Research Policy 51 (2022) 104420

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

This article forms part of the Special Issue on The Governance of AI

OK computer: Worker perceptions of algorithmic recruitment☆

Elena Fumagalli[a], Sarah Rezaei[b,*], Anna Salomons[a]

[a] Utrecht University, School of Economics, Netherlands
[b] University of Innsbruck, Department of Public Finance, Austria

ARTICLE INFO

JEL classification:
C9
J24
M51
O33

Keywords:
Algorithmic evaluation
Technological change
Online labor market
Online experiment

ABSTRACT

We provide evidence on how workers on an online platform perceive algorithmic versus human recruitment through two incentivized experiments designed to elicit willingness to pay for human or algorithmic evaluation. In particular, we test how information on workers' performance affects their recruiter choice and whether the algorithmic recruiter is perceived as more or less gender-biased than the human one. We find that workers do perceive human and algorithmic evaluation differently, even though both recruiters are given the same inputs in our controlled setting. Specifically, human recruiters are perceived to be more error-prone evaluators and place more weight on personal characteristics, whereas algorithmic recruiters are seen as placing more weight on task performance. Consistent with these perceptions, workers with good task performance relative to others prefer algorithmic evaluation, whereas those with lower task performance prefer human evaluation. We also find suggestive evidence that perceived differences in gender bias drive preferences for human versus algorithmic recruitment.

1. Introduction

Algorithmic predictions are increasingly used in decision-making as a result of ongoing advances in artificial intelligence (AI) and data infrastructure, and labor markets are an important domain for such applications. Specifically, algorithmic input is widely used when hiring or otherwise evaluating workers (Ajunwa and Greene, 2019), for example through automated CV-screening and other algorithmic assessment services (WSJ, 2012; Carey and Smith, 2016; Raghavan et al., 2020).[1] Broadly put, regression-based or machine-learning models are used to predict workers' future job performance, leveraging data on current and previous job applicants and/or existing workers at the firm.

Despite a large body of work evaluating the performance of algorithms in predicting outcomes across a wide range of settings, much less is known about how people interact with these technologies. For labor markets in particular, studies on the use of algorithms have focused on the employer perspective[2]: as yet, there is only limited evidence on how *workers* perceive algorithmic versus human evaluation, as we discuss below. However, this question is critical for understanding how the increasing use of algorithms in the labor market affects worker welfare and organizational commitment.[3] Worker welfare also matters for governments aiming to regulate the use of algorithms in labor markets.

☆ Helpful input from Dominik Prugger on an earlier version of this project is gratefully acknowledged. We are grateful to Maria Savona and two anonymous reviewers for helpful comments, and to Lucky Belder and Nina Bontje for helping to interpret legal provisions concerning algorithmic versus human evaluation. This work was supported by a Utrecht University Institutions seed money grant. Salomons thanks Instituut Gak, Netherlands for financial support.

* Corresponding author.
E-mail addresses: e.fumagalli@uu.nl (E. Fumagalli), sarah.rezaei@uibk.ac.at (S. Rezaei), a.m.salomons@uu.nl (A. Salomons).

[1] Algorithmic recommendations are also being used to guide workers' job search (e.g. see Belot et al. 2018, Goos et al. 2019); and for predicting worker turnover to maximize worker retention by for example IBM's Watson Analytics (IBM, 2016).

[2] This literature suggests that algorithmic screening is a valuable hiring tool for firms, because it saves on hiring costs and can lead to a better prediction of applicants' job performance, a higher match quality, and increased worker retention. Specifically, firms receiving algorithmically recommended applicants have higher fill rates for vacancies without crowding out non-recommended job applicants (Horton, 2017); and algorithmically recommended candidates are more likely to pass interviews and receive a job offer, more likely to accept job offers when extended, and also more productive once hired as employees (Cowgill, 2020). Combined with the low (marginal) cost of automated versus human evaluation, these findings help explain the broad appeal of these technologies to employers.

[3] The broader relationship between worker welfare and technology use is the subject of a related literature in innovation studies which has investigated the effect of access and use of technologies on workers' job satisfaction (Castellacci and Viñas-Bardolet, 2019), mood (Venkatesh and Speier, 1999), and well-being (Lohmann, 2015; Pénard et al., 2013; Graham and Nikolova, 2013). See Castellacci and Tveito (2018) for a review of the literature on how Internet use affects well-being and see Erdogan et al. (2012) for a review of the literature on the relationship between life satisfaction and the work domain.

https://doi.org/10.1016/j.respol.2021.104420
Received 30 August 2020; Received in revised form 22 October 2021; Accepted 29 October 2021
Available online 18 November 2021

In this paper, we study workers' perceptions of and preferences for algorithmic evaluation, as compared to human evaluation. In particular, we mimic a hiring setting, using workers from Amazon's online labor market Mechanical Turk (MTurk) as subjects. We design experiments to uncover causal evidence about the factors driving preferences for human versus algorithmic prediction, allowing us to abstract from considerations that are unrelated to differences in evaluation. To our knowledge, we provide the first incentivized analyses of workers' preferences for algorithmic versus human evaluation, what these preferences are based on, and how they differ across different worker subgroups.

Workers are registered in the U.S. or Canada and selected on quality.[4] After performing a job task, we elicit workers' willingness to pay for their preferred recruiter, human or algorithmic. Importantly, we inform workers that both recruiters have access to the same information: performance on a job test as well as a set of pre-determined personal characteristics (age, education, gender, and ethnicity) that are usually available on (or can be inferred from) a CV. Further, all recruitment decisions are communicated online, at the same time, and without in-person contact with human recruiters. This allows us to isolate as much as possible any preferences based on recruiter decision-making, rather than other factors such as time preferences, (dis)tastes for interpersonal interaction, or concerns about data privacy.

We go beyond describing incentivized preferences by implementing two experiments in this hiring setting. These interventions are aimed at understanding the causal impacts on recruiter choice of the two major inputs in recruitment decisions: workers' observed task performance and their personal characteristics (specifically gender). In the first experiment (*information experiment*), we study the effect of task performance on recruiter preference by randomly providing workers with information about their task performance compared to others. In the second experiment (*gender bias experiment*), we focus on gender bias by having workers compete against a fictitious worker whose gender is randomized. This allows us to obtain worker perceptions of recruiters' relative gender bias.

This paper makes three contributions which are related to two strands of literature as well as current policy discussions. Our first contribution is to study algorithmic versus human evaluation from the perspective of workers. This is related to a literature which considers how users, consumers, and other decision-makers perceive and use algorithmic input. Some studies document aversion to using algorithmic prediction (Dietvorst et al., 2015, 2018; Lee, 2018; Yeomans et al., 2019; Newman et al., 2020), particularly when algorithms are observed making mistakes (Dietvorst et al., 2015). However, acceptance of algorithmic input is increased when decision-makers have the option to modify the algorithm's forecast (Dietvorst et al., 2018) or are given more information about the functioning of the algorithm (Yeomans et al., 2019). Other studies find algorithmic input is valued more than recommendations from an external human adviser in a range of settings, although decision-makers still prefer their own judgment over an algorithm's, to the detriment of prediction accuracy (Logg et al., 2019; Hoffman et al., 2018), and particularly in the hiring context, workers prefer being judged by a human rather than an algorithm (Dineen et al., 2004). However, this preference may depend on the nature of the task: individuals perceive algorithms as less fair and reliable than humans when human skills are required, and equally reliable when mechanical skills are required (Lee, 2018).

In contrast to this literature, we consider algorithmic versus human evaluation from the perspective of workers rather than decision-makers in the hiring process. This is an important distinction because recruitment is different from other scenarios where algorithmic or human input is chosen: in previously studied settings, decision-makers were asked to maximize prediction accuracy, but workers will instead want to maximize the evaluation score the (human or algorithmic) recruiter gives them. That is, it is not necessarily in the worker's interest to choose the recruiter with the most accurate prediction of their ability. To our knowledge, ours is the first study to elicit incentivized preferences of workers for algorithmic or human evaluation.

This worker perspective is central to current policy discussions on the use of algorithmic decision-making. The European Commission's new proposal on regulating the use of AI (European Commission, 2021) is a clear example of this, and the proposal specifically classifies AI for job screening and evaluation as high-risk: "*AI systems used in employment, workers management and access to self-employment, notably for the recruitment and selection of persons, for making decisions on promotion and termination and for task allocation, monitoring or evaluation of persons in work-related contractual relationships, should also be classified as high-risk, since those systems may appreciably impact future career prospects and livelihoods of these persons.*" Similarly, various states in the U.S. have (pending) legislation requiring data disclosure for automated decision-making in hiring (State of Illinois, 2021), and for using acquisition methods that minimize the risk of adverse and discriminatory impacts resulting from the design and application of automated decision systems (State of California, 2021; State of Vermont, 2021). Further, the European Union has already included provisions aimed at protecting people from being subjected to fully-automated decision-making under certain conditions in 2016 General Data Protection Regulation (GDPR) legislation.[5] These legislative efforts targeted at protecting those who are algorithmically evaluated underscore that not only the decision-makers but also those subjected to algorithmic decisions should be studied.

