessment takes place. Thus, by focusing on the human-AI interaction, our study investigates the use of AI tools more generally, and not just for a particular AI tool. This is because the only difference between the treatment and control in our supply-side experiments is the information about the AI evaluation and not the particular working of the specific AI tool. In particular, we did not provide evaluators with information about the (un)biasedness of the AI tool used. As such we believe that applicants and evaluators made decisions using their preexisting beliefs about AI tools more generally.

This study provides insights on the possible removal of barriers in recruitment women face when entering male-dominated tech jobs when AI tools are introduced. It is likely that the introduction of AI tools will impact not just recruitment but other environments where barriers exist. For example, it is conceivable that AI tools will assist in the early identification of talent and thus perhaps encourage women to obtain a tech degree. It is also conceivable that AI tools will assist employers in the assessment of hired employees and this may improve women's chances for career advancement. Further research is needed to measure these potential changes.

References

Acikgoz, Y., Davison, K.H., Compagnone, M., & Laske, M. Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment*, 28(4): 399-416. (2020).

Adamovic, M, Cooney-O'Donoghue, D., Avery, M., Leibbrandt, A., & Watson-Lynn, E. "Artificial Intelligence in Recruitment: Friend or Foe for Diversity & Inclusion?" (2022).

Agan, A. Y., Davenport, D., Ludwig, J., & Mullainathan, S. "Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias." *National Bureau of Economic Research (No. w30981).* (2023)

Agan, A. Y., & Starr, S. Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics*, 133(1), 191-235. (2018).

Aguirre, J., Matta, J., & Montoya, A.M. "Joining the Old Boys' Club: Women's Returns to Majoring in Technology and Engineering." Mimeo. (2022)

Altonji, J. G., & Blank, R.M. Race and gender in the labor market. *Handbook of Labor Economics*, 3: 3143-3259. (1999).

Anderson, D. M., & Haupert, M.J. "Employment and statistical discrimination: A hands-on experiment." (1999).

Arrow, K. J. "The Theory of Discrimination." In Discrimination in Labor Markets, edited by Orley Ashenfelter and Albert Rees. Princeton University Press, 3–33.(1973).

Bai, B., Dai, H., Zhang, D. J., Zhang, F., & Hu, H. The impacts of algorithmic work assignment on fairness perceptions and productivity: Evidence from field experiments. *Manufacturing & Service Operations Management*, 24(6), 3060-3078. (2022).

Bagues, M., Sylos-Labini, M., & Zinovyeva, N. Does the gender composition of scientific committees' matter? *American Economic Review*, 107(4), 1207-38. (2017).

Bagues, M. F., & Esteve-Volart, B. Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment. *The Review of Economic Studies*, 77(4), 1301-1328. (2010).

Banerjee, R., Ibanez, M., Riener, G., & Sahoo, S. (2021). "Affirmative action and application strategies: Evidence from field experiments in Columbia." *DICE Discussion Paper* (No. 362). (2021).

Bao, Z. & Huang, D. "Can Artificial Intelligence Improve Gender Equality? Evidence from a Natural Experiment" Available at SSRN: https://ssrn.com/abstract=4202239. (2022).

Bartoš, V., Bauer, M., Chytilová, J., & Matějka, F. (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, *106*(6), 1437-75.

Beasley, M.A. & Fischer, M.J. Why they leave: The impact of stereotype threat on the attrition of women and minorities from science, math and engineering majors. *Social Psychology of Education* 15.4: 427-448. (2012)

Becker, G. S. The Economics of Discrimination. Chicago: The University of Chicago Press. (1957).

Benson, A., Board, S., & Meyer-ter-Vehn, M. "Discrimination in Hiring: Evidence from Retail Sales". Available at SSRN: https://ssrn.com/abstract=4179847 (2022).

Bertrand, M, & Duflo, E. Field experiments on discrimination. *Handbook of Economic Field Experiments* 1: 309-393. (2017).

Bloodhart, B., Balgopal, M.M., Casper, A.M.A., McMeeking, L.B.S., & Fischer, E.V. Outperforming yet undervalued: Undergraduate women in STEM. *Plos one* 15, no. 6: e0234685. (2020).

Bohren, A. J., Hull, P., & Imas, A. "Systemic Discrimination: Theory and Measurement". *Mimeo.* (2023).

Bohren, J. A., Imas, A. & Rosenberg, M. The Dynamics of Discrimination: Theory and Evidence. *American Economic Review*. 109:10. (2019).

Bohren, J. A., Haggag, K., Imas, A., & Pope, D.G. Inaccurate Statistical Discrimination: An Identification Problem. *Review of Economics and Statistics*. (2022).

Bureau of Labor Statistics, U.S. Department of Labor, *Occupational Outlook Handbook*, Web Developers and Digital Designers, at https://www.bls.gov/ooh/computer-and-information-technology/web-developers.htm (visited *March 03, 2023*). (2022a).

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239

Bureau of Labor Statistics, U.S. Department of Labor, *Occupational Outlook Handbook*, Computer and Information Technology Occupations, at https://www.bls.gov/ooh/computer-and-information-technology/home.htm (visited *March 03, 2023*). (2022b).

of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33.2: 220-239. (2020)

Cain, G.G. The economic analysis of labor market discrimination: A survey. *Handbook of Labor Economics* 1: 693-785. (1986).

Campos-Mercade, P, & Mengel, F. Non-Bayesian statistical discrimination. *Management Science* (2023).

Czibor, E., Jimenez-Gomez, D., & List, J.A. "The Dozen Things Experimental Economists Should Do (More of)." *National Bureau of Economic Research*. No 25451. (2019).

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. Productivity and selection of human capital with machine learning. *American Economic Review* 106, no. 5: 124-27. (2016).

Chugunova, M., & Sele, D. "We and it: An interdisciplinary review of the experimental evidence on human-machine interaction." *Max Planck Institute for Innovation & Competition Research Paper* 20-15 (2020).

Cohen, T. How to leverage artificial intelligence to meet your diversity goals. *Strategic HR Review*. (2019).

Cook, A., Ingersoll, A. R., & Glass, C. Gender gaps at the top: Does board composition affect executive compensation? *Human Relations*, 72(8), 1292-1314. (2019).

Cowgill, B. "Bias and productivity in humans and algorithms: Theory and evidence from resume screening." *Columbia Business School, Columbia University* 29 (2018a).

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

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, M., Hakimov, R., & Kübler, D. "Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence." (2022).

Dastin, J. “Amazon scraps secret AI recruiting tool that showed bias against women.” *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. (2018).

Delfino, A. “Breaking gender barriers: Experimental evidence on men in pink-collar jobs.” *IZA Discussion Paper* 14083. (2021).

Deschamps, P. “Gender Quotas in Hiring Committees: a Boon or a Bane for Women?” *Sciences Po* (No. 82). (2018).

Dianat, A., Echenique, F., & Yariv, L. Statistical discrimination and affirmative action in the lab. *Games and Economic Behavior* 132: 41-58. (2022).

Dietvorst, B. J., Simmons, J.P., & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144.1: 114. (2015).

Dietvorst, B. J., Simmons, J.P., & Massey, C. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64.3: 1155-1170. (2018).

Doleac, J. L., & Hansen, B. The unintended consequences of “ban the box”: Statistical discrimination and employment outcomes when criminal histories are hidden. *Journal of Labor Economics* 38.2: 321-374. (2020).

Domínguez, J.J. "The Effectiveness of Committee Quotas: The Role of Group Dynamics." *Available at SSRN 4188778* (2021).

Dua, A., Ellingrud, K., Hancock, B., Luby, R., Madgavkar, A., & Pemberton, S. *Freelance, side hustles, and gigs: Many more Americans have become independent workers*. McKinsey & Company. Retrieved January 25, 2023, from https://www.mckinsey.com/featured-insights/sustainable-inclusive-growth/future-of-america/freelance-side-hustles-and-gigs-many-more-americans-have-become-independent-workers (2022).

Fehr, E., Gächter, S., & Kirchsteiger, G. Reciprocity as a contract enforcement device: Experimental evidence. *Econometrica.* 833-860. (1997).

Feld, J., Ip, E., Leibbrandt, A., & Vecci, J. "Identifying and overcoming gender barriers in tech: A field experiment on inaccurate statistical discrimination." (2022).

Feld, J., Salamanca, N., & Hamermesh, D.S. "Endophilia or exophobia: Beyond discrimination." *The Economic Journal* 126.594: 1503-1527. (2016).

Fernandez-Mateo, I & Fernandez, R. Bending the Pipeline? Executive Search and Gender Inequality in Hiring for Top Management Jobs. *Management Science*. 62. 10.1287/mnsc.2015.2315. (2016).

