Contents

I. Introduction

The screening of job applicants using data-driven technologies has grown rapidly in the U.S. and is starting to pick up pace in Europe.[1] An emerging literature suggests that while internet job search was considered uneffective in the early 2000's (Kuhn and Skuterud, 2004; Kuhn and Mansour, 2014), modern algorithmic screening technologies are indeed valuable hiring tools for firms. Horton (2017) evaluates an experiment conducted on an online labor market where some firms receive algorithmically recommended applicants. Such treated firms display higher fill rates for vacancies without crowding out non-recommended job applicants. More general, Hoffman et al. (2017) show that managers deciding against the recommendations of an automated hiring test result in lower average tenures of workers hired than managers who followed the recommendations.

However, ethical concerns are being raised that these algorithms may perpetuate or exacerbate existing labor market biases, exclude vulnerable groups from the labor market, provide no legal basis for appeals, and are poorly understood by the persons being judged by them (Barocas and Selbst, 2016).[2] A full understanding of the usage and possible implications of such algorithmic hiring technologies hinges, therefore, crucially on the supply-side: the extent to which and reasons why job applicants consent to algorithmic screening, and/or change their behavior in response to such settings.

I study the decision process of job applicants when offered a choice between algorithmic and human evaluation, using an incentivized behavioral experiment conducted online. In doing so, I contribute to the literature in at least four ways. First, I show that job applicants react strongly to changes in available information about a recruitment method: After seeing the algorithmic recruiter discriminate against women, female participants overtly opt for the human, while males choose the algorithmic recruiter. In the existing and above-mentioned studies concerning algorithmic hiring recommendations, the decision process of job applicants is either largely ignored, or it is assumed that job applicants do not behave strategically themselves (Cowgill, 2018). Second, I show that underlying beliefs about the recruitment methods and payoff-maximizing behavior as well as perceived discrimination explain the choice between human and algorithmic recruiter. The extent to which participants react to communicated weights about an algorithmic recruiter seems to depend on prior assumptions and participants discounting information based on perceived fairness and discrimination. Third, my experimental design is a novel contribution to the experimental economist's toolkit when aiming to study how people perceive technology. This behavioral approach is complementary to findings that emphasize the potential of using machine-based algorithms to mitigate errors and biases in human decision making (Kleinberg et al., 2017).

The superiority of algorithmic judgement in forecasting and prediction has long been established (Dawes et al., 1989) and has recently found new attention and confirmation with the rise of machine learning models (Mullainathan and Spiess, 2017; Dietvorst

[1] See for example an articles in Forbes: Forbes and The New York Times or an academic summary of the development of automated hiring platforms in (Ajunwa and Greene, 2019).

[2] For a popularized summary of these argument, see for example (O'Neil, 2017)

et al., 2015). At the same time, all existing behavioral research looking at how people choose between algorithmic or human prediction had the participant deciding whether or not to use the algorithmic prediction, or which prediction they would rather rely on. Instead, in this paper, I ask the question how people decide if they are the subject of either human or algorithmic prediction, in the specific context of job applicants. Existing behavioral experiments investigate how individuals choose between algorithmic and human prediction. However, they partly display contradictory results ranging from algorithmic appreciation (Logg et al., 2019) to algorithmic aversion (Dietvorst et al., 2015). In case of algorithmic aversion, Prahl and Van Swol (2017) show that individuals discount algorithmic advice more than human advice, after seeing it err. Further, Dietvorst et al. (2016) show that having the option to slightly modify an algorithm's forecast, considerably increases its acceptance. Yeomans et al. (2017) use a behavioral experiment conducted on the same online labor market as is used in this experiment and show that the acceptance of algorithmic advice also depends on the perception of individuals, for example to what extent people understand the functioning of the algorithm at hand.

The experimental design of this paper has two distinct parts. First, participants submit personal characteristics and perform an incentivized task ten times. The results are then used as a training data set to construct an algorithm that predicts the average performance of the task, based on personal characteristics and on the performance in one observable round. In the second part of the experiment, new participants are recruited and repeat the first part. Next, and critically, they are informed that either the constructed algorithm or a human recruiter will evaluate their average performance. The participants then choose their preferred recruitment method. The choice is incentivized as the participants' payment depends on the performance that their chosen recruiter predicts. As potential drivers of their choice, I elicit the participants' beliefs regarding how well both methods would predict them and ten qualitative beliefs about both recruitment methods. At the same time, participants are randomized to either the baseline or the treatment group of the experiment, in which I alter the representation of the algorithmic recruiter. Participants in the baseline group receive a sparse and neutral description of both recruitment methods. The treatment communicates the functioning of the algorithmic recruiter and two of its weights, resulting from the training data set: It statistically discriminates against females compared to males and assigns a positive weight to the observed performance in one round.