Our second contribution is to go beyond eliciting preferences to uncover causal evidence on two potential determinants of workers' recruiter choice: their task performance, and their gender. Any such impacts respectively result from perceived differences in the extent to which recruiters consider task performance, and perceived differences in recruiters' gender bias. This is related to a literature studying bias in algorithms, either in isolation or compared to human bias (see Cowgill and Tucker, 2020; Köchling and Wehner, 2020, for an overview). This literature is in part motivated by the concern that algorithms could perpetuate or worsen existing biases through having biased objectives or biased training data (see Obermeyer et al., 2019), or through built-in financial incentives (Datta et al., 2015; Lambrecht and Tucker, 2019).[6] The concern about algorithmic bias in worker evaluation settings also permeates policy discussions: for example, the European Commission (2021) proposal states that "*Throughout the recruitment process and in the evaluation, promotion, or retention of persons in work-related contractual relationships, such systems may perpetuate historical patterns of*

[4] Specifically, having completed at least 100 MTurk tasks with an approval rating of at least 95%.

[5] Article 22(1) of the GDPR states that "*The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention". Profiling is defined as "Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular, to analyze or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behavior, location or movements.*" A 2021 lawsuit of workers against the ride-sharing firm Uber (Rechtbank Amsterdam, 2021) cites these GDPR protections, arguing they had been removed from the platform based on automated decisions made by fraud detection software — in this case, however, Uber was able to show humans were meaningfully involved such that there was no fully-automated decision-making.

[6] For example, female job seekers are found to be presented with fewer ads for high-paying jobs (Datta et al., 2015), and fewer jobs in science, technology, engineering, and math fields (Lambrecht and Tucker, 2019) as a result of advertising algorithms maximizing cost-effectiveness.

discrimination, for example against women, certain age groups, persons with disabilities, or persons of certain racial or ethnic origins or sexual orientation." While the literature investigates the existence of bias in algorithmic (and human) decision-making outcomes, we study to what extent perceived differences in bias between human and algorithmic evaluation determine recruiter preferences for those subject to these decisions. This is related to a recent set of studies which document bias perceptions in different settings (Saxena et al., 2019; Newman et al., 2020; Wang et al., 2020). Distinct from these studies, we consider a specific hiring scenario and also investigate how perceived differences in bias causally impact worker choices.

We find that workers do perceive human and algorithmic evaluation differently and are willing to pay to obtain their preferred recruiter, even though both recruiters are given the same inputs for evaluation in our controlled setting. Specifically, human recruiters are perceived to process the information available to them in a more biased and error-prone way, and to place more weight on personal characteristics, whereas algorithmic recruiters are seen as more transparent and as placing more weight on task performance. Consistent with these perceptions, we uncover causal evidence that workers with good task performance relative to others prefer algorithmic recruiters, whereas those with lower task performance prefer human recruiters. In terms of the relative gender bias in recruitment, workers are found to choose human or algorithmic recruitment in part because they perceive the human recruiter as more strongly favoring men with worse task performance.

This paper is organized as follows. Section 2 outlines our experimental set-up. Section 3 describes the data and Section 4 presents our results. We discuss our findings in the context of governance of artificial intelligence in Section 5, and Section 6 concludes.

2. Recruiter choice experiment

We run two experiments mimicking a hiring setting in which a firm tries to predict job applicants' ability at performing a task and where workers have the possibility to choose their favorite recruiter: algorithmic or human (recruiter choice experiments).[7] Experiments are conducted using oTree (Chen et al., 2016). Specifically, we recruit workers to perform a real-effort number-finding task. In this task, developed by Buser et al. (2020), workers have to find the two unique numbers which add up to 100 out of a 3 × 3 matrix containing nine numbers between 1 and 99. We chose this task because how fast a worker gives the correct answer is a well-defined measure of performance. Further, speed varies across workers but the task is not prohibitively difficult nor time-consuming. Importantly, workers cannot cheat on this task even in a non-monitored environment.[8]

We recruit workers from Amazon's Mechanical Turk (henceforth MTurk), an online labor market where employers offer tasks to a large pool of potential workers. MTurk was founded in 2005 for crowdsourcing small labor tasks but has been widely used for social experiments since then (see Paolacci et al., 2010). Compared to laboratory experiments, online labor markets have the important advantage that a large number of diverse workers can be recruited at a reasonable cost. Past research has shown that the outcomes of standard experimental games (ranging from prisoner's dilemma to interactive multi-period public goods games) do not differ substantially between workers on MTurk and in economic lab settings (Horton et al., 2011; Arechar et al., 2018). However, since MTurk workers (henceforth, workers) can leave the experiment at any time, the literature calls for participation restrictions both before and after the experiment: we outline ours in Section 3.

We run two recruiter choice experiments: an information experiment and a gender bias experiment. Choices in both experiments are incentivized. The former is a piece-rate experiment aimed at uncovering the effect of task performance. By contrast, the latter is a tournament experiment where the worker competes with a fictitious competitor, aimed at uncovering the effect of gender bias.

Each experiment has a different treatment and a different set of incentives. In what follows, we outline the two experimental set-ups, explain how recruiters are designed, and describe the resulting data.

2.1. Experimental set-up

The information experiment consists of six phases. In phase one, workers enter demographic information: age, gender, ethnicity, and education level.

In phase two, workers are shown the number-finding task, and given an example to familiarize themselves: this is a non-incentivized practice round. We inform workers that there are 10 subsequent rounds of the same task (with different numbers) and that the time in one of these rounds will be selected and sent to the recruiter. Workers are given 90 s to play each round, and are informed when their chosen answer is incorrect: they can use any remaining time to find the correct pair of numbers. After completing the 10 rounds, workers are informed that their time in round 9 is selected. We call it recruiter-observed task time (henceforth observed task time). The workers are shown their observed task time (in seconds) as well as their average task time across all rounds.

Our information experiment consists of two related treatments where treated workers receive additional information. In treatment 1a, we inform workers about their observed task time relative to that of past workers. Specifically, we show the median observed time of past workers and tell workers whether they were faster or slower than that. In treatment 1b, in addition to the information provided in treatment 1a, we inform workers about their hypothetical pay-off if it were only based on their observed task time (i.e. ignoring personal characteristics). In both treatments, workers in the control group do not receive any additional information apart from their observed task time and their average task time across all rounds. By comparing treated workers' recruiter choice to that of control group workers, treatments 1a and 1b help us understand to what extent workers choose a particular recruiter based on their task performance.

In phase three, workers are told that a firm wants to hire workers who are good at the number-finding task, and will use recruiters to predict workers' average task time based on their characteristics (age, gender, ethnicity, and education level) as well as their observed task time. This set of recruiter inputs is chosen to mimic characteristics usually contained on (or inferred from) real-world applicant CVs, combined with a recruitment task which is designed to inform on workers' ability for the job they are being considered for. Two types of recruiters are described: an algorithmic recruiter, and a human recruiter.[9] Workers are also informed that both recruiters are given the same set of inputs, but may give different predictions of the worker's average task time. After this explanation, workers are asked to choose which recruiter they would prefer to be evaluated by. To incentivize this choice, workers are informed that they will receive a payment (up to $2, piece rate) based on the recruiter's prediction of their average task time.

In phase four, we assess workers' willingness to pay for their preferred recruiters. This is achieved by first assigning each worker to

[7] Our experiments have been given clearance by the Ethical Review Committee of Utrecht University's Faculty of Law, Economics, and Governance. Further experimental details can be found in Appendix A.1.

[8] Niederle and Vesterlund (2007) use a similar task but with adding numbers rather than finding the correct pair: however, without monitoring, workers could cheat by using a calculator.

[9] See Appendix A.1 for details on the recruiter descriptions.

the recruiter they did *not* select. Next, workers are given $0.40 (corresponding to 13%–20% of the maximum task payment) and asked to state their willingness to pay (between $0 and $0.40) to obtain their preferred recruiter. Following Becker et al. (1964), workers are informed that an amount between $0 and $0.40 is randomly drawn: if the stated willingness to pay is equal to or greater than the drawn amount, workers are able to get their preferred recruiter and must pay the drawn amount, while any amount not spent is theirs to keep. Conversely, if the stated willingness to pay is lower than the drawn amount, the recruiter is unchanged (and workers retain the $0.40).

In phase five, workers are asked to estimate the average task time each recruiter (i.e. algorithmic and human) predicts for them. To incentivize this phase, one of the estimates is randomly chosen for payment. The more accurate the estimate, the higher the pay-off: payment declines linearly in the mean squared prediction error.[10] This payment is up to $0.5.

In the sixth and final phase, workers answer several survey questions about their beliefs concerning human and algorithmic recruiters, about their risk preferences (see also Dohmen et al., 2011), and a short version of big five personality traits (Gerlitz and Schupp, 2005). At the end of phase six, workers receive a $0.1 participation fee.