Flory, J. A., Leibbrandt, A., & List, J. A. Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, *82*(1), 122-155. (2015)

Flory, J. A., Leibbrandt, A., Rott, C., & Stoddard, O. Increasing Workplace Diversity Evidence from a Recruiting Experiment at a Fortune 500 Company. *Journal of Human Resources*, *56*(1), 73-92. (2021)

Flory, J., Leibbrandt, A., Rott, C., & Stoddard, O.. Leadership Signals and “Growth Mindset”: A Natural Field Experiment in Attracting Minorities to High-Profile Positions. *Management Science*. (forthcoming)

Fouad, N. A., & Santana, M.C. SCCT and underrepresented populations in STEM fields: Moving the needle. *Journal of Career Assessment* 25.1: 24-39. (2017).

Fry, R, Kennedy, B., & Funk, C. STEM jobs see uneven progress in increasing gender, racial and ethnic diversity. *Pew Research Center Science & Society* (2021).

Fryer, R.G., Goeree, J.K., & Holt, C.A. Experience-based discrimination: Classroom games. *The Journal of Economic Education* 36.2: 160-170. (2005).

Gächter, S., & Fehr, E.. Reciprocity and contract enforcement. *Handbook of Experimental Economics Results* 1: 319-324. (2008).

Galer, S. "How to stop HIRING Bias: Don't Let AI take Over HR". *Forbes*. https://www.forbes.com/sites/sap/2019/11/12/how-to-stop-hiring-bias-dont-let-ai-take-over-hr/?sh=4f0be9b2e0fb. (2019, November 12).

Gee, K. "In Unilever's radical hiring experiment, resumes are out, algorithms are in." *Wall Street Journal* 26 (2017).

Gee, L.K. The more you know: Information effects on job application rates in a large field experiment. *Management Science* 65.5: 2077-2094. (2019).

Giusta, M.D., & Bosworth, S. Bias and discrimination: what do we know?. *Oxford Review of Economic Policy* 36.4: 925-943. (2020).

Glaeser, E.L., Hillisiii, A., Kimiv, H., Kominersv, S.D., & Lucavi, M. "How Does Compliance Affect the Returns to Algorithms? Evidence from Boston's Restaurant Inspectors." (2019).

Glover, D, Pallais, A., & Pariente, W. Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics* 132.3: 1219-1260. (2017).

Goldin, C., & Rouse, C. Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review* 90.4: 715-741. (2000).

Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., & Ur, B.. "An empirical study on the perceived fairness of realistic, imperfect machine learning models." In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 392-402. (2020).

Haegele I. "The Broken Rung: Gender and the Leadership Gap." Mimeo. (2023)

Heilmann, C. "Council post: Artificial Intelligence and recruiting: A candidate's perspective." *Forbes*. Retrieved February 7, 2023, from https://www.forbes.com/sites/forbescoachescouncil/2018/06/22/artificial-intelligence-and-recruiting-a-candidates-perspective/?sh=7fd337807a88. (2018).

Holmes, A. "Ai could be the key to ending discrimination in hiring, but experts warn it can be just as biased as humans." *Business Insider*. Available https://www.businessinsider.com/ai-hiring-tools-biased-as-humans-experts-warn-2019-10. (2019, October 8).

Houser, K. A. Can AI Solve the Diversity Problem in the Tech Industry: Mitigating Noise and Bias in Employment Decision-Making. *Stan. Tech. L. Rev.*, *22*, 290. (2019).

Hsieh, C., Hurst, E., Jones, C.I., & Klenow, P.J. The allocation of talent and us economic growth. *Econometrica* 87, no. 5: 1439-1474. (2019).

Jussupow, E., Benbasat, I., & Heinzl, A. "Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion." (2020).

Kessler, J. B., Low, C., & Shan, X. "Lowering the playing field: Discrimination through sequential spillover effects." mimeo. (2022).

Köchling, A. & Wehner, M.C. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* 13.3: 795-848. (2020).

Kordzadeh, N & Ghasemaghaei, M. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems* 31.3: 388-409. (2022).

Kulp, T. “Ai and hiring bias: Why you need to teach your robots well.” *HRExecutive.com*. https://hrexecutive.com/ai-and-hiring-bias-why-you-need-to-teach-your-robots-well/.(2021).

Lee, M.K. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5.1: 2053951718756684. (2018).

Lee, M.K., & Rich, K. "Who Is Included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust." *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. (2021).

Lee, M.K., & Baykal, S. "Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division." *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. (2017).

Leibbrandt, A., & List, J.A. “When Equal Employment Opportunity Statements Backfire: Field Experimental Evidence on Job-Entry Decisions.” *NBER Working paper #25035* (2018).

Lengnick-Hall, M. L., Neely, A. R., & Stone, C. B. Human resource management in the digital age: Big data, HR analytics and artificial intelligence. In *Management and technological challenges in the digital age* (pp. 1-30). CRC Press. (2018)

Li, D., Raymond, L.R., &, Bergman, P. “Hiring as exploration.” *National Bureau of Economic Research* No. w27736. (2020).

LinkedIn. Report Highlights Top Global Trends in Recruiting. Available https://news.linkedin.com/2018/1/global-recruiting-trends-2018. (2018)

List, J A. The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics* 119.1: 49-89. (2004).

Logg, J.M., Minson, J.A., & Moore, D.A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151: 90-103. (2019).

Makarova, E., Aeschlimann, B., & Herzog, W. Why is the pipeline leaking? Experiences of young women in STEM vocational education and training and their adjustment strategies. *Empirical Research in Vocational Education and Training* 8.1: 1-18. (2016).

Malik, A., Srikanth, N. R., & Budhwar, P. Digitisation, artificial intelligence (AI) and HRM. *Human Resource Management: Strategic and International Perspectives*, 88. (2020).

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. "Implications of AI (un-) fairness in higher education admissions: the effects of perceived AI (un-) fairness on exit, voice and organizational reputation." In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 122-130. (2020).

Meister, J. "Ten HR trends in the age of artificial intelligence." *Forbes,* available at: www. forbes. com/sites/jeannemeister/2019/01/08/ten-hr-trends-in-the-age-of-artificial-intelligence (2019).

Metz, C. “Who is making sure the A.I. machines AREN'T RACIST?” *The New York Times*. https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html. (2021).

Miller, C. The Persistent Effect of Temporary Affirmative Action. *American Economic Journal: Applied Economics* 9 (3):152–90. (2017).

Miller, T. "Explainable artificial intelligence: What were you thinking?." *Artificial Intelligence: For Better or Worse.* 19-38. (2019)

Mirowska, A., & Mesnet, L. Preferring the devil you know: Potential applicant reactions to artificial intelligence evaluation of interviews. *Human Resource Management Journal*. (2021).

Mocanu, T. Designing Gender Equity: Evidence from Hiring Practices and Committees. Mimeo. Available https://tatianamocanu.github.io/jm/mocanu_jmp_hiring.pdf (2023).

Murciano-Goroff, R. Missing women in tech: The labor market for highly skilled software engineers. *Management Science* 68.5: 3262-3281. (2022).
Neumark, D. Experimental Research on Labor Market Discrimination. *Journal of Economic Literature*, 56(3): 799-866. (2018).
Newman, D.T., Fast, N.J., & Harmon, D.J. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160: 149-167. (2020).
Niederle, M, Segal, C., & Vesterlund, L. How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness. Management Science 59 (1):1–16. (2013).
Opitz, S., Sliwka, D., Vogelsang, T. & Zimmermann, T. "The Targeted Assignment of Incentive Schemes." Available at SSRN: https://ssrn.com/abstract=4077778, (2022).
Paola, M., & Scoppa, V. Gender Discrimination and Evaluators' Gender: Evidence from Italian Academia. Economica, 82(325), 162–188. (2015).
Patty, J.W., & Penn, M.G. "Algorithmic fairness and statistical discrimination." *arXiv preprint arXiv:2208.08341* (2022).
Phelps, E.S. The statistical theory of racism and sexism. *The American Economic Review* 62.4: 659-661. (1972).
Roberson, L. & Kulik, C.T. "Stereotype threat at work." *Academy of Management Perspectives* 21.2: 24-40. (2007).
Rodgers, W.M. Handbook on the Economics of Discrimination. Edward Elgar Publishing, (2009).
Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review*, *107*(5), 141-45.
Sassler, S., Glass, J., Levitte, Y., & Michelmore, K.M. The missing women in STEM? Assessing gender differentials in the factors associated with transition to first jobs. *Social science research* 63: 192-208. (2017).
Serra-Garcia, M. & Gneezy, U. "Improving Human Deception Detection Using Algorithmic Feedback." (2023).
Sharkey, N. The impact of gender and race bias in AI. *Humanitarian Law and Policy* (2018).
Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, *61*(4), 66-83. (2019).
SHRM. "SHRM Survey Findings: The Importance of Social Media for Recruiters and Job Seekers." *HR Today Trends and Forecasting*. https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/SHRM-Ascendo-Resources-Social-Media-Recruitment.pdf (2015).
Spencer, S.J., Logel, C., & Davies, P.G. Stereotype threat. *Annual Review of Psychology* 67, no. 1: 415-437. (2016).
Stahl, A. "7 fast-growing industries for Freelancers." *Forbes*. Retrieved January 25, 2023, from https://www.forbes.com/sites/ashleystahl/2021/01/07/7-fast-growing-industries-for-freelancers/?sh=3292c0495fef (2021).
Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, *9*(2), 20539517221115189. (2022).
Stevenson, M.T., & Doleac, J.L. "Algorithmic risk assessment in the hands of humans." *Available at SSRN 3489440* (2022).