The notion that an algorithmic recruiter shows signs of a gender bias is not a too far-fetched one. Unchecked algorithms trained on data sets inhibiting gender differences in outcomes, like the one I received by running an auxiliary experiment, will mostly reproduce and extrapolate existing gender biases.[3] In practice, Facebook's advertising algorithms have shown to be skewed along gender and racial lines for ads on employment and

[3] An emerging and multidisciplinary body of research investigates ways to identify and remove biases from algorithmic forecasts and predictions by either running hypothetical data-simulations and testing the results for biases (Feldman et al., 2015; Kleinberg et al., 2018a,b) or subjecting algorithms to additional fairness constraints (Chouldechova, 2017; Dwork et al., 2012; Corbett-Davies et al., 2017). On the other hand, stricter laws against algorithmic discrimination have been passed or are in the process of introduction, for example Europe's General Data Protection Legislation (GDPR) or new legislation recently brought forward in the U.S. Senate.

housing opportunities (Ali et al., 2019). Similar gender-biased results for personalized employment ads can be obtained even when the algorithm in question was specifically designed to act gender-neutrally (Lambrecht and Tucker, 2019).

The remainder of this paper is organized as follows: First, I describe the overall experimental design and the online labor market used to recruit participants. Section three presents the description of the two auxiliary experiments and describes the characteristics of the resulting algorithmic and human recruiter. Next, the experimental design of the main experiment and methodology used to analyze results is explained. Section five then shows the main results of the experiment and analyzes which factors are driving the choice between the algorithmic and human recruiter. Next, the key results are discussed with regard to their internal and external validity. I conclude with discussing the findings in respect to the current debate about algorithmic fairness and avenues for future research on the topic of algorithmic recommendations in general and job applicants recommendations in particular.

II. The Current Study

A. *Conceptual Overview and Related Work*

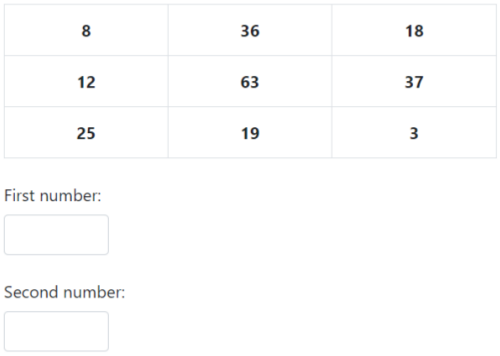
Conceptually, I experimentally mimic a hiring setting in which the employer can observe the job applicants' ability only once. The employer consequently has to predict how well this applicant will perform, on average, in the future. The applicant can then decide if his or her capability is predicted by either a human or algorithmic recruiter. To construct both recruitment methods I conduct two auxiliary experiments before new participants are presented with choosing a recruitment method in the main experiment. The results of the first experiment (henceforth called experiment A) are used to construct a simple algorithm which predicts the average performance of participants. This resembles a common real-life practice for hiring algorithms, which use a training data set of past hires to train an algorithm that then extrapolates the results to new potential hires.[4] For the second auxiliary experiment (henceforth called experiment B) new participants are recruited who predict the average performance of the participants in experiment A and who serve as the human recruiter within the main experiment. This procedure was chosen as matching participants from the main experiment simultaneously to a human recruiter in real time was logistically not possible.

In the main experiment, participants thus receive real predictions after choosing either the algorithmic or the human recruiter. This allows to incentivizing their choice: Their final payoff increases with the average performance predicted by their chosen recruitment method. The main experiment additionally elicits underlying beliefs of participants regarding both recruitment methods. First, I ask how well each method would predict their performance and second how confident they are in their belief. All three experiments were conducted using oTree as experimental software (Chen et al., 2016). The instructions for all three experiments can be found in appendix C.

[4] A detailled description of the functioning of predictive hiring algorithms can be found in Rewriting the rules for the digital age: 2017 Deloitte Global Human Capital Trends.