Phases one, two, four, and six of the gender bias experiment are identical to those of the information experiment; phases three and five differ. For the gender bias experiment, we change our experiment to a tournament set-up. Specifically, in phase three we assign each worker to compete against a fictitious worker whose education, ethnicity, and age are the same as the worker's. However, we randomly vary the gender of the competitor as well as their observed task time, in a 2-by-2 design. To make these tournaments close races, we set the observed task time of the competitor to either 1.18 s faster or slower than the worker's. Treated workers are those assigned to a different-gender competitor. After completing the number-finding task, workers are informed about their observed task time as well as the observed task time and personal characteristics of their competitor and ask to choose which recruiter will evaluate both competitors. To incentivize this choice, workers receive up to $1 based on their predicted average time in the number-finding task and $2 bonus if they are predicted to be faster than the competitor.[11]

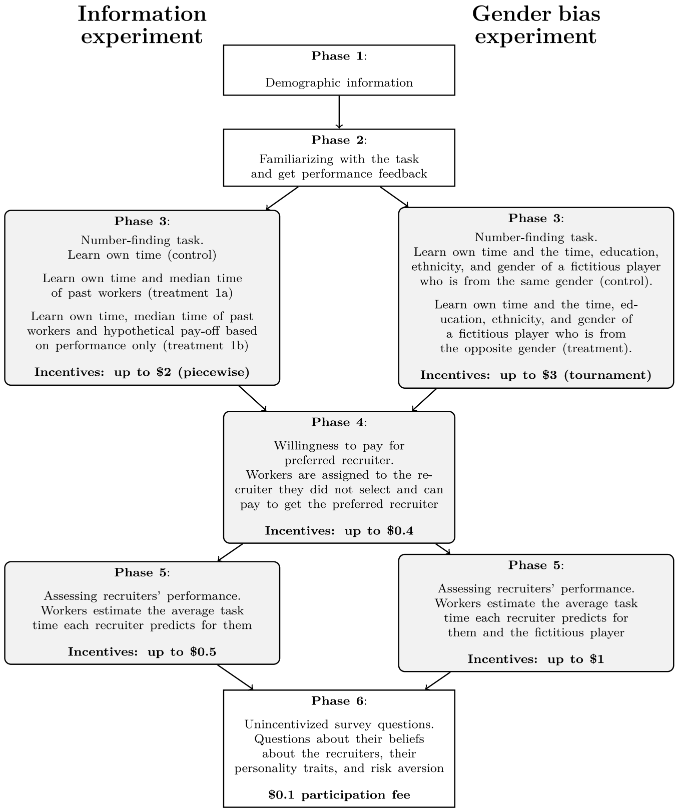
In phase five, workers are asked to estimate the average task time each recruiter (i.e. algorithmic and human) predicts not only for them but also for their competitor. The payment for this phase is up to $1.

The gender bias experiment allows us to study the effect of competitor gender on recruiter choice. Specifically, by comparing the recruiter choice of workers matched with a different-gender competitor to those with a same-gender competitor, we can study whether workers perceive one type of recruiter to be more gender biased than the other. Note that the difference relative to same-gender competitors nets out any overall impact of the competitive setting on recruiter preference.

The structure of the experiments and their incentive schemes are described in Fig. 1. The distribution of the total earnings in the full sample and by experiment is shown in Figure A.8.

2.2. Recruiter training

To avoid deception in our experiments, we use actual human and algorithmic recruiters trained on past data. To achieve this, we collected data from two pilot experiments: an MTurk experiment with 345 MTurk workers on April 17 and August 5, 2019, and a lab experiment with 22 participants on June 13, 2019.



Information experiment
Gender bias experiment
Phase 1:
Demographic information
Phase 2:
Familiarizing with the task and get performance feedback
Phase 3:
Number-finding task.
Learn own time (control)
Learn own time and median time of past workers (treatment 1a)
Learn own time, median time of past workers and hypothetical pay-off based on performance only (treatment 1b)
Incentives: up to $2 (piecewise)
Phase 3:
Number-finding task.
Learn own time and the time, education, ethnicity, and gender of a fictitious player who is from the same gender (control).
Learn own time and the time, education, ethnicity, and gender of a fictitious player who is from the opposite gender (treatment).
Incentives: up to $3 (tournament)
Phase 4:
Willingness to pay for preferred recruiter.
Workers are assigned to the recruiter they did not select and can pay to get the preferred recruiter
Incentives: up to $0.4
Phase 5:
Assessing recruiters' performance.
Workers estimate the average task time each recruiter predicts for them
Incentives: up to $0.5
Phase 5:
Assessing recruiters' performance.
Workers estimate the average task time each recruiter predicts for them and the fictitious player
Incentives: up to $1
Phase 6:
Unincentivized survey questions.
Questions about their beliefs about the recruiters, their personality traits, and risk aversion
$0.1 participation fee

Fig. 1. Experimental design. *Note*: gray boxes represent the incentivized phases.

The algorithmic recruiter is designed as coefficients from an OLS regression of average task time on workers' age, gender, ethnicity, education level, and observed task time, across all 345 workers of the MTurk pilot experiment. These coefficients are then used to predict the average task time of workers who obtain the algorithmic recruiter in the information and gender bias experiments.

For the human recruiter, we conducted a lab experiment in the Experimental Laboratory for Sociology and Economics (ELSE) at Utrecht University. Here, 22 participants were paid to act as recruiters and evaluate the average task time of 83 of the 345 MTurk workers. The human recruiters first performed two rounds of the number-finding task to understand it. Next, each of them individually evaluated either 41 or 42 MTurk workers. All recruiters received 3 euros as a show-up fee and up to 8 euros for their recruitment performance. This performance is defined as the squared distance of their prediction from the actual average task time of the workers they evaluated: after completing the prediction task, one of these predictions was randomly selected for payment. We use an OLS regression with interaction terms for each recruiter to compute 22 sets of coefficients for predicting workers' average task time. One of these sets of coefficients is randomly assigned and used to predict the average task time of workers who obtain the human recruiter in the information and the gender bias experiments.

Both recruiter types assign positive weights to observed task time, and also give non-zero weights to personal characteristics. Both algorithmic and human recruiters tend to predict faster average task time for those with fast observed task time, and for male and younger workers. Specifically, out of 22 human recruiters, 13 place more weight on task performance than does the algorithm, 17 predict a faster average task time for male workers, and 16 predict a faster average task time for younger workers. Human recruiters (20 out of 22) also tend to predict faster average task times for more highly educated workers, while the algorithm does not. Non-whites are predicted to have a faster average

[10] See Appendix Table A.1 for a detailed payment table.

[11] This design is adjusted from Niederle and Vesterlund (2007), where the worker who loses the tournament receives no payment. We pay a baseline pay of up to $1 to ensure a minimum payment to all workers.

task time by the algorithm, while predictions across human recruiters vary widely. In particular, 10 out of 22 human recruiters predict a slower average task time for non-white workers than white workers. Coefficients used for both the algorithmic and human recruiters are reported in Appendix A.1.5. Note that these weights are not our object of study: we only employ real algorithmic and human recruiters to avoid deceiving participants.

Figure A.5 shows the kernel density of prediction errors for both human and algorithmic recruiters, defined as the worker-level difference between average task time predicted by the recruiter and the observed task time of the worker in the main experiments. The graph shows that the algorithmic recruiter has a lower variance. This is not surprising since the human recruiter is randomly assigned from a set of 22 recruiters. Both distributions have a mean close to zero, but on average the human scores the workers slightly more favorably than does the algorithm.

3. Data description

We impose three ex-ante restrictions on experiment participation (see Mason and Suri, 2012; Goodman et al., 2013; Paolacci and Chandler, 2014): workers have to be registered in the U.S. or Canada, have an internal MTurk approval rating for past tasks of over 95%, and have at least 100 completed tasks.[12] We allow every worker to participate at most once, and exclude workers who did not complete the entire experiment (609 workers in total). We also eliminate a small number of workers who on average take more than 55 out of the allotted 90 s to complete the tasks (21 workers in total): these workers likely just let time run out on most tasks, which makes their data uninformative.

In addition to the ex-ante restrictions, we ask two attention check questions in phase three of the information and gender bias experiments: (1) "The algorithm and the human recruiter make the hiring decision based on the same information" (correct answer = yes); (2) "The algorithmic recruiter may assign a different score to you than the human recruiter" (correct answer = yes). All results in Sections 4 and 5 are complemented with a robustness check on the subsample of workers who gave the correct answer to both questions.