Tambe, P., Cappelli, P., & Yakubovich, V. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *61*(4), 15-42. (2019).

Upadhyay, A. K., & Khandelwal, K. Applying artificial intelligence: implications for recruitment. *Strategic HR Review*. (2018).

US Bureau of Labor Statistics. "Women in the labor force: A databook." (2021).

Van Veelen, R., Derks, B., & Endedijk, M.D. Double trouble: How being outnumbered and negatively stereotyped threatens career outcomes of women in STEM. *Frontiers in psychology* 10: 150. (2019).

Vassilopoulou, J., Kyriakidou, O., Özbilgin, M. F., & Groutsis, D. Scientism as illusion in HR algorithms: Towards a framework for algorithmic hygiene for bias proofing. *Human Resource Management Journal*. (2022).

Von Krogh, G. Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries* (2018).

Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., & Trichina, E. Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. *The International Journal of Human Resource Management*, *33*(6), 1237-1266. (2022).

Wachter-Boettcher, S. "AI recruiting tools do not eliminate bias." *Time*. https://time.com/4993431/ai-recruiting-tools-do-not-eliminate-bias/. (2017).

Wall Street Journal. "Employers, investors take notice of AI tools to speed job recruitment." Retrieved February 7, 2023, from https://www.wsj.com/articles/employers-investors-take-notice-of-ai-tools-to-speed-job-recruitment-11641599629 (2022)

Wang, A. J. "Procedural justice and risk-assessment algorithms." *Available at SSRN 3170136* (2018).

Xue, Y, & Larson, R.C. STEM crisis or STEM surplus? Yes and yes. *Monthly Labor Review.* (2015).

Yarger, L., Payton, F.C., & Neupane, B. Algorithmic equity in the hiring of underrepresented IT job candidates. *Online information review* 44.2: 383-395. (2019).

Zielinski, D. "Addressing artificial intelligence-based hiring concerns." *SHRM*. https://www.shrm.org/hr-today/news/hr-magazine/summer2020/pages/artificial-intelligence-based-hiring-concerns.aspx. (2020).

Zhang, L. & Yencha, C. Examining perceptions towards hiring algorithms. *Technology in Society* 68: 101848. (2022).

Appendix A. Supplemental Figures and Tables

Figure A.1: Example of Test Interface

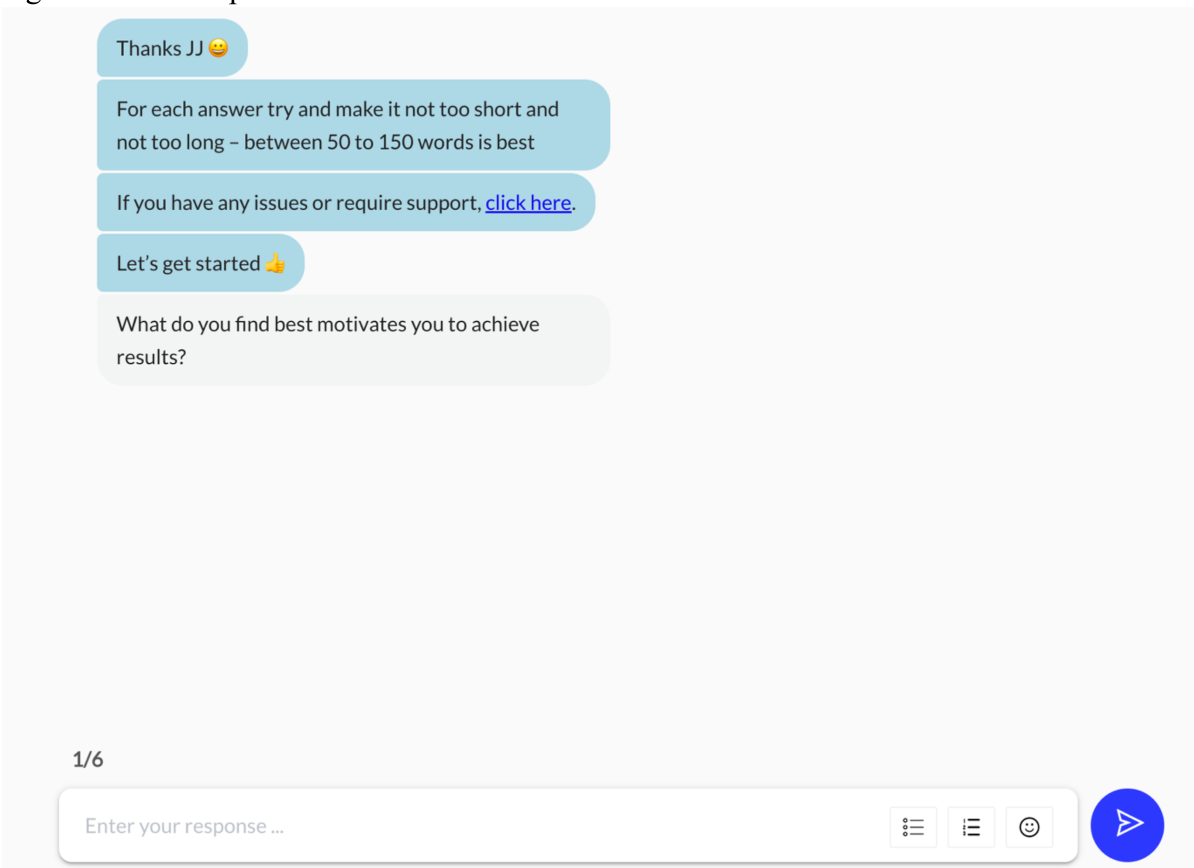
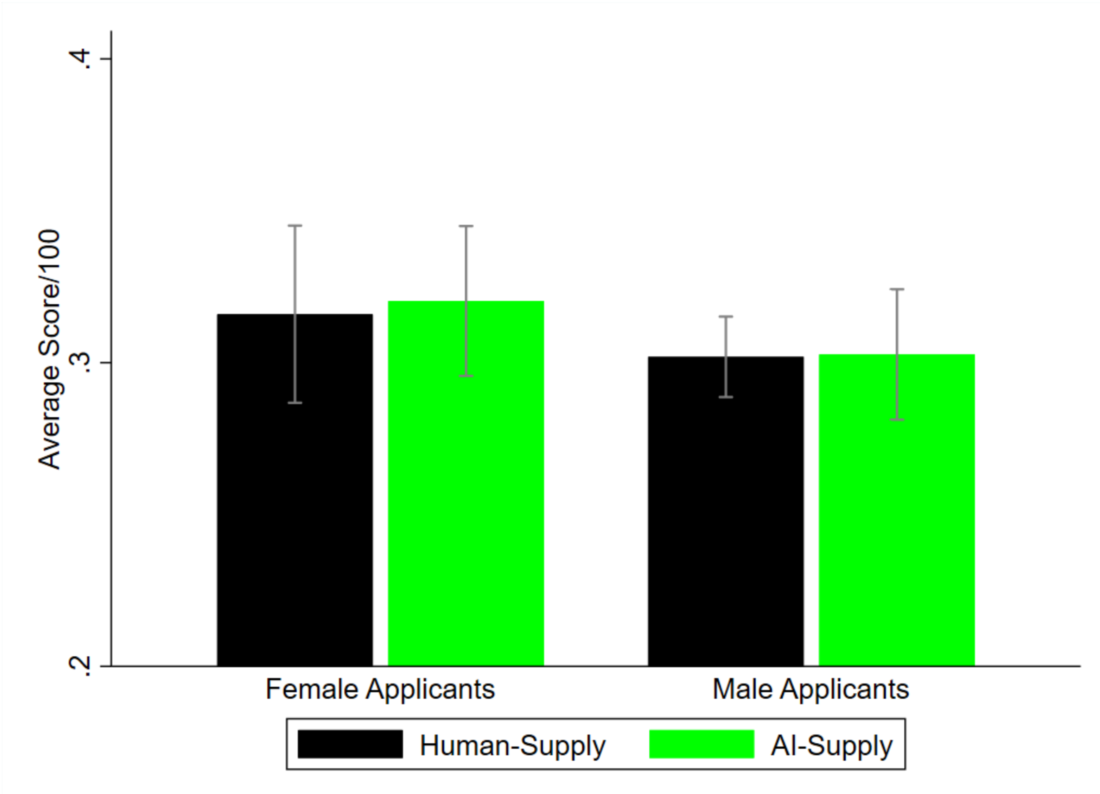
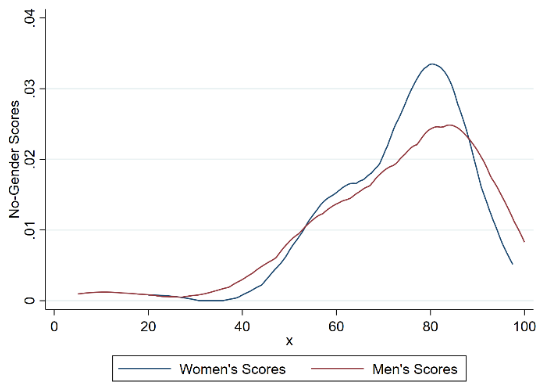