The real-effort task I use throughout the experiment is a number finding exercise, where participants have to report the unique two numbers which add up to 100 out of a three by three matrix containing nine random numbers between one and 99. Performance of the task is then measured as the number of seconds it takes participants to enter these two numbers. The maximum allotted time is 90 seconds. The task is an adaption from an experiment by Niederle and Vesterlund (2007), in which participants have to a add two-digit numbers. Changing the task from adding numbers to finding the correct pair was mainly chosen to prevent cheating within the online environment of the experiment.[5]

Figure 1. : One of the ten number-finding exercises



In all three experiments, participants submit their gender (female being 0, male being 1), age (in five-year intervals), most recent level of education and ethnicity. This information is then later used as input to construct the algorithm, and is given to the human recruiters to make their predictions. This is similar to real-world examples where hiring algorithms base their decisions on various information other than the task at hand.[6] Throughout all three experiments, no deception of participants was used.

The experiment is designed to show the extent and reasons to which people decide between *being evaluated* by a human or algorithmic recruiter. It is thus most similar to existing studies looking at how people decide between *using* algorithmic or human recommendations or forecasters. Dietvorst et al. (2015) experimentally manipulate the degree to which the accuracy of an algorithmic forecaster is observable and show that participants choose the algorithmic prediction less after seeing it err, even though the algorithm is the superior forecaster. Logg et al. (2019) manipulate the presentation of the

[5]Adding number could, for example, easily be cheated by participants using a calculator. Other real-effort tasks commonly employed by experimental economists (Charness et al., 2018) such as transcription of text are on the other hand very dependent on the participants online devices. The chosen task is also similar to (Heyman and Ariely, 2004), who also a number-finding exercise of numbers 1-100. Within their task, their is however no correct solution and the authors measure after how many seconds participants give up

[6]For example, the company Ideal extracts and interprets the information and text of an applicants resume and assigns them to five different degrees of quality

algorithmic device (jointly or separately) and the extent to which participants have prior experience with forecasting. Most similar to the design used in this paper is Yeomans et al. (2017) which lets participants choose between a human and algorithmic recommender system making forecasts and show that participants are reluctant to use the more accurate algorithmic forecaster, mostly because they do not understand or feel familiar with it. Further, the authors show that participants respond considerably to different degrees of transparency in explaining the algorithm. All three studies use, similar to the present research, a set of training data on which they build a simple recommender algorithm and recruit participants on the online labor market of Amazon's Mechanical Turk.

B. Doing Research on Amazon's Mechanical Turk

The main experiment and one of the auxiliary experiments were conducted on Amazon's Mechanical Turk (henceforth MTurk), an online labor market where employers offer real-wage tasks and exercises to a large pool of potential workers. MTurk was founded in 2005 mainly as a possibility to crowdsource small labor tasks but has been used widely for social experiments since then. For an extensive discussion of the functioning of MTurk and its usage within the social sciences, see Paolacci et al. (2010).

Economic experiments on such online labor markets lack some of the standard control measures in experimental economics, such as controlling the recruitment process of participants or directly supervising the experiment in a laboratory. However, past research on MTurk found that outcomes of experiments replicating standard experimental games on MTurk display no profound differences to economic laboratories. Horton et al. (2011) showe that participants recruited on MTurk behave qualitatively and quantitatively comparable to standard laboratory participants in cooperative games such as the prisoner's dilemma. Further, they respond to priming consistent with behavioral theory and behave in a similar way to a canonical experiment on framing done by Tversky and Kahneman (1981).

Arechar et al. (2018) conduct a more logistically complex interactive multi-period public goods game and conclude that standard results from laboratory studies are replicated using participants from MTurk. However, as participants on MTurk can leave the experiment at any given time and as the recruitment process cannot be as closely controlled as in laboratory experiments, the authors call for a careful selection of recruitment restrictions, both, before and after the experiment. Following further suggestions on the sampling of participants on MTurk (Paolacci and Chandler, 2014; Mason and Suri, 2012; Goodman et al., 2013), I impose three ex-ante restrictions which participants can register for the experiment: they have to be registered in the U.S. or Canada, have an internal MTurk approval rating for past taks of over 95 % and have at least 50 completed tasks.[7] For the main experiment, I also restricted registration for participants who already participated in experiment A. Ex-post, I exclude participants from the main experiment who

[7]These three measures are also recommended for avoiding bots as participants, a problem that received considerable attention within the last year, see this blog post summarizing the issue: http://turkrequesters.blogspot.com/2018/08/the-bot-problem-on-mturk.html

failed either one of four restrictions: not completing the entire experiment, entering nonsensical answers in both of the free-text boxes of the main experiment and failing either of the two sets of attention checks. All exclusion criteria were set before the start of the experiment.