Experimental data was collected between April and July 2020, in 17 separate sessions. In total, we have 1725 valid workers across both recruiter choice experiments. Out of these workers, 1056 participated in the information experiment (with 681 treated), and 669 participated in the gender bias experiments (with 348 treated). Table 1 shows descriptive statistics on each of these four sets of workers. Male is a dummy variable indicating whether the worker self-identifies as male. Age is a continuous variable constructed using a variable collected in bands by taking the median point of each band (less than 15 years old; 15–19; 20–24; 25–29; 30–34; 35–39; 40–44; 45–49; 50–54; 55–59; 60–64; 65 years or older). Non-white is a dummy variable indicating the worker self-identifies as either: Hispanic or Latino; black or African American; Asian; American Indian or Alaska native; Middle Eastern or North African; Native Hawaiian or other Pacific Island; or Some race or ethnicity other than white. Higher educated is a dummy variable indicating the worker completed at least some college at the undergraduate level (possibilities were: some secondary education; completed secondary education; trade/technical/vocational training; some undergraduate education (college or university); completed undergraduate education; some postgraduate education; completed postgraduate education). Risk propensity is a categorical variable measuring how much the respondent is willing to take risk: this variable ranges from 0 (not at all willing to take risks) to 10 (very willing to take risks). Finally, Conscientiousness, Agreeableness, Neuroticism, Openness, and Extroversion are indices capturing the big-five personality traits, constructed from a set of 15 questions[13]: the minimum score for each trait is 1 and the maximum 15.

Table 1 shows that on average, workers take around 20 to 25 s to solve the round-9 task — this performance will be observed by recruiters. Across all ten rounds, they take 18 to 21 s on average. Slightly more than half of the workers are male, and around 30% are non-white. Workers' average age is around 37, and two-thirds have at least some undergraduate education. It is reassuring that the large majority of workers answer each of the attention questions correctly: however, our Appendix contains robustness checks for all specifications where we only retain the subsample of workers who answered both questions correctly. This sample contains 747 out of 1,056 = 71% of all workers for the information experiment and 407 out of 669 = 61% of all workers for the gender bias experiment.

Overall, the observable characteristics are well-balanced between treated and control groups. However, since there are some minor differences, we also report specifications where we control for all observables, including risk propensity and personality traits. Lastly, our data were collected over a number of sessions taking place on different dates: since worker characteristics can vary by date, we include fixed effects for the date of data collection in all our specifications (even the ones without additional controls). Our conclusions are not affected by this.

4. Results

To study how workers perceive algorithmic versus human recruitment, we start by documenting qualitative survey evidence of recruiter perceptions (Section 4.1). Next, we go beyond survey measures to study workers' incentivized recruitment preferences by analyzing recruiter choice and willingness to pay for one's preferred recruiter (Section 4.2). Lastly, Section 4.3 presents causal evidence on what drives recruiter choice, based on our two experimental interventions.

4.1. Recruiter perceptions

Fig. 2 shows how the algorithmic and human recruiter are perceived across nine different dimensions: fairness, discrimination, prediction accuracy (i.e. not being error-prone), transparency, simplicity, familiarity, speed, and as giving importance to workers' characteristics or as giving importance to workers' observed task performance. The figure presents results where the five-point Likert scale is collapsed into three categories: disagree, neutral, and agree. This highlights that compared to the human recruiter, the algorithmic recruiter is typically perceived as more fair, more transparent, simpler, faster, and as placing a higher weight on workers' task performance. On the contrary, the human recruiter is perceived as more discriminatory, more error-prone, more familiar, and as placing a higher weight on workers' personal characteristics.[14]

These basic survey results suggest both recruiter types are perceived differently by MTurk workers. We now turn to consider to what extent these perceptions are accompanied by different recruiter preferences and the correlates of these choices.

[12] These three measures are also recommended for avoiding bots as participants.

[13] These questions are outlined in Appendix A.1.6.

[14] To avoid framing, we randomize the order in which the statements are presented: this means that some workers are asked to agree or disagree with statements presented as "The human recruiter will give more weight to my personal characteristics than the algorithmic recruiter" (left panel) and others with statements presented as "The algorithmic recruiter will give more weight to my personal characteristics than the human recruiter" (right panel). Although groups who see the human mentioned first tend to choose the neutral option more frequently, results are qualitatively consistent in both groups.

Table 1
Worker descriptives.

Sample:	Information experiment		Gender bias experiment	
	Treated (1)	Control (2)	Treated (3)	Control (4)
Observed task time (round 9)	21.04 (13.87)	20.27 (14.05)	25.64 (17.66)	25.49 (18.40)
Average task time (all rounds)	18.42 (7.75)	17.43 (8.28)	21.11 (9.19)	20.76 (9.07)
Male	0.55 (0.50)	0.51 (0.50)	0.57 (0.50)	0.57 (0.50)
Age	37.01 (12.27)	37.17 (11.32)	36.40 (10.45)	36.66 (11.12)
Non-white	0.26 (0.44)	0.29 (0.45)	0.32 (0.47)	0.28 (0.45)
Highly educated	0.64 (0.48)	0.65 (0.48)	0.67 (0.47)	0.67 (0.47)
Risk propensity	5.34 (2.68)	4.99 (2.57)	6.26 (2.69)	6.29 (2.55)
Conscientiousness	11.70 (2.17)	11.87 (2.20)	11.32 (2.08)	11.27 (2.10)
Agreeableness	11.39 (2.22)	11.37 (2.23)	11.00 (2.06)	10.93 (2.13)
Neuroticism	8.63 (3.26)	8.40 (3.30)	8.61 (2.82)	8.60 (2.88)
Openness	11.39 (2.42)	11.31 (2.44)	11.67 (2.17)	11.37 (2.21)
Extraversion	8.46 (3.04)	8.10 (3.23)	9.17 (2.62)	8.89 (2.66)
First attention question correct	0.80 (0.40)	0.86 (0.34)	0.85 (0.36)	0.82 (0.38)
Second attention question correct	0.87 (0.34)	0.88 (0.32)	0.74 (0.44)	0.73 (0.44)
N	681	375	348	321

Notes: Means reported, standard deviations in parentheses. Risk propensity can range from 0 (most risk-averse) to 10 (most risk-loving). Big-five personality traits can range from 1 (lowest) to 15 (highest).

4.2. Recruiter preferences

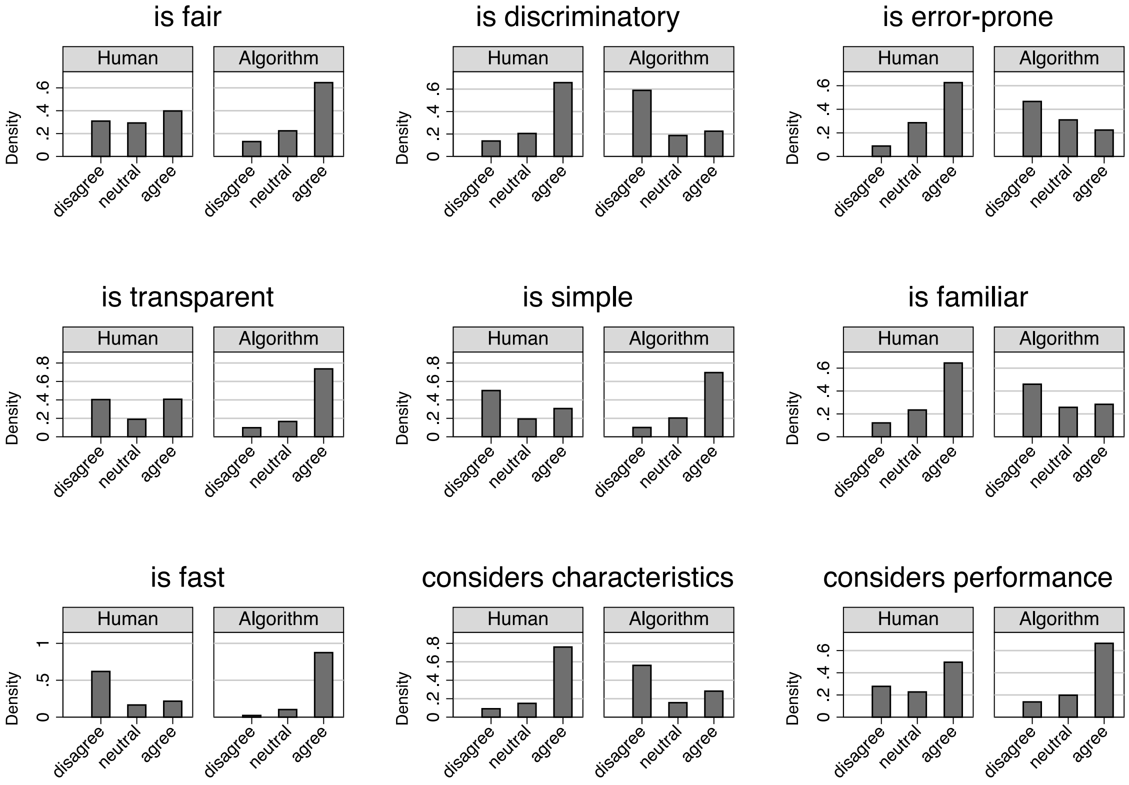
We measure workers' recruiter preferences in two main ways: their incentivized choice of recruiter, as well as their willingness to pay for this preferred recruiter. Our first finding, shown in the first panel of Fig. 3, is that workers do not unanimously prefer one recruiter over the other: around 50% of workers prefer the algorithm, and around 50% prefer the human. Moreover, the second panel shows that close to 60% of all workers are willing to pay to have their preferred recruiters evaluate them. This means that, while some workers are indifferent as measured by their willingness to pay, most are not. On average, workers give up 29% of the allocated recruiter choice budget to obtain their favorite recruiter (this rises to 47% of the budget among those with non-zero willingness to pay). Further, willingness to pay is higher for those who prefer the human recruiter as compared to those who prefer the algorithmic recruiter, as shown in the third and fourth panels of Fig. 3. All in all, this highlights that workers have different recruiter preferences, but that those who are assigned an algorithmic recruiter when they prefer the human have larger welfare losses than those who are assigned a human recruiter when they prefer the algorithm. Taken at face value, this suggests that GDPR protections have a place in providing the right to human evaluation for those who prefer it.