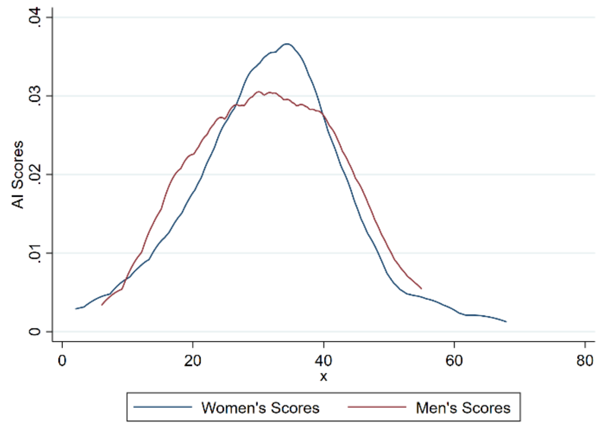
Figure A.2: AI-Generated Score by Gender and Treatment



Notes: The figure represents the average assessment score generated by the AI. The left two columns illustrate the behaviour of female applicants and the next two columns represent behaviour of male applicants. Confidence intervals on each bar illustrate significance at the 10% level.

Figure A.3: Distributions of Evaluations for Male and Female Candidates, for the gender blind treatment and the AI algorithm.





Notes: This figure illustrates the distribution of women and mens scores. The left panel shows the distribution for the no gender treatment and the right figure illustrates the pure AI score.

Figure A.4 Actual and Estimated Fraction Male Across Evaluation Quantiles, by Evaluation Type

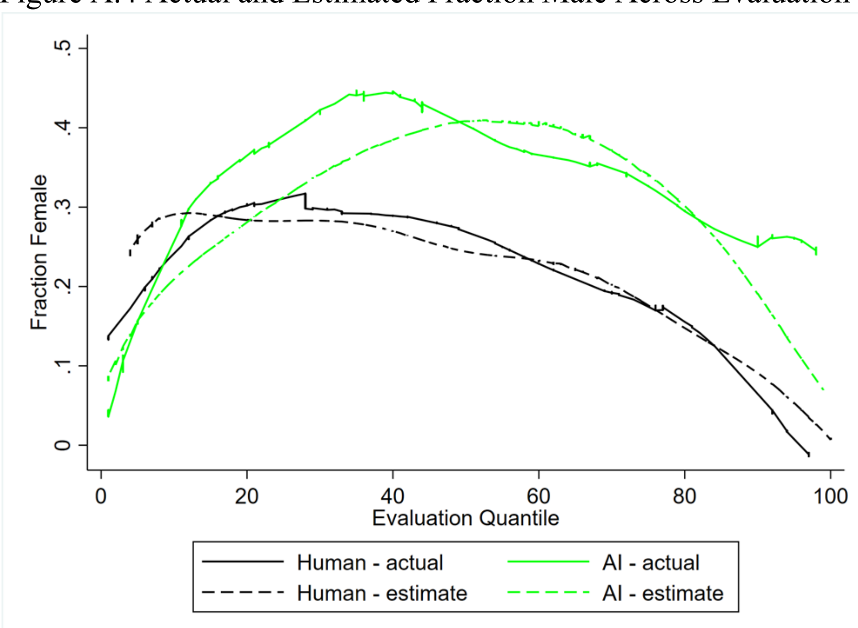


Table A.1: Experiment 1 Summary Statistics

	(1) N	(2) mean	(3) Sd	(4) min	(5) max
Male	723	0.761	0.427	0	1
currently studying	723	0.182	0.385	0	1
currently employed	723	0.555	0.497	0	1
Education					
Less than High school	723	0.069	0.083	0	1
High school	723	0.057	0.231	0	1
Some college	723	0.188	0.391	0	1
2-year college	723	0.112	0.316	0	1
4-year college	723	0.512	0.500	0	1
Postgrad	723	0.124	0.330	0	1
Years web development experience	723	3.793	4.378	0	45

Notes: This Table reports the summary statistics for the sample of applicants in Experiment 1.

Table A.2: Regressions of Completed Applicant Characteristics by Gender and Treatment

Models	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Web Design Training								Experience with Programming Language					
	University	Non-University Class	Self-Taught	Years of Web Design Experience	4 Year Degree Holder	Java	HTML	CSS	Python	PHP	C#	React	JavaScript	Angular
AI Evaluation	0.091	-0.026	0.082	0.589	0.066	-0.004	0.006	0.010	-0.033	0.066	0.015	-0.053	0.007	0.019
	(0.064)	(0.063)	(0.057)	(0.580)	(0.064)	(0.062)	(0.021)	(0.021)	(0.065)	(0.062)	(0.053)	(0.062)	(0.025)	(0.048)
Female Applicant	0.140**	-0.076	-0.096	-0.225	-0.014	0.146**	-0.032	-0.012	-0.087	-0.081	-0.064	-0.187***	-0.097**	-0.058
	(0.070)	(0.067)	(0.069)	(0.509)	(0.071)	(0.071)	(0.033)	(0.029)	(0.070)	(0.060)	(0.051)	(0.070)	(0.046)	(0.044)
AI Evaluation × Female Applicant	-0.230*	0.110	-0.190	-0.851	-0.123	-0.218*	0.027	-0.111	0.023	-0.079	-0.009	0.053	0.046	-0.083
	(0.124)	(0.121)	(0.120)	(0.889)	(0.124)	(0.118)	(0.047)	(0.070)	(0.123)	(0.106)	(0.092)	(0.123)	(0.070)	(0.0671)
Constant	0.470***	0.404***	0.674***	3.311***	0.483***	0.370***	0.970***	0.965***	0.509***	0.300***	0.204***	0.687***	0.957***	0.152***
	(0.033)	(0.033)	(0.031)	(0.255)	(0.033)	(0.032)	(0.011)	(0.012)	(0.033)	(0.030)	(0.027)	(0.031)	(0.014)	(0.0238)
N	410	410	410	410	410	410	410	410	410	410	410	410	410	410

Notes: We use an OLS to estimate the models. AI Evaluation is a dummy indicating that the candidate was in the AI-Supply treatment. Female Applicant is a dummy indicating the candidate being considered is female. The dependent variable an indicator for whether the candidate has the characteristic in the column header. The sample includes only those candidates who completed the application. Data are from the experiment 1. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A.3 Regressions of Non-Completers' Characteristics by Gender and Treatment