III. Auxiliary Experiments

A. *Experiment A: Training the Algorithmic Recruiter*

Participant Recruitment and Sample Characteristics

Eigthy-three participants were recruited on MTurk and completed the experiment A. Out of the 83 participants, 59 % are female and 41 % male. A majority of participants, 61 %, obtained either some undergraduate education or more. 50 % of participants are in the age category of 25 - 35 and 74.7 % of all participants identified themselves as White. The average speed with which the participants solved all ten number finding exercises was 19.33 seconds and the total average payoff was US Dollar ($) 3.53 while the experiment lasted around 20 minutes, which is well above the average payoff on MTurk (Horton et al., 2011)

Experimental Design

The main goal of experiment A was to generate a data set to train the algorithmic recruiter. The participants first submitted their personal characteristics of gender, age, education and ethnicity, and then completed ten rounds of the number-finding task. All participants received a show-up fee of $ 0.10 and were informed that their additional payoff depends on their performance in one out of the ten exercises. (For clarity, I henceforth call this round "observed round") Specifically, it was communicated that their earnings were determined by how much time the participant needed to solve the chosen number-finding exercise ranging from $ 2.50 (0 seconds completion time) to $ 0 (90 seconds). The algorithm which is used in the main experiment then estimates how the personal characteristics of gender, age, education, ethnicity, and the performance in the observed round affect the average performance using a simple ordinary least squares (OLS) regression. This echoes the general idea that the employer can observe the productivity of an applicant only once and has to then predict the average performance. While there are many methods which yield a better predictive performance than OLS regression,[8] the aim of this research is not to identify the best predictive method, but to study how participants react to it. In this regard, OLS regression offers the benefits of an easy-to-understand method, whose weights are straightforward to communicate.

[8]In practice, most firms use a combination of supervised machine learning algorithms, see (Mullainathan and Spiess, 2017) for a recent overview of such methods.

RESULTS

Column (1) of Table 1 reports the weights of the OLS regression

$$(1) \qquad Time_{Alg/Hum_i} = \beta_0 + \beta_1 Male_i + \beta_2 Age_i + \beta_3 Educ_i + \beta_4 NonWhite_i + \beta_5 Observed_i + \varepsilon_i$$

where $Time_{Alg/Hum_i}$ is the average time that participant i needed to solve all exercises, estimated for either the algorithmic (alg) - or human (hum) recruiter. Further, $Male_i$ indicates if participant i is male or not, Age_i is an ordinal variable for the age of participant i in 5-year bins, $Educ_i$ is an ordinal variable describing the highest achieved level of education of participant i, $NonWhite_i$ indicates if participant i identifies as being on-White and $Observed_i$ is the time it took participant i to solve the observed round of the number-finding task. Column (1) thus shows the weights that the algorithmic recruiter in the main experiment will assign to participants. In particular, the algorithm assigns a faster (thus better) average speed to participants who are male, younger, less educated, White and have a faster time in the observed round. It thus resembles a gender-biased algorithm, statistically discriminating against women.[9]

B. Experiment B: Obtaining Predictions by Human Recruiters

PARTICIPANT RECRUITMENT AND SAMPLE CHARACTERISTICS

Twenty-two participants were recruited from Utrecht University's Experimental Laboratory for Sociology and Economics (ELSE) for experiment B. Participants of experiment B earned on average € 10.35 while the whole experiment lasted about 40 minutes. Out of the 22 participants, 16 were female, 6 male, 17 the majority, 16, was in the age category of 20-25 and 16 were obtaining or already obtained an undergraduate education, while 3 obtained a graduate's education. 15 identified themselves as White and 7 as non-White.