Our set-up also allows descriptively studying what factors underlie these preferences. First, we consider to what extent workers choose the recruiter that they expect will rate their performance more favorably: although our design is set up to minimize other factors such as preferences for in-person interaction, we cannot rule out that workers still prefer a specific type of recruiter for other reasons, such as familiarity. Second, we can study the role of observed task performance: to the extent that workers perceive the algorithm to give more weight to performance, and be less error-prone as a predictor, we would expect those with better performance to prefer algorithmic recruitment. Lastly, we can study whether worker characteristics such as gender, age, ethnicity, education level, propensity to take risks, or big-five personality type are correlated with recruiter preferences.

For the descriptive analysis of this section, we focus on the information experiment and consider the full sample of valid workers.[15] Results are presented in Table 2. Columns (1), (2), and (3) show the correlates of choosing the human recruiter, whereas columns (4), (5), and (6) show the correlates of willingness to pay for one's preferred recruiter (as a percentage of the total allocated budget), whether human or algorithm. For each dependent variable we present three specifications: in the first, reported in columns (1) and (4), the independent variables are two categorical variables equal to one if, respectively, the worker expects the algorithm or the human to rate them more favorably (the omitted category being equal expected ratings); in the second, reported in columns (2) and (5), we add the recruiter-observed task performance (henceforth "observed task performance"); and in the third, reported in columns (3) and (6), we also control for individual worker characteristics.

Table 2 highlights several findings. First, workers do choose the recruiter that they believe will rate them most favorably: when they

[15] In Appendix Tables A.7 through A.9, we present the same analysis for the gender bias experiment, as well as a robustness check on the subsample of workers who gave correct answers to the two attention checks. Results are very similar.


is fair

is discriminatory

is error-prone

is transparent

is simple

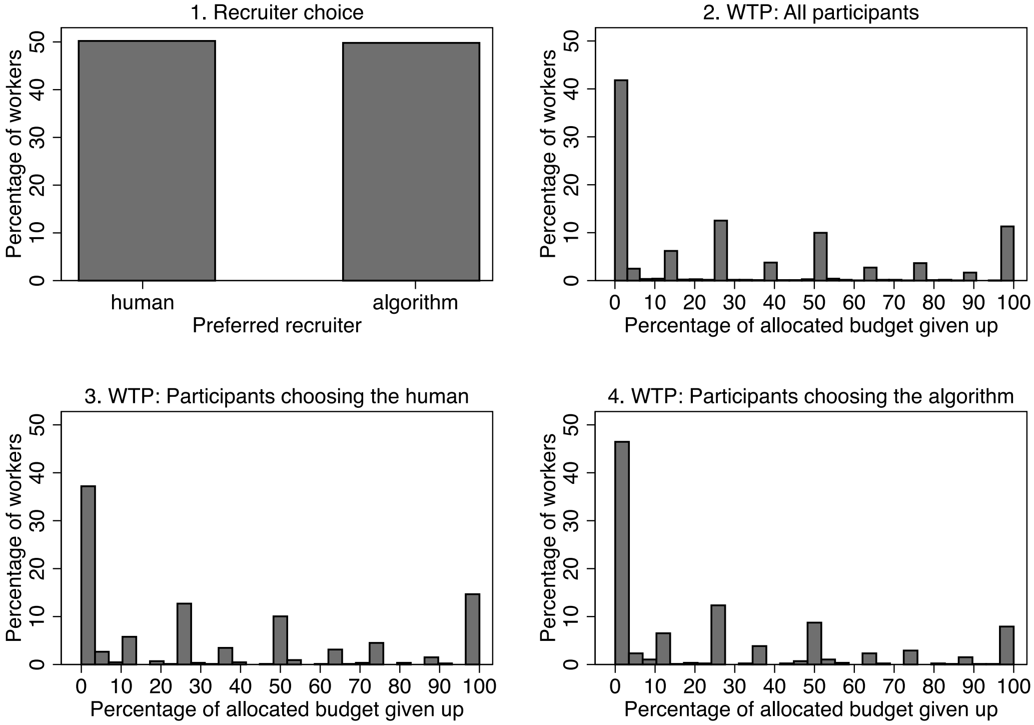
is familiar

is fast

considers characteristics

considers performance

Fig. 2. Perception of the recruiter.


1. Recruiter choice

2. WTP: All participants

3. WTP: Participants choosing the human

4. WTP: Participants choosing the algorithm

Fig. 3. Recruiter choice and willingness to pay (WTP).

expect the human to be more favorable, they prefer the human recruiter; and when they expect the algorithm to be more favorable, they prefer the algorithmic recruiter. This suggests our incentive design is successful in eliciting preferences based on expected evaluation scores. This is also evidenced by workers' willingness to pay for the recruiter that they believe rates them more favorably.[16] However, preferences are asymmetric: our third specification shows that workers who expect

[16] Note that we have designed the incentives such that rational workers will choose the recruiter which they *believe* will give them a higher rating. In the

the human to be in their favor are 15.6 percentage points more likely to choose the human recruiter, whereas workers who expect the algorithm to be in their favor are only 7.8 percentage points more likely to choose the algorithm. Similarly, preferring the human recruiter is a significant predictor of willingness to pay, even when controlling for all other covariates.

The second key finding is that the observed task performance is a significant predictor of recruiter preference. We define observed task performance as the negative of time spent solving round 9, standardized across all workers of both the information and the gender bias experiment. As such, a one-unit increase in observed task performance indicates the worker solved the round-9 task one standard deviation faster. Table 2, column (3) shows that workers who perform one standard deviation better in the observed task are 7.2 percentage points less likely to choose the human recruiter. Further, those with better observed task performance have a lower willingness to pay for their preferred recruiter (see column (6)). Taken together, this means it is particularly workers with poor observed task performances who prefer being evaluated by a human recruiter. This is consistent with a belief, documented in the previous section, that human recruiters will give less weight to task performance, and are more error-prone predictors. Specifically, less weight on task performance will benefit those with poor performance, and the competitive element inherent in recruiting may suggest to workers that adding noise to the prediction is beneficial to how they compare to others.

The third set of findings concerns worker characteristics: overall, these do not seem particularly important. Their inclusion does not substantially change the magnitude of the coefficients of the categorical variables indicating whether the algorithm or the human is favorable, and only slightly decreases the effect of the observed task performance. Including these characteristics does increase the overall precision and explanatory power of the model. All else equal, male and non-white workers are somewhat less likely to choose the human recruiter, while extroverted workers are more likely to choose the human recruiter. There are no differences in recruiter choice by worker age or education level. Further, risk-loving workers have a higher willingness to pay for their preferred recruiters. This is likely due to the uncertainty inherent in the auction which assigns the final recruiter.[17]

4.3. Recruiter choice: causal evidence

So far we have shown descriptive, albeit incentivized, evidence of workers' preferences over human and algorithmic recruitment: we now turn to the results of our two experimental interventions. The treatments are designed to identify the causal effects of the two sets of performance predictors: observed performance in the selected task, and observable characteristics, specifically gender.

4.3.1. Information experiment

In the first experiment, we study the impact of randomly making workers aware of their round-9 task performance compared to the median round-9 performance of past workers: this information treatment allows us to uncover the impact of observed task performance on recruiter preference. Specifically, if workers who are made aware of having a low (high) observed task performance are more likely to select the human (algorithmic) recruiter, the association we uncovered earlier is truly caused by workers' expectations of recruiter evaluations, and not driven by the performance difference per se (or by omitted variables).

To study this, we compare treated workers to control group workers of the same observed task performance level: the only difference between these groups is that the former has been made aware of their observed performance relative to the median. Our estimating equation is:

$$y_i = \alpha + \beta \times treat_i + \psi_i + \varepsilon_i \tag{1}$$

where i indexes individual workers. The dependent variable y_i is either a dummy for preferring the human recruiter or willingness to pay for one's preferred recruiter as a percentage of the total budget allocated to this part. $treat_i$ is a dummy equal to one when workers have been informed about their observed task performance relative to the median of past workers. We estimate Eq. (1) separately for those with observed task performance below and above the median since the information signal differs for them: the former know they are low observed task performers whereas the latter know they are high observed task performers. We also estimate this model with a set of controls contained in ψ_i. These are the observed task performance (as before, measured as the negative of the number of seconds needed to solve the round-9 task), dummies for whether the worker expects the human or algorithmic recruiter to assign them a better score (the omitted category being equal expected scores across recruiters), worker characteristics (gender, age, ethnicity, and education level), big-five personality traits and risk propensity, as well as fixed effects for the date of data collection. In the willingness to pay equation, we also control for a dummy variable equal to one if the worker prefers the human recruiter.