Models	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Web Design Training							Experience with Programming Language						
	University	Non-University Class	Self-Taught	Years of Web Design Experience	4 Year Degree Holder	Java	HTML	CSS	Python	PHP	C#	React	JavaScript	Angular
AI Evaluation	0.093	-0.135**	0.022	0.511	0.063	-0.050	-0.020	-0.036	-0.017	-0.026	0.044	-0.102	0.015	-0.016
	(0.065)	(0.063)	(0.061)	(0.717)	(0.067)	(0.067)	(0.021)	(0.027)	(0.067)	(0.065)	(0.060)	(0.066)	(0.037)	(0.056)
Female Applicant	-0.080	-0.091	-0.172**	-1.006*	0.130*	-0.083	-0.037	-0.030	-0.150**	-0.121*	-0.062	-0.119	-0.007	-0.122**
	(0.077)	(0.073)	(0.075)	(0.576)	(0.075)	(0.076)	(0.030)	(0.030)	(0.075)	(0.071)	(0.061)	(0.076)	(0.046)	(0.054)
AI Evaluation × Female Applicant	0.091	0.135	-0.006	-0.428	-0.230	0.117	0.003	0.019	0.117	0.076	0.106	-0.014	-0.048	0.033
	(0.153)	(0.151)	(0.157)	(1.145)	(0.159)	(0.159)	(0.0738)	(0.0757)	(0.159)	(0.151)	(0.145)	(0.157)	(0.103)	(0.113)
Constant	0.563***	0.424***	0.689***	4.456***	0.503***	0.483***	0.987***	0.980***	0.517***	0.404***	0.245***	0.636***	0.907***	0.238***
	(0.041)	(0.041)	(0.038)	(0.406)	(0.041)	(0.041)	(0.009)	(0.011)	(0.041)	(0.040)	(0.035)	(0.039)	(0.024)	(0.035)
N	316	316	316	316	316	316	316	316	316	316	316	316	316	316

Notes: We use an OLS to estimate the models. AI Evaluation is a dummy indicating that the candidate was in the AI-Supply treatment. Female Applicant is a dummy indicating the candidate being considered is female. The dependent variable an indicator for whether the candidate has the characteristic in the column header. The sample includes only those candidates who did not complete the application. Data are from the experiment 1. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A.4: Application Results, Application Completion by Gender

Models	(1) Application Completion	(2) Application Completion	(3) Application Completion	(4) Application Completion	(5) Application Completion	(6) Application Completion	(7) Application Completion	(8) Application Completion
AI Supply	-0.127***	-0.121***	-0.118***	-0.124***	-0.120***	-0.125***	-0.131***	-0.117**
	(0.0457)	(0.046)	(0.0456)	(0.0458)	(0.0457)	(0.0454)	(0.0463)	(0.046)
Female Applicant	-0.088*	-0.084	-0.094*	-0.085*	-0.090*	-0.091*	-0.084	-0.090*
	(0.0515)	(0.0523)	(0.0515)	(0.0514)	(0.0528)	(0.0511)	(0.0521)	(0.0532)
AI Supply × Female Applicant	0.305***	0.297***	0.293***	0.296***	0.288***	0.298***	0.302***	0.265***
	(0.0921)	(0.0926)	(0.0918)	(0.0922)	(0.0928)	(0.0923)	(0.0931)	(0.0956)
Controls								
Web Design Training	N	Y	N	N	N	N	N	Y
Years of Experience	N	N	Y	N	N	N	N	Y
4 Year Degree Holder	N	N	N	Y	N	N	N	Y
Programming Languages Known	N	N	N	N	Y	N	N	Y
Time to Receive Interview	N	N	N	N	N	Y	N	Y
Race of Applicant	N	N	N	N	N	N	Y	Y
Constant	0.604***	0.630***	0.650***	0.626***	0.603***	0.670***	0.626***	0.780***
	(0.0251)	(0.0449)	(0.0295)	(0.0308)	(0.11)	(0.0343)	(0.0617)	(0.136)
N	726	726	726	726	726	726	726	726

Notes: We use an OLS to estimate the models. The first column reports estimate without controls and controls are added in the second column. The dependent variable is an indicator variable whether the applicant completed the interview assessment. The variable AI Supply is equal to one if the applicant was randomly assigned to the *AI Supply Treatment*. Data are from the experiment 1. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A.5: Survey Results

	Women			Men			
Sample reporting…	Human	AI	Diff (Human-AI)	Human	AI	Diff (Human-AI)	N
Survey Sample - High Status	0.59	0.69	-0.10	0.57	0.53	0.04	129
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.04)	
Survey Sample - High Value	0.93	0.92	0.02	0.83	0.79	0.04	129
	(0.03)	(0.04)	(0.05)	(0.05)	(0.05)	(0.06)	
Survey Sample - Wage	32.11	30.05	2.05	32.00	32.16	-0.16	114
	(2.52)	(2.32)	(1.74)	(2.26)	(2.23)	(1.51)	
Experimental Sample - High Status	0.52	0.38	0.13	0.37	0.34	0.03	410
	(0.06)	(0.08)	(0.11)	(0.03)	(0.05)	(0.06)	
Experimental Sample - High Value	0.81	0.68	0.14	0.75	0.52	0.22***	410
	(0.05)	(0.08)	(0.09)	(0.02)	(0.06)	0.06	

Notes: We present average and standard errors for how women and men report in either the Experimental or General Survey what they think about human and AI evaluation, as well as the difference with t-tests. In the General Survey, reporting "High Status" is saying a job recruited with a particular type of evaluation would be either "high status" or "very high status" rather than "low status", "very low status", or "neutral" .In the General Survey, reporting "High Value" is saying a job recruited with a particular type of evaluation would be either "important" or "very important" rather than "not important" or "neutral". In the General Survey, wage is the hourly wage anticipated for the job with that evaluation method, with the top and bottom 5% truncated (resulting in a range of values from $10 to $100 per hour). In the Experimental Survey, reporting "High Status" is saying a job recruited with a particular type of evaluation would be either "high status" or "very high status" rather than "low status", "very low status", or "neutral" .In the Experimental Survey, reporting "High Value" is saying a job recruited with a particular type of evaluation would be either "high value" or "very high value" rather than "low value", "very low value", or "neutral". Data are from the experiment 1. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A.6: Balance between the sample used in Experiment 2 and the full sample of applicants

	(1) Full	(2) Demand Sample	(3) Diff
Currently studying	0.141	0.164	0.02
Currently employed	0.495	0.522	0.027
Education			
Less than High school	0.000	0.006	0.00
High school	0.111	0.075	0.036
Some college	0.212	0.208	0.004
2-year college	0.131	0.106	0.026
4-year college	0.434	0.491	0.057
Postgrad	0.111	0.113	0.002
Years web development experience	3.429	3.343	0.085
F-Stat		0.44 p-value=0.898	

Note: This Table reports the difference in characteristics between the full sample and the demand sample. The first column reports the characteristics for the full sample of applicants and column 2 reports the characteristics for the AD demand subsample. The full sample is restricted to those who complete the assessment.

Table A.7: Experiment 2 Summary Statistics

	(1) N	(2) mean	(3) sd	(4) min	(5) max
Male	507	0.665	0.472	0	1
Ethnicity: White or Caucasian	507	0.769	0.422	0	1
Ethnicity: Asian	507	0.095	0.293	0	1
Ethnicity: African American	507	0.94	0.293	0	1
Age	507	42.78	12.32	19	77
Currently employed	507	0.957	0.204	0	1
Education					
High school	507	0.055	0.228	0	1
Some college	507	0.122	0.328	0	1
2-year college	507	0.091	0.288	0	1
4-year college	507	0.432	0.496	0	1
Postgrad	507	0.300	0.459	0	1
Role in the Tech sector					
Managers	507	0.250	0.434	0	1
Senior Managers (Director, hiring manager etc.)	507	0.221	0.415	0	1
Software developer or engineer	507	0.162	0.369	0	1
Consultant or general tech (e.g., multiple roles)	507	0.219	0.414	0	1
Other	507	0.280	0.449	0	1
Responsible for hiring	507	0.838	0.369	0	1

Note: This table reports summary statistics for experiment 2.

Table A.8: Human Evaluators vs Artificial Intelligence

Models	(1)	(2)	(3)	(4)	(5)	(6)
		Human-Demand			AI-Demand	
Female Applicant	-2.897***	-2.531**	-3.025**	-1.065	0.182	-0.398
	(1.093)	(1.243)	(1.224)	(1.497)	(1.787)	(1.719)
Applicant Controls	N	Y	Y	N	Y	Y
AI Score Control	N	N	Y	N	N	Y
Constant	74.51***	61.52***	57.42***	56.80***	41.78***	28.24***
	(1.088)	(4.895)	(4.983)	(1.784)	(8.128)	(8.360)
N	805	805	805	591	591	591

Notes: We use an OLS to estimate the models with robust standard errors and evaluator fixed effects. The first and fourth columns reports estimate without controls; the second and fifth columns report estimates with applicant controls included; the third and sixth columns include an additional control for the AI-produced score. Applicant controls include indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of experience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), time between providing initial information and receiving the email with the interview invitation, and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander). The dependent variable is the score given by evaluators in the Human-Demand (columns 1-3) and AI-Demand (columns 4-6) treatments. N is the number of observations in each treatment. As discussed in Section 5.1, the number of observations differ between the two treatments. Data are from the experiment 2. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A.9: Regressions by Quantile