EXPERIMENTAL DESIGN

The participants' main task was to predict the average time of the MTurk participants from experiment A. Specifically, all participants from experiment B first performed two training rounds of the same number-finding tasks to understand the task at hand. Next, they were given the personal characteristics of the 83 MTurk participants (gender, age, education and ethnicity) and the time of the observed round from experiment A. With

[9]The algorithmic recruiter also statistically discriminates against older, more educated and non-White people. In this paper, I focus on gender-bias due to two reasons: First, communicating only one discriminatory weight within the treatment leads to a better observable treatment effect than including the weights of ethnicity, education and age as well. Second, out of the two personal characteristics which are the focus of the debate on algorithmic fairness: gender and ethnicity, gender has the advantage of a nearly 50-50 distribution within the sample, ensuring that the resulting effects are not too small.

Table 1—: Weights for the algorithmic and human recruiter (averaged)

	Algorithmic recruiter (1)	Human recruiter (2)
Constant	6.430**	5.062***
	(3.283)	(1.094)
Male	-4.096**	0.193
	(1.242)	(0.526)
Age	0.749**	0.456***
	(0.252)	(0.111)
Education	0.715†	-0.253
	(0.391)	(0.165)
Non-White	1.028	-0.702
	(1.432)	(0.617)
Observed Round	0.531***	0.515***
	(0.030)	(0.013)
Observations	83	83
R^2	0.844	0.958

Note: This table reports OLS regressions where the dependent variable indicates how many seconds participants needed to solve the number-finding exercises on average. Column (1) estimates an OLS regression equation based on the five covariates. Column (2) estimates a regression equation, based on the average speed that the laboratory participants in the second auxiliary experiment predicted. Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

this information, they had to predict the average performance of the participants in experiment A, that is, the average time it took to solve all ten number-finding exercises. All laboratory participants received €3.00 as a show-up fee and were instructed that they could earn an additional amount of up to €8.00 depending on how close their prediction was to the actual average performance of the MTurk workers. This was incentivized using a quadratic scoring rule Gneiting and Raftery (2007) between the actual average time and the prediction. After inserting a prediction for all 83 MTurk workers, one of these predictions was randomly selected for payment. The randomization was done by the participants themselves by drawing one of 83 numbers out of a closed envelope.

RESULTS

Column (2) of Table 1 shows the same regression equation as column (1), but for the average predictions submitted by participants of experiment B. On average, the human recruiter also assigns a faster (thus better) average speed to participants who are younger and have a faster time in the observed round. In comparison to the algorithmic recruiter, however, participants from experiment B did not assign a significant weight on gender. The two methods also assign different signs for both the coefficients of education and the dichotomous variable for ethnicity, without the coefficients being statistically significant.

IV. Main Experimnet

A. Participant Recruitment and Sample Characteristics

One hundred seventy participants started the main experiment through the MTurk interface. Due to difficulties in data collection, participants from the baseline and treatment group had to be recruited separately.[10] Out of the 170 participants who started the experiment, 21 were excluded ex-post because they did not fully answer all questions, and one further participant was excluded as she submitted nonsensical answers for both open-ended questions.[11]

The resulting 148 participants received an average payoff of $ 3.70 and spent an average time of 21 minutes to complete the experiment.[12] Participants in the entire sample were 53 % female. The median participant was 30-35 years old (age category 5), while the most frequent age group of was 25-30 years olds (34 participants). Most of the participants, 92, reported having received some undergraduate education or as their highest received education. 22% percent of participants identified as non-White. On average, participants needed 17.89 seconds for the observed round and an average of 16.56 seconds for all ten rounds.

B. Experimental Design

Identical to experiment A, participants first submit their personal characteristics of gender, age, education, and ethnicity, and are then instructed to perform ten iterations of the same number-finding exercises and payoff-incentivization as in experiment A.

Next, participants are informed that either an algorithmic or a human recruiter will predict their average time of solving the exercises and that both recruitment methods base their prediction on the same data: the personal characteristics (gender, age, education, and ethnicity) and the performance in the observed round. Participants are incentivized to choose their preferred recruiter by instructing them that their payoff decreases if their chosen recruiter predicts a higher (thus worse) average time.[13] Three attention checks ensure that participants understand the experimental design. If participants choose the algorithmic recruiter, their average speed of solving the exercises is calculated using the OLS regression weights from column (1) of Table 1. If participants choose the human recruiter, their average speed is determined by matching them to one of the predictions

[10] In the first recruitment session, where I recruited the treatment and control group at the same time, participants in the treatment group seemed not to understand the information-rich description of the algorithm. Confusion arose as a *higher* average speed predicted by the algorithm has to be understood as a *less beneficial* prediction for the participant. In the second recruitment process, I thus adopted a more straightforward approach and named a lower time "better" and a higher time "worse". The treated participants were recruited three weeks after the first recruitment session.