Panel A of Table 3 confirms that workers who are made aware that their observed task performance is low are indeed more likely to select the human recruiter compared to a control group of low performers not made aware of this. Panel B shows those being made aware of being a low performer are also more willing to pay for their preferred recruiter. This confirms a causal link between low observed performance and preference for human recruitment. However, we again find some asymmetry: workers who are made aware of having a high observed task performance do not substantially alter their recruiter choice, and have a smaller increase in their willingness to pay as a result of this information. These results are consistent with workers perceiving the human recruiter to place less weight on task performance since this is more likely to benefit those not able to send a positive performance signal.[18]

4.3.2. Gender bias experiment

In the second experiment, we assign workers to compete with a fictitious competitor: the worker who obtains the best predicted average task performance wins the tournament. The fictitious competitor is randomly presented as having either the same gender as the worker (control group) or not (treatment group); and differs only slightly in terms of observed task performance (1.18 s faster or slower). All the other characteristics are the same as the worker's. By comparing recruiter choice between men and women for workers paired with same-gender competitors to that of workers paired with different-gender competitors, we can uncover any perceived differences in gender discrimination across the recruitment methods.

Specifically, we use a Differences-in-Differences framework, allowing us to study whether being paired with a different-gender competitor has a different effect for men and women and the sign of the difference. We modify Eq. (1) as follows:

$$y_i = \alpha + \beta_1 \times treat_i + \beta_2 \times male_i + \beta_3 \times treat_i \times male_i + \psi_i + \varepsilon_i \tag{2}$$

absence of any preferences or beliefs, the rational thing to do would be not to pay for one's chosen recruiter.

[17] Appendix Table A.8 presents results on the subsample of workers who answered the attention check questions correctly: results are very similar to those obtained for the full sample of workers.

[18] In our Appendix, we present two robustness checks on these findings: first we estimate the model on a subsample of workers who gave the correct answer to both attention check questions (see Table A.10); second, we estimate the model for information treatments 1a and 1b separately (see Table A.11). The robustness checks confirm the results from our main specification.

Table 2
Correlates of recruiter choice and willingness to pay (WTP).

	Dependent variable:					
	Preference for human recruiter			WTP for preferred recruiter		
	(1)	(2)	(3)	(4)	(5)	(6)
Human favorable	0.183*** (4.09)	0.165*** (3.71)	0.156*** (3.52)	11.82*** (4.11)	10.95*** (3.83)	11.16*** (3.97)
Algorithm favorable	−0.071* (−1.71)	−0.080* (−1.93)	−0.078* (−1.90)	11.60*** (4.36)	10.98*** (4.16)	11.72*** (4.51)
Observed task performance		−0.082*** (−4.81)	−0.072*** (−4.19)		−5.053*** (−4.62)	−3.825*** (−3.48)
Male			−0.074** (−2.36)			−1.019 (−0.51)
Age			0.000 (0.07)			0.124 (1.49)
Non-white			−0.066* (−1.92)			4.032* (1.86)
Highly educated			0.045 (1.45)			0.019 (0.01)
Risk propensity			−0.002 (−0.34)			2.408*** (6.09)
Conscientiousness			−0.005 (−0.58)			−0.622 (−1.26)
Agreeableness			0.002 (0.29)			−0.292 (−0.63)
Neuroticism			0.009* (1.69)			0.737** (2.22)
Openness			0.010 (1.48)			−0.124 (−0.30)
Extroversion			0.020*** (3.85)			0.215 (0.64)
Prefer human				5.187*** (2.63)	3.839* (1.94)	3.647* (1.85)
Constant	0.458*** (12.77)	0.477*** (13.37)	0.189 (1.17)	12.69*** (5.16)	14.51*** (5.88)	0.254 (0.02)
N	1,056	1,056	1,056	1,056	1,056	1,056
R^2	0.065	0.086	0.117	0.053	0.072	0.121
adj. R^2	0.055	0.074	0.097	0.041	0.059	0.100

Notes: Estimates for the information experiment. Willingness to pay is measured as a percentage of the total allocated budget. The variable *Human (/ Algorithm) favorable* is defined as a dummy variable that is equal to one when the worker expects the human (/ algorithm) to rate them more favorably: the omitted category is workers who believe both recruiters will rate them equally. Observed task performance is measured as the negative of the number of seconds the worker took to solve the observed task (i.e. round 9), standardized to have a zero mean and unit standard deviation. All models are estimated with OLS and control for the date of data collection. *t* statistics in parentheses; * p<0.1 ** p<0.05, *** p<0.01.

where $treat_i$ is a dummy variable equal to one if worker i has been treated, i.e., if i has been paired with a different-gender competitor, and $male_i$ is a dummy variable equal to one if i is a male. Therefore, β_3 captures whether the treatment differently affects men and women: it is equal to the effect that being paired with a different-gender competitor has on men minus the effect that being paired with a different-gender competitor has on women. We also estimate this model with a set of controls contained in ψ_i. These are dummies for whether the worker expects the human or algorithmic recruiter to assign them a better score than the competitor (the omitted category being equal expected scores across recruiters), worker characteristics (gender, age, ethnicity, and education level), big-five personality traits and risk propensity, as well as fixed effects for the date of data collection.

We estimate Eq. (2) separately for workers paired with a (slightly) slower or a (slightly) faster competitor. Even though these performance differences are only 1.18 s in either direction, our previous analysis suggests that observed task performance is a strong predictor of recruiter choice. When using preference for human recruitment as the dependent variable, we would expect $\beta_3 > 0$ if human recruiters are perceived to be more biased in favor of men than algorithmic recruiters. Conversely, we would expect $\beta_3 < 0$ if human recruiters are perceived to be less biased in favor of men than algorithmic recruiters. For differences in perceived gender bias to be important for recruiter choice, we would expect $\beta_3 \neq 0$ whenever we use willingness to pay for one's favored recruiter is the dependent variable.

Results are reported in Table 4. This highlights that some perceived difference in gender bias across recruiters exists: workers perceive human recruiters to be more favorable to men who are slower than their competitors. Specifically, the coefficients in columns (3) and (4) suggest that male workers who are slower than their female competitors are 21 to 24 percentage points more likely to choose human recruitment as compared to female workers who are slower than their male competitors. This is a very sizeable effect, suggesting this perceived difference in gender bias is important for recruiter choice. However, perceived bias is less pervasive. For one, no differences are found for those who are faster than their competitor (columns (1) and (2)), suggesting there is no overall perceived bias against women in human relative to algorithmic recruitment. Further, columns (5) through (8) show workers are not differentially willing to pay to obtain their favored recruiter.[19] The lack of willingness to pay suggests differences in gender

[19] We find the same when looking at willingness to pay for both men and women when paired with a different-gender competitor, i.e. only using a single difference rather than differences-in-differences estimator.

Table 3
The impact of information on task performance on recruiter preference and willingness to pay.

A. Dependent variable: Preference for human recruiter		
	(1)	(2)
Treated, low observed task performance	0.178*** (3.30)	0.140*** (2.63)
N	595	595
	(3)	(4)
Treated, high observed task performance	0.011 (0.19)	0.002 (0.04)
N	461	461
B. Dependent variable: Willingness to pay for preferred recruiter		
	(5)	(6)
Treated, low observed task performance	9.866*** (2.70)	7.768** (2.16)
N	595	595
	(7)	(8)
Treated, high observed task performance	3.141 (0.91)	2.854 (0.84)
N	461	461
Additional controls	No	Yes

Notes: Estimates for the information experiment. Treated workers who were made aware they had a low observed task performance are compared to control group workers with low observed task performance. Treated workers who were made aware they had a high observed task performance are compared to control group workers with high observed task performance. All models are estimated with OLS and control for the date of data collection. Additional controls are ψ_i from Eq. (1). t statistics in parentheses; * p<0.1 ** p<0.05, *** p<0.01.

bias are potentially not perceived as a first-order concern for recruiter choice — this is in contrast to the results for observed task performance uncovered previously. Besides being statistically insignificant, the sign on willingness to pay is actually negative rather than positive for the subsample where perceived gender bias was found: those slower at the task than their different-gender competitors. All in all, we interpret the results from this experiment as providing some evidence of perceived differences in gender bias across human and algorithmic recruitment in our setting, which are deserving of follow-up study.[20]

5. Discussion

Although our results do not necessarily generalize to a full hiring process, the setting we mimic is closely related to a first stage of CV screening combined with some standardized job or aptitude test. Further, our controlled set-up allows us to uncover causal evidence and largely abstract from factors that are unrelated to differences in evaluation (such as general preferences for human interaction).