Models	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	50th Quantile				75th Quantile				90th Quantile			
	Human-Demand		AI-Demand		Human-Demand		AI-Demand		Human-Demand		AI-Demand	
Female Applicant	0.195	-0.587	0.879	0.399	-2.662**	-3.225***	1.581	2.223	-3.307***	-3.020**	5.394**	4.618*
	(1.272)	(1.392)	(2.904)	(2.668)	(1.080)	(1.049)	(2.997)	(3.074)	(1.239)	(1.479)	(2.297)	(2.360)
Applicant Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
AI Score	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Constant	58.92***	52.61***	50.30***	29.21***	75.62***	74.59***	65.42***	59.13***	79.55***	76.99***	77.32***	78.36***
	(5.859)	(5.838)	(9.519)	(7.842)	(3.429)	(4.686)	(11.77)	(14.58)	(3.039)	(4.811)	(18.89)	(16.50)
N	805	805	591	591	805	805	591	591	805	805	591	591

Notes: We use quantile regression to estimate the models with robust standard errors. The odd columns report estimates with applicant controls included; the even columns include an additional control for the AI-produced score. Applicant controls include indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of experience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), time between providing initial information and receiving the email with the interview invitation, and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander). The dependent variable is the score given by evaluators in the *Human-Demand* (columns 1-2, 5-6, 9-10) and AI-Demand (columns 3-4, 7-8, 11-12) treatments. Columns 1-4 present quantile regressions at the median, 5-8 at the 75th percentile, and 9-12 at the 90th percentile. Data are from the experiment 2. Significance levels are *** p<0.01, ** p<0.05, * p<0.1

Table A.10: Regressions of in the No-Name Treatment and Full Sample

Models	(1)	(2)	(3)	(4)	(5)	(6)
	No-Name Treatment			Evaluation		
Female Applicant	-0.177	2.373	2.077			
	(1.466)	(1.916)	(1.875)			
Female Applicant				-0.177	1.989	1.206
				(1.463)	(1.602)	(1.547)
Human-Demand				0.863	1.377	1.138
				(1.848)	(1.801)	(1.779)
AI-Demand				-16.84***	-16.26***	-16.42***
				(2.324)	(2.280)	(2.258)
Female Applicant × Human-Demand				-2.720	-4.093**	-3.879**
				(1.826)	(1.868)	(1.808)
Female Applicant × AI-Demand				-0.888	-2.241	-1.792
				(2.091)	(2.064)	(1.959)
Gender Gap - Human-Demand vs. AI-Demand				p=.32	p=.30	p=.22
Applicant Controls	N	Y	Y	N	Y	Y
AI Score	N	N	Y	N	N	Y
Constant	73.65***	62.67***	58.87***	73.65***	60.12***	53.87***
	(1.497)	(8.414)	(8.547)	(1.494)	(4.219)	(4.254)

Notes: We use an OLS to estimate the models with robust standard errors clustered at the evaluator level. The first and fourth columns reports estimate without controls; the second and fifth columns report estimates with applicant controls included; the third and sixth columns include an additional control for the AI-produced score. Applicant controls include indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of experience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), time between providing initial information and receiving the email with the interview invitation, and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander). The dependent variable is the score given by evaluators in the No-Name treatment (columns 1-3) and in all treatments (columns 4-6). Gender Gap - *Human-Demand* vs. *AI-Demand* presents the result of the test of equivalence between the Female Applicant X *Human-Demand* and Female Applicant X *AI-Demand* coefficients. Data are from the experiment 2. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A.11: Results in the No-Name Treatment and Full Sample by Quantile

Models	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	No-Name Treatment						Evaluation					
	50th Quantile		75th Quantile		90th Quantile		50th Quantile		75th Quantile		90th Quantile	
Female Applicant	0.980	1.803	-0.678	-0.436	-0.559	-0.377						
	(1.952)	(1.785)	(1.349)	(1.038)	(1.201)	(1.485)						
Female Applicant							0.921	-0.144	-0.251	-0.620	0.497	-0.0403
							(1.848)	(1.710)	(1.214)	(0.847)	(1.217)	(1.345)
Human-Demand							-1.516	-2.700	-2.093**	-1.904**	-1.24e-14	0.00476
							(1.808)	(1.661)	(0.902)	(0.808)	(1.226)	(1.131)
AI-Demand							-22.57***	-22.94***	-12.25***	-12.97***	-6.381***	-6.492***
							(1.949)	(1.938)	(1.540)	(1.204)	(2.323)	(1.931)
Female Applicant × Human-Demand							-1.122 (2.240)	-0.131 (2.088)	-1.907 (1.516)	-2.550** (1.201)	-3.347* (1.761)	-3.595* (1.932)
Female Applicant × AI-Demand							0.280 (3.153)	0.590 (3.249)	-1.076 (2.671)	-0.241 (2.681)	3.259 (3.234)	3.027 (2.845)
Gender Gap - Human-Demand vs. AI-Demand							p=.65	p=.81	p=.75	p=.39	p=.05**	p=.02**
Applicant Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
AI Score	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Constant	74.10***	70.97***	85.68***	81.99***	89.57***	86.30***	68.27***	63.02***	78.06***	74.47***	87.24***	83.53***
	(6.945)	(8.121)	(4.665)	(5.588)	(6.815)	(5.231)	(3.733)	(4.253)	(2.369)	(2.327)	(3.822)	(3.819)

N	621	621	621	621	621	621	2017	2017	2017	2017	2017	2017

Notes: We use quantile regression to estimate the models with robust standard errors. The odd columns report estimates with applicant controls included; the even columns include an additional control for the AI-produced score. Applicant controls include indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of experience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), time between providing initial information and receiving the email with the interview invitation, and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander). The dependent variable is the score given by evaluators in the No-Name treatment (columns 1-6) and all treatments (columns 7-12) treatments. Columns 1-2 and 7-8 present quantile regressions at the median, 3-4 and 9-10 at the 75th percentile, and 9-10 and 11-12 at the 90th percentile. Gender Gap - *Human-Demand* vs. AI-Demand presents the result of the test of equivalence between the Female Applicant X *Human-Demand* and Female Applicant X *AI-Demand* coefficients. Data are from the experiment 2. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Table A.12: Deviation from the AI Score

Models	(1)	(2)	(3)	(4)	(5)	(6)
	All Applicants		Female Applicants		Male Applicants	
	Dev. From AI	Abs. Dev. From AI	Dev. From AI	Abs. Dev. From AI	Dev. From AI	Abs. Dev. From AI
Biased	8.865***	6.547***	9.185***	6.733**	8.532***	6.354**
	(2.023)	(1.799)	(2.984)	(2.627)	(2.742)	(2.468)
Male Evaluator	2.946	2.318	5.327*	3.492	0.550	1.137
	(2.087)	(1.856)	(3.071)	(2.705)	(2.834)	(2.551)
Above Median Age	-0.146	-0.689	-2.447	-3.398	2.136	2.008
	(2.054)	(1.827)	(3.027)	(2.666)	(2.786)	(2.507)
Web Developer Hiring Experience	12.24***	12.07***	13.30***	13.02***	11.17***	11.11***
	(2.039)	(1.814)	(3.006)	(2.647)	(2.765)	(2.488)
Constant	14.36***	18.41***	12.89***	18.42***	15.86***	18.41***
	(2.080)	(1.851)	(3.052)	(2.688)	(2.835)	(2.551)
N	591	591	295	295	296	296

Notes: We use an OLS to estimate the models. The dependent variable in odd columns is the deviation of the *AI-Demand* score from the AI score. The dependent variable in even columns is the absolute value of this deviation. Biased is a binary variable where 1 is evaluators identified as believing men are better at web development than women, 0 otherwise. Male Evaluator is a binary variable where 1 is male evaluators, 0 otherwise. Above Median Age is a binary variable where 1 is evaluators born before the median birth year of 1981, 0 otherwise. Web Developer Hiring Experience is a binary variable where 1 is evaluators who do report having prior experience hiring web developers, 0 otherwise. The sample is restricted to evaluators in the *AI-Demand* treatment. The sample in columns 1-2 are all evaluations in the *AI-Demand* treatment. The sample in columns 3-4 are evaluations of female applicants in the *AI-Demand* treatment. The sample in columns 5-6 are evaluations of male applicants in the *AI-Demand* treatment. Data are from the experiment 2. Significance levels are *** p<0.01, ** p<0.05, * p<0.1.

Appendix B. Job Ad

Web Developer for leading international organization in the education sector

Job Information

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

Job Description

We are looking for a Web Developer to create a minimalist website that attracts organizations to us and enables the purchase of our innovative product.

In this role, you will have the opportunity to bring in your creativity, talent and drive. Develop and design a website that stands out and improves your portfolio.