Our evidence suggests a number of recommendations for the governance of Artificial Intelligence (AI). For one, the current policy discussion with respect to evaluation is largely focused on bias with respect to immutable worker characteristics such as gender or race: however, our findings suggest the weight recruiters give to task performance is of foremost importance from the perspective of workers being evaluated.[21] Thus, while increased prediction accuracy could improve the match quality between workers and jobs, our results suggest it could also be perceived to give applicants less of a chance based on their previous performance (e.g. as reflected on their CV), or their performance in a job test. In our experiments, all workers can signal their performance equally (albeit noisily), but in the real world, this need not be the case. For example, workers with recent unemployment spells or skills that are not formally credentialed may not be given a chance, potentially leading to further entrenchment of labor market disadvantages. Further, there are many cases where job performance is difficult to predict ex ante, or depends on factors not easily measured in individual tests, such as social skills (Deming and Weidmann, 2020).

Within the context of our experiments, we can provide preliminary evidence for this mechanism by studying workers' recruiter choice as a function of the quality of their performance signal. Specifically, we estimate the following model:

$$\text{human}_i = \alpha + \beta_1 \times \text{bad signal}_i + \zeta_i + u_i \tag{3}$$

where i indexes individual workers. The dependent variable human_i is a dummy for preferring the human recruiter. The variable bad signal is the standardized difference between the number of seconds the worker took to solve the round-9 task and the number of seconds the worker took to solve all observed tasks on average. Average task time here captures workers' actual performance in this task, which is unobserved to recruiters but which they try to predict; whereas performance in round 9 is the performance signal observed by the recruiter. Workers who have a higher value for this difference are sending a bad signal, as the time they spent to solve the observed task is higher than their actual average task time. We also estimate this model with a set of controls contained in ζ_i. These are worker characteristics (gender, age, ethnicity, and education level), big-five personality traits and risk propensity, as well as fixed effects for the date of data collection.[22] Results in Table 5 highlight that workers who send a worse performance signal have a stronger preference for human recruitment.[23] This suggests the ability to signal job performance could be a determinant of worker preferences for algorithmic versus human recruitment.[24]

These findings are an argument in favor of using algorithms not only to select successful candidates but also to increase exploration in hiring, as has recently been proposed (Li et al., 2020). Specifically, Li et al. (2020) show that exploration can be used in hiring settings by designing an algorithm which gives a premium to learning about underrepresented groups. Our results suggest that it may be important to encourage algorithmic exploration for groups with different abilities to signal task performance. These considerations could become even more important in a setting where most training data is generated by algorithmic decisions: given the lower prediction noise in such data, fewer candidates with poor observed task performance will be hired, resulting in further entrenchment. Our findings suggest that more research is needed to design algorithms that are not only focused on selecting the best candidates based on a pool of previous successes, but which give applicants a chance to prove themselves in the job. To our knowledge, these considerations have not yet featured in recent legislative proposals.

A final policy-relevant finding is that while the GDPR is focused on avoiding fully automated decision-making, particularly in high-risk cases such as hiring, our results suggest many workers would prefer to be evaluated by an algorithm when given the choice, at least in an initial screening phase. This highlights that for the governance of

[20] Results of the robustness check on the subsample of workers with correct answers to both attention checks are presented in Appendix Table A.12 and lead to the same conclusions.

[21] If anything, our results suggest human recruiters are perceived as more gender-biased. This is consistent with a literature finding that algorithms can attenuate biases found in human decision-making (e.g., see Kleinberg et al., 2017).

[22] We limit our analysis to the information experiment, where the pay-offs from recruiter choice depend only on the worker and not also on the fictitious competitor, for whom we do not have an average task ability.

[23] Appendix table A.13 reports the results for the subsample of workers who passed the attention checks. Results are qualitatively identical.

[24] This could be studied further using an experiment where only a randomly selected subset of workers are informed about their observed task performance compared to their average ability.

Table 4
Perceived relative recruiter gender bias, by task performance.

	A. Dependent variable: Preference for human recruiter			
	Faster than competitor		Slower than competitor	
	(1)	(2)	(3)	(4)
Treated × male	−0.012 (−0.11)	−0.060 (−0.55)	0.208* (1.94)	0.244** (2.27)
N	329	329	340	340
	B. Dependent variable: Willingness to pay			
	Faster than competitor		Slower than competitor	
	(5)	(6)	(7)	(8)
Treated × male	5.346 (0.63)	2.295 (0.28)	−5.147 (−0.64)	−3.789 (−0.49)
N	329	329	340	340
Additional controls	No	Yes	No	Yes

Notes: Estimates for the gender bias experiment. Treated workers are those paired with a different-gender competitor. All models are estimated with OLS and control for the date of data collection. Additional controls are ψ_i from Eq. (2). t statistics in parentheses; * p<0.1 ** p<0.05, *** p<0.01.

AI in labor markets, any issues with AI decision-making should be compared to the counterfactual of human decision-making, rather than be considered in isolation. This also applies to concerns about bias. However, our finding that workers are more willing to pay for human than algorithmic recruitment does provide a basis for placing emphasis on the right to be evaluated by a human, as the GDPR currently does.

As with any experiment, our findings are limited to the specific setting we examine. Several dimensions seem most relevant here. First, the MTurk population is one very familiar with online environments, and therefore perhaps also more amenable to algorithmic evaluation than the public at large. If this is the case, we may understate overall preferences and willingness to pay for human evaluation: however, even in this particular population, we find significant willingness to pay based on workers' expectations of the recruiter's evaluation score.

Second, our finding that perceived relative gender bias is not a first-order concern for recruiter choice could be context-specific for a number of reasons. Firstly, our number-finding task and the target of finding average ability at number-finding may be relatively gender-neutral as compared to the real-world labor market. On the one hand, our math-based task is likely more male- than female-stereotyped, and both algorithmic and human recruiters predict a lower ability for women. However, a setting more high-powered to detect gender bias could be to directly contrast female- to male-stereotyped tasks, as in Sarsons et al. (2020). Comparing outcomes for different tasks in our hiring setting could be an interesting extension of our work. Further, expected gender bias in human recruitment may play out more through in-person interactions than in the remote setting provided here. This means our setting could artificially lower perceived differences in gender bias between recruiters.[25] And lastly, our setting has not considered other potential dimensions of bias, such as by ethnicity: however, the gender bias experiment we designed can be easily modified to study this.

Third, preferences for the use of algorithms in hiring settings may depend on the nature of the job: evidence from non-labor market settings suggests that tasks where human qualities are important are perceived to be better judged by humans than machines (Lee, 2018; Gruber et al., 2020). Our numerical task could clearly be performed well by machines: this implies the preference and willingness to pay for human recruitment we find could be higher for tasks requiring soft and/or social skills. This could be studied in our setting by replacing the numerical task with one involving such skills or allowing big-five personality characteristics to enter into the recruitment decision. However, our recruitment process does already give a role to human judgment by allowing characteristics other than task performance as an evaluation input.

6. Conclusion

Workers are increasingly faced with algorithmic evaluation in job screening: this practice is receiving substantial attention from policymakers, and is deserving of study because of its potential effects on worker welfare.

We mimic a hiring setting on the MTurk platform to elicit workers' preferences for algorithmic or human evaluation in recruitment. We find that individuals do perceive human and algorithmic evaluation differently, and are willing to pay for their preferred recruiter even though both recruiters are given the same inputs in our controlled setting. Specifically, human recruiters are perceived to be more error-prone and biased evaluators, and place more weight on personal characteristics, whereas algorithmic recruiters are seen as more transparent and placing more weight on task performance. Consistent with these perceptions, workers with good observed task performance relative to others prefer algorithmic evaluation, whereas those with lower observed task performance prefer human evaluation. After all, when observed task performance is low, less weight on performance is beneficial, and noisier prediction improves the chance of being ranked above one's labor market competitor(s). We also uncover evidence that perceived differences in recruiters' gender bias matter for preferences over human and algorithmic recruitment, because human recruiters are perceived to be more biased in favor of males with worse task performance.

Overall, these results suggest that widespread adoption of algorithmic recruitment would have heterogeneous effects across workers and possibly reduce some workers' welfare, even while improving prediction accuracy. Further, workers who send a worse signal relative to their actual task ability have the strongest preference for human recruitment. This implies that the quality of performance measurement, including but not limited to signal quality, can have an important effect on workers' preference for human recruitment. This highlights that more emphasis should be placed on prediction accuracy (rather than only bias) for understanding workers' perceptions of algorithmic evaluation. This finding suggests a role for designing algorithms which purposefully explore candidates with noisier performance signals (Li et al., 2020).

Since algorithmic evaluation from the perspective of workers is only beginning to be studied, we see many opportunities for follow-up research. An important next step would be to consider these issues in a field experiment. Further, while we have focused on evaluation

[25] It should also be noted that our experiment informs only on relative bias: it does not rule out that both recruiters are perceived to be equally gender biased. However, relative bias is the appropriate metric for recruiter choice.