Responsibilities

- Create and discuss wireframes to decide on layout
- Write code for applications
- Run functionality tests
- Develop software documentation
- Maximize webpage visibility
- Provide your feedback and thoughts on the projects.

You will know you are successful in this role if you

- Enjoy website design
- Are able to design a beautiful website front end
- Have solid knowledge in JavaScript, HTML & CSS
- Enjoy working independently

How to Apply

To apply, please complete the application form
https://monash.az1.qualtrics.com/jfe/form/SV_cUtfMbnXV1D608S
by 30th of October.

Disclaimer

By applying, you acknowledge that your information may be used for assessment purposes.

Appendix C: Survey of US Labor Force

In addition to our tech sample, we also elicited responses from 124 non-tech respondents. Similarly to our tech sample, we find that women are significantly more worried about bias from human evaluation than from AI (t-test, diff=0.19, p=.00) and are more worried about bias from human evaluation than men are (t-test, diff=0.11, p=.19). However, in this sample we also find that men are more worried about bias from humans than AI, deviating from our tech sample (t-test, diff=0.28, p=.00). This could be because some of the men in this sample will be in female-dominated industries where they may anticipate bias against themselves, though we do not have the power to show these results by industry. We also find in this non-tech sample that women anticipate that jobs using human recruiting, rather than AI recruiting, will be higher status (t-test, diff=0.10, p=.03), higher value (t-test, diff=0.12, p=.02), and higher paying (t-test, diff=2.22, p=.02). Men in the non-tech sample do not derive information about these aspects from the evaluation type.

Appendix D: Demand Side Conceptual Framework

Suppose there are two groups of workers, $g \in \{M, W\}$, where M is men and W is women. These workers can have one of two ability levels, $A \in \{A_H, A_L\}$, where H is high and L is low. Regardless of what the underlying probabilities of men and women being high ability, suppose evaluators' uninformed priors are that men are more likely to be high ability than women, i.e. $\hat{P}_M^U(A = A_H) - B = \hat{P}_W^U(A = A_H)$, where bias B>0 and the probabilities are all between 0 and 1. As we will show in section 4.3, this is in line with the beliefs held by our uninformed evaluators, i.e. those who do not have the information from the AI about the candidates' abilities. For clarity's, we will shorten the notation so that $\hat{P}_g^x = \hat{P}_g^x(A = A_H)$.

Suppose evaluators get a signal about a candidate i of gender g's ability $S_i \in \{S_H, S_L\}$ where

$P(S_i = S_H) = \begin{cases} \alpha_g \; if \; A_i = A_H \\ \beta_g \; if \; A_i = A_L \end{cases}$ and $P(S_i = S_L) = \begin{cases} 1 - \alpha_g \; if \; A_i = A_H \\ 1 - \beta_g \; if \; A_i = A_L \end{cases}$

Thus, if an evaluator sees a signal $S_i = S_H$, the evaluator becomes informed (i.e. x=I), and their posterior about that individual is

$$\hat{P}_g^I(A_H) = \frac{\alpha_g \hat{P}_g^U(A_H)}{\alpha_g \hat{P}_g^U(A_H) + \beta_g(1 - \hat{P}_g^U(A_H))}$$

We can think of α_g as the true positive rate of this information, with β_g being the false positive rate, within a gender. By allowing this information structure to vary across genders, we can allow signals to be more or less informative about men and women.

When considering the extent to which this information can possibly debias the evaluators' beliefs, we want to understand the conditions of α and β that lead to $\hat{P}_M^I(A_H) - \hat{P}_W^I(A_H) \to 0$, i.e. that upon getting a positive signal about an individual, you will have the same posterior about that individual whether it is a man or a woman. Let's first consider the case for which the signal structure is equivalent for men and women, i.e. $\alpha_M = \alpha_W = \alpha$ and $\beta_M = \beta_W = \beta$. For this, we identify that

$$\hat{P}_M^I(A_H) - \hat{P}_W^I(A_H) = \frac{\alpha \hat{P}_M^U(A_H)}{\alpha \hat{P}_M^U(A_H) + \beta\left(1 - \hat{P}_M^U(A_H)\right)} - \frac{\alpha \hat{P}_W^U(A_H)}{\alpha \hat{P}_W^U(A_H) + \beta(1 - \hat{P}_W^U(A_H))}$$

Let us consider a set of options. First, we can consider what happens as $\alpha \to 1$, i.e. the true positive rate goes to 1. Given a particular level of β, as $\alpha \to 1$,

$$\hat{P}_M^I(A_H) - \hat{P}_W^I(A_H) \to \frac{\hat{P}_M^U(A_H)}{\hat{P}_M^U(A_H) + \beta\left(1 - \hat{P}_M^U(A_H)\right)} - \frac{\hat{P}_W^U(A_H)}{\hat{P}_W^U(A_H) + \beta(1 - \hat{P}_W^U(A_H))}$$

which, while less than the original level of bias B, is positive and non-zero. This suggests that while increasing the true positive rate of the information will decrease the bias in evaluators' posteriors for those who received a high signal, it will never draw the bias to 0.

This example indicates that the gap between the posteriors evaluators hold for men and women after receiving a high signal really comes from the $\beta\left(1 - \widehat{P_g^U}\right)$ terms in the denominators. So, if instead we consider what happens as $\beta \to 0$, we can see that

$$\hat{P}_M^I(A_H) - \hat{P}_W^I(A_H) \to \frac{\alpha\hat{P}_M^U(A_H)}{\alpha\hat{P}_M^U(A_H) + 0 * \left(1 - \hat{P}_M^U(A_H)\right)} - \frac{\alpha\hat{P}_W^U(A_H)}{\alpha\hat{P}_W^U(A_H) + 0 * \left(1 - \hat{P}_W^U(A_H)\right)} = 0$$

or, as the false positive rate decreases to zero, we are able to conclude from a positive signal the same degree of information when the individual is a man or a woman.

In our study, we primarily concern ourselves with the right tail of the distribution of evaluations, or, here, those who are believed to be high ability, as these are the individuals who are most likely to be considered for employment. However, you may wonder how these information structures may instead matter for disparities in beliefs after receiving a low signal, $S_i = S_H$, which draws posteriors towards the left side of the distribution. In that case, the degree to which α and β can debias the distribution is flipped, with $\beta \to 0$ resulting in a bias smaller than B, but still positive, whereas if $\alpha \to 1$ you tend towards unbiasedness. Thus, the extent to which you can anticipate information debiasing evaluators' posterior beliefs depends on the false positive rate in the face of a high signal and the true positive rate in the face of a low signal.

AI, in this context, provides such information. In effect, it is a piece of information that either provides a high or low signal about the applicant's ability. As such, we see that, in this case where evaluators believe the signaling structure is equivalent between men and women, they will become totally unbiased in response to a positive (negative) signal only as the false positive (true positive) rate goes to zero (one).

However, there is great concern that AI is biased against minority groups, i.e. that there is not the same signal structure across men and women. Specifically, there is concern that the AI will provide more positives, both true and false, for men than women. This could be modeled as:

$$\alpha_W = x\alpha_M$$
$$\beta_W = y\beta_M$$

where $0 < x, y < 1$, essentially scaling down the chance of getting a positive outcome if one is a woman. How does this impact our insights as to the relationship between the information signal and the (un)biasedness of the final outcome? We still find that as $\beta \to 0$, the posteriors become unbiased in response to a high signal. The same is true for the response to a low signal when $\alpha \to 1$. Now, we can consider how changing x and y impacts the level of bias.

Consider

$$\hat{P}_M^I(A_H) - \hat{P}_W^I(A_H) = \frac{\alpha_M \hat{P}_M^U(A_H)}{\alpha_M \hat{P}_M^U(A_H) + \beta_M\left(1 - \hat{P}_M^U(A_H)\right)} - \frac{\alpha_W \hat{P}_W^U(A_H)}{\alpha_w \hat{P}_W^U(A_H) + \beta_w\left(1 - \hat{P}_W^U(A_H)\right)}$$
$$= \frac{\alpha_M \hat{P}_M^U(A_H)}{\alpha_M \hat{P}_M^U(A_H) + \beta_M\left(1 - \hat{P}_M^U(A_H)\right)} - \frac{x\alpha_M \hat{P}_W^U(A_H)}{x\alpha_M \hat{P}_W^U(A_H) + y\beta_M(1 - \hat{P}_W^U(A_H))}$$

As x decreases, i.e. there is a relatively lower true positive rate for women than for men, the value of $\frac{x\alpha_M \hat{P}_W^U(A_H)}{x\alpha_M \hat{P}_g^U(A_H) + y\beta_M(1-\hat{P}_F^U(A_H))}$ decreases, resulting in a larger gap between the posteriors, i.e. larger bias in the posteriors after a high signal. There are even some conditions under which, if x is low enough, the bias in posteriors is actually greater than the initial bias in priors.