Table 5
Performance signals and recruiter choice.

	Dependent variable: Preference for human recruiter	
	(1)	(2)
Bad signal	0.051*** (2.97)	0.044*** (2.61)
N	1,056	1,056
R^2	0.024	0.064
adj. R^2	0.013	0.045
Additional controls	No	Yes

Notes: Estimates for the information experiment. Bad signal is the difference between the number of seconds the worker took to solve round 9 and the number of seconds the worker took to solve all observed tasks on average; standardized to have a zero mean and unit standard deviation. All models are estimated with OLS and control for the date of data collection. Additional controls are gender, age, race, education, personality characteristics and risk aversion. *t* statistics in parentheses; * p<0.1 ** p<0.05, *** p<0.01.

differences for driving preferences over human and algorithmic recruitment, future studies could consider other dimensions of interest such as data privacy concerns or (dis)amenities from human interaction or judgment.

CRediT authorship contribution statement

Elena Fumagalli: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition. **Sarah Rezaei:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing - review & editing, Funding acquisition. **Anna Salomons:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.respol.2021.104420.

References

Ajunwa, I., Greene, D., 2019. Platforms at work: Automated hiring platforms and other new intermediaries in the organization of work. In: Work and Labor in the Digital Age. Emerald Publishing Limited, pp. 61–91.

Arechar, A.A., Gächter, S., Molleman, L., 2018. Conducting interactive experiments online. Exp. Econ. 21 (1), 99–131.

Becker, G.M., DeGroot, M.H., Marschak, J., 1964. Measuring utility by a single-response sequential method. Behav. Sci. 9 (3), 226–232.

Belot, M., Kircher, P., Muller, P., 2018. Providing advice to jobseekers at low cost: An experimental study on online advice. Rev. Econom. Stud. 86 (4), 1411–1447.

Buser, T., Niederle, M., Oosterbeek, H., 2020. Can competitiveness predict education and labor market outcomes? evidence from incentivized choice and survey measures. Working Paper.

Carey, D., Smith, M., 2016. How companies are using simulations, competitions, and analytics to hire. https://hbr.org/2016/04/how-companies-are-using-simulations-competitions-and-analytics-to-hire.

Castellacci, F., Tveito, V., 2018. Internet use and well-being: A survey and a theoretical framework. Res. Policy 47 (1), 308–325.

Castellacci, F., Viñas-Bardolet, C., 2019. Internet use and job satisfaction. Comput. Hum. Behav. 90, 141–152.

Chen, D.L., Schonger, M., Wickens, C., 2016. oTree—An open-source platform for laboratory, online, and field experiments. J. Behav. Exp. Finance 9, 88–97.

Cowgill, B., 2020. Bias and productivity in humans and algorithms: Theory and evidence from resume screening. Working Paper, Columbia Business School.

Cowgill, B., Tucker, C., 2020. Algorithmic fairness and economics. Working Paper. https://ssrn.com/abstract=3361280.

Datta, A., Tschantz, M.C., Datta, A., 2015. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. Proc. Priv. Enhancing Technol. 2015 (1), 92–112.

Deming, D., Weidmann, B., 2020. Team players: How social skills improve group performance. Working Paper.

Dietvorst, B.J., Simmons, J.P., Massey, C., 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. J. Exp. Psychol. [Gen.] 144 (1), 114.

Dietvorst, B.J., Simmons, J.P., Massey, C., 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Manage. Sci. 64 (3), 1155–1170.

Dineen, B.R., Noe, R.A., Wang, C., 2004. Perceived fairness of web-based applicant screening procedures: Weighing the rules of justice and the role of individual differences. Hum. Resour. Manag. 43 (2–3), 127–145.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G.G., 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. J. Eur. Econom. Assoc. 9 (3), 522–550.

Erdogan, B., Bauer, T.N., Truxillo, D.M., Mansfield, L.R., 2012. Whistle while you work: A review of the life satisfaction literature. J. Manage. 38 (4), 1038–1083.

European Commission, 2021. Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

Gerlitz, J.-Y., Schupp, J., 2005. Zur Erhebung der Big-Five-basierten persoenlichkeitsmerkmale im SOEP. DIW Res. Notes 4, 2005.

Goodman, J.K., Cryder, C.E., Cheema, A., 2013. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. J. Behav. Decis. Mak. 26 (3), 213–224.

Goos, M., Rademakers, E., Salomons, A., Willekens, B., 2019. Markets for jobs and their task overlap. Labour Econ. 61.

Graham, C., Nikolova, M., 2013. Does access to information technology make people happier? Insights from well-being surveys from around the world. J. Socio-Econ. 44, 126–139.

Gruber, J., Handel, B.R., Kina, S.H., Kolstad, J.T., 2020. Managing intelligence: Skilled experts and AI in markets for complex products. NBER Working Paper 27038.

Hoffman, M., Kahn, L.B., Li, D., 2018. Discretion in hiring. Q. J. Econ. 133 (2), 765–800.

Horton, J.J., 2017. The effects of algorithmic labor market recommendations: Evidence from a field experiment. J. Labor Econ. 35 (2), 345–385.

Horton, J.J., Rand, D.G., Zeckhauser, R.J., 2011. The online laboratory: Conducting experiments in a real labor market. Exp. Econ. 14 (3), 399–425.

IBM, 2016. Watson analytics use case for HR: Retaining valuable employees. https://www.ibm.com/blogs/business-analytics/watson-analytics-use-case-for-hr-retaining-valuable-employees/.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2017. Human decisions and machine predictions. Q. J. Econ. 133 (1), 237–293.

Köchling, A., Wehner, M.C., 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Bus. Res. 1–54.

Lambrecht, A., Tucker, C., 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manage. Sci. 65 (7), 2966–2981.

Lee, M.K., 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data Soc. 5 (1), 2053951718756684.

Li, D., Raymond, L., Bergman, P., 2020. Hiring as exploration. SSRN Working Paper 3630630.

Logg, J.M., Minson, J.A., Moore, D.A., 2019. Algorithm appreciation: People prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. 151, 90–103.

Lohmann, S., 2015. Information technologies and subjective well-being: does the Internet raise material aspirations? Oxf. Econ. Pap. 67 (3), 740–759.

Mason, W., Suri, S., 2012. Conducting behavioral research on amazon's mechanical turk. Behav. Res. Methods 44 (1), 1–23.

Newman, D.T., Fast, N.J., Harmon, D.J., 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. Organ. Behav. Hum. Decis. Process. 160, 149–167.

Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? Do men compete too much? Q. J. Econ. 122 (3), 1067–1101.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366 (6464), 447–453.

Paolacci, G., Chandler, J., 2014. Inside the turk: Understanding mechanical turk as a participant pool. Curr. Dir. Psychol. Sci. 23 (3), 184–188.

Paolacci, G., Chandler, J., Ipeirotis, P.G., 2010. Running experiments on amazon mechanical turk. Judgm. Decis. Mak. 5 (5), 411–419.

Pénard, T., Poussing, N., Suire, R., 2013. Does the internet make people happier? J. Socio-Econ. 46, 105–116.

Raghavan, M., Barocas, S., Kleinberg, J., Levy, K., 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 469–481.

Rechtbank Amsterdam, 2021. Case C/13/692003 / HA RK 20-302. http://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBAMS:2021:1018.

Sarsons, H., Gerxhani, K., Reuben, E., Schram, A., 2020. Gender differences in recognition for group work. J. Polit. Econ. Forthcoming.

Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., Liu, Y., 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 99–106.

State of California, 2021. Act 2021 CA A 13. https://custom.statenet.com/public/resources.cgi?id=ID:bill:CA2021000A13&ciq=ncsl&client_md=d0b9abc5423ba59c83de196a4321a141&mode=current_text.

State of Illinois, 2021. Act 2021 IL H 53. https://custom.statenet.com/public/resources.cgi?id=ID:bill:IL2021000H53&ciq=ncsl&client_md=cf812e17e7ae023eba694938c9628eea&mode=current_text.

State of Vermont, 2021. Act 2021 VT H 263. https://custom.statenet.com/public/resources.cgi?id=ID:bill:VT2021000H263&ciq=ncsl&client_md=7fcb043c609beb468b49a53ca7e6dee1&mode=current_text.

Venkatesh, V., Speier, C., 1999. Computer technology training in the workplace: A longitudinal investigation of the effect of mood. Organ. Behav. Hum. Decis. Process. 79 (1), 1–28.

Wang, R., Harper, F.M., Zhu, H., 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14.

WSJ, 2012. Your Resume Vs. Oblivion. The Wall Street Journal, https://www.wsj.com/articles/SB10001424052970204624204577178941034941330.

Yeomans, M., Shah, A., Mullainathan, S., Kleinberg, J., 2019. Making sense of recommendations. J. Behav. Decis. Mak. 32 (4), 403–414.