If instead we consider what happens as y decreases, or that the false positive rate for women decreases relative to men, we then find that the value of $\frac{x\alpha_M \hat{P}_W^U(A_H)}{x\alpha_M \hat{P}_g^U(A_H) + y\beta_M(1-\hat{P}_F^U(A_H))}$ increases, resulting in less bias in the posteriors. This makes sense, as this indicates a stronger information from a high signal for women than for men.

Interestingly, when $x = y$, or we restrict the disparities in the true and false positives to move together, we find that this bias is actually unimportant, and one is left with the same structure as if there were no differences in information structure between men and women. Thus, when considering the impact of bias in AI, one really needs to consider whether the bias against women is occurring more for the true or false positives – if more for the true, then there is concern that will generate larger biases in the posteriors; if more for the false, this will actually diminish biases in the posteriors; and if equally for true and false, then it actually is unimportant when understanding how this information will impact disparities in the posteriors.

From here we can consider what happens in the case that the evaluators believe that the information structure is unbiased, but it is actually biased, i.e. they believe $x = y = 1$, but it is actually the case that $0 < x, y < 1$. Consider the example in which the evaluator believes $y = 1$ but actually $y < 1$. Then, the evaluator does not recognize that the false positive rate is lower for women than for men, and is equally dubious of a positive signal from a man and a woman when they should really trust a positive signal from a woman more. In that case, they will under-evaluate women relative to men. On the other hand, if the evaluator believes $x = 1$ when actually $x < 1$, they don't recognize that the true positive rate is lower for women than for men, indicating that a high signal is less informative for women. In this case, they will actually evaluate women too highly relative to men, compared to what they would do if they knew that the information structure was biased.

Appendix E: Description of Market Sample Construction

In our design, we have a pool of applicants who applied in either the *Human-Supply* or *AI-Supply* treatment, and a set of evaluations of applicants by evaluators in either the *Human-Demand* or *AI-Demand* treatment. However, there is not a perfect match between applicants and evaluations: in order to control for applicant treatment across the evaluator treatments, as well as to have multiple evaluations per applicant that was evaluated to generate stability in estimates, all female applicants and just over 200 males were selected to be shown to evaluators (and recall each evaluator was shown 2 male and 2 female profiles). As such, our resulting applicant pool that we have evaluations for was disproportionately female across both treatments, but more so for the applicants in the human treatment.

To rescale the distribution of evaluations to reflect the gender distribution of applicants, we use a replication exercise to selectively replicate evaluation observations in order to achieve the proper gender distribution. Because the randomized selection of applications for the evaluation section maintained the distribution of applicant characteristics within gender and treatment, by using a randomized selection of applicants to be replicated we can again maintain that distribution of applicant characteristics. Additionally, we chose to amplify the sample size through this replication to 10000 in each treatment in order to reduce the role of random chance generating the distributions we find.

Specifically, we identified the percentage chance an individual from each gender-treatment group would have to be selected to generate the same gender distribution of applicants in each treatment in a 50-person subsample of the 289 applicants that were used in the evaluation. Each applicant observation here contains the average evaluation score that applicant received from all evaluators by treatment, i.e. the applicant will have a score for each treatment they were evaluated under, as well as their AI-generated score. We then used these selection probabilities to randomly select 200 50-applicant subsamples for each treatment from the entire pool of evaluated applicants. These 200 subsamples were added together to form the entire group to be evaluated in this section. It maintains the distribution of characteristics by gender and treatment of both the full applicant sample and the sample of applicants used in the evaluation section, while returning the sample of applicants used in the evaluation section, with their scores, to the gender distribution of applicants in the full applicant sample.

Appendix F: Description of Estimating Evaluations with Biased AI

To model the way evaluators respond to the AI evaluation score, we use the data from the Market Sample Analysis (see Appendix E) to regress, separately for male and female applicants, an applicant's *AI-Demand* score on all of the information evaluators saw in the resumes (i.e. education, years of experience, coding languages, etc.), the AI score, and the average *Human-Demand* score received by that applicant in the full-profile.[25] Specifically, we run the regression:

$$AIDemand_{ig} = \beta_{0g} + \beta_{1g} AIScore_{ig} + \beta_{2g} HumanDemand_{ig} + \gamma_g X_{ig} + \epsilon_{ig}$$

with $g \in (M, F)$ or gender being male or female, where $AIDemand_{ig}$ is the average *AI-Demand* score provided to an individual applicant i of gender g, $AIScore_{ig}$ is the AI-generated score of applicant i of gender g, $HumanDemand_{ig}$ is the average *Human-Demand* score provided to applicant i of gender g, X_{ig} is a vector of variables including indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of experience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander) held by applicant i of gender g, and ϵ_{ig} is the residual.

From there, for each applicant, we predict what the *AI-Demand* score should be for each applicant based on their characteristics. Specifically, we calculate:

$$PredAIDemand_{ig} = \beta_{0g} + \beta_{1g} AIScore_{ig} + \beta_{2g} HumanDemand_{ig} + \gamma_g X_{ig}$$

given the characteristics of applicant i of gender g and the values of $\beta_{0g}, \beta_{1g}, \beta_{2g}$, and γ_g estimated above matching the applicant's gender g. This gives us a baseline of what our estimates look like without bias. To generate our biased AI estimates, we calculate the following values based on $b \in (0.1, 0.25, 0.5, 1)$, or a 10%, 25%, 50%, or 100% bias respectively:

$$b\_AIDemand_{iF} = \beta_{0F} + \beta_{1F}(AIScore_{iF} * (1 - b)) + \beta_{2F} HumanDemand_{iF} + \gamma_F X_{iF}$$

$$b\_AIDemand_{iM} = \beta_{0M} + \beta_{1M} AIScore_{iM} + \beta_{2M} HumanDemand_{iM} + \gamma_M X_{iM}$$

[25] The inclusion of the applicant's *Human-Demand* score is to capture any information about the application that is not captured already in the resume information as we have it in the regression, as well as the quality of the interview answers as judged by the applicant. Using instead the *No-Name* score or not including this type of term at all does not substantially impact the findings.

or, in other words, women's predicted *AI-Demand* scores are re-calculated using an AI-generated score scaled down by the amount of the bias, while men's predicted scores remain calculated using their original AI-generated scores.

We do a similar process for the *Human-Demand* scores. First, we estimate the coefficients of the following regression:

$$HumanDemand_{ig} = \beta_{3g} + \beta_{4g} AIScore_{ig} + \eta_g X_{ig} + \epsilon_{ig}$$

with $g \in (M, F)$ or gender being male or female, where $HumanDemand_{ig}$ is the average *Human-Demand* score provided to applicant i of gender g, $AIScore_{ig}$ is the AI-generated score of applicant i of gender g, X_{ig} is a vector of variables including indicators for the type of web design training (University courses, non-university courses, and/or self-taught), years of experience in web design, an indicator for holding a 4 year university degree, indicators for the type of programming languages known (Java, HTML, CSS, Python, PHP, C#, React, JavaScript, and/or Angular), and indicators for the race of the applicant (White or Caucasian, Black or African American, Hispanic or Latino/a, Native American or Alaska Native, Asian, and/or Native Hawaiian or Pacific Islander) held by applicant i of gender g, and ϵ_{ig} is the residual. This is then used to generate predicted *Human-Demand* scores:

$$PredHumanDemand_{ig} = \beta_{3g} + \beta_{4g} AIScore_{ig} + \eta_g X_{ig}$$

given the characteristics of applicant i of gender g and the values of β_{3g}, β_{4g}, and η_g estimated above matching the applicant's gender g. Again, by using this predicted value rather than the actual values, any comparisons we make between the biased AI predictions and the *Human-Demand* evaluations are not coming through one set of values having gone through this prediction method.

Finally, we do the same process as in section 5 to calculate the fraction of the n^{th} quantile that is female across the different groups, using the *Human-Supply* application behavior for the *PredHumanDemand* estimates and the *AI-Supply* application behavior for the *PredAIDemand* and *b_AIDemand* estimates.[26]

[26] The comparison of the actual and estimated values can be found in Appendix Figure A.4. The estimation procedure provides estimates very close to the actual outcomes for the *Human-Supply*/*Human-Demand* group, but less-so for the *AI-Supply*/*AI-Demand* group. However, the latter is close enough that we can continue our back-of-the-envelope calculation, particularly since the estimation procedure results in less impact of AI on diversity at the right tail than is actually the case, already shrinking the gap we are trying to close with the bias added to the AI scores.