[11] No participant failed the third exclusion criteria, that is not giving the correct answer for either set of attention checks.

[12] The median time of completion is seven minutes, indicating a right-skewed distribution. The participants on MTurk did not need to finish some parts of the experiment within a specific time. For a further discussion on time spent per page or decision, see section V.A

[13] As in the number-finding exercise, the maximum payoff of $ 1.50 , equivalent to a predicted average speed of zero seconds, decreases linearly for each additional predicted second.

made in experiment B.[14]

I then elicit the participants' underlying beliefs about both recruitment methods. After having made their choice, but before they know their assigned prediction, I ask which average speed they think both the algorithmic and the human recruiter would assign them, ranging from zero to 90 seconds. This is incentivized by instructing participants that their payoff is increasing the closer their stated belief is to the actual prediction using a quadratic scoring rule Gneiting and Raftery (2007). Participants receive $ 1.00 minus the squared distance between believed and actual prediction, randomly chosen to be the belief about either the human or algorithmic recruiter. The participants also submit how confident they are in these beliefs, ranging from 0% (not confident at all) to 100% (fully confident).

The control and treatment group of the experiment differs only in the representation of the algorithmic recruiter. In the baseline of the experiment, both the algorithm and the human recruiter are presented sparsely and neutrally. While the human recruiter does not receive an extra description, the algorithm is described as: *"an automated program"*. In the treatment session of the experiment, I use an information-rich representation, which participants read before choosing their recruiter and before submitting their subjective beliefs:

"Think of the algorithmic recruiter as a tool that can poll hundreds of people and determine how many seconds each person needed to solve the exercises on average. This way, the algorithm can learn which characteristics lead to a specific average time and how gender, age, education, ethnicity, and performance in round 9 affect the average time of solving the exercises. The algorithm was trained on past data and will compare your characteristics to past participants'. Thus, it will calculate your average time based on how similar participants performed in the past.
In particular, the algorithm predicts:
- a better average performance for male participants (4.1 seconds faster) relative to female ones,
- a better average performance for participants who had a better performance in round 9 (0.5 seconds faster for every 1 second they were faster in round 9)".

Thus, the description in the treatment has two parts. The first explains the functioning of the algorithmic recruiter, while the second communicates how gender and the time of the observed round (round 9) affect the predicted average time. The first part of the description is adopted from Yeomans et al. (2017), where the description was shown leading to greater preference for an algorithmic recommender system. In comparison to Yeomans et al. (2017), I importantly include the information that the algorithmic recruiter assigns a better prediction to males relative to females and to participants who had a faster

[14]Specifically, the process first matches participants of the main experiment to one of the participants in experiment A, such that they have the same gender and a minimal distance between age, education, ethnicity, and their performance in the ninth round. Then, one of the 23 predictions made by the ELSE lab participants about this specific participant is selected randomly. This process was chosen as it guarantees that the human predictions are made about a similar participants, while keeping the variability of human judgement by not taking averages

time in the observed round.

After the results and payoffs are communicated to the participants, I elicit their qualitative perceptions about both recruiters. This is done along the dimensions of fairness, transparency, simplicity, their ability to understand the recruiter, the variability of a recruiter's prediction, their familiarity with both recruitment methods, to what extent either method might discriminate against them and which recruitment method is more likely to make errors. I elicit all these categories by using a five-point Likert scale ranging from "1 - Strongly disagree" to "5 - Strongly agree".[15]

Last, I elicit information about the participants' personalities, which will serve to further understand the decision process behind choosing either the human or algorithmic recruiter and the formation of beliefs. First, I measure risk aversion, using an incentivized five-item lottery design as in (Eckel and Grossman, 2002) for which I use the ready-made oTree code from Holzmeister (2017). Second, I measure personality traits using a 15 item questionnaire on extraversion, conscientiousness, agreeableness, neuroticism, and openness to experience, commonly referred to as the "Five-Factor model" John et al. (1999). This compact form of personality elicitation shows strong robustness of self-reported measures of personality dimensions, especially for young and middle-aged adults (Lang et al., 2011). [16]

The treatment is designed to change the way how participants perceive the algorithmic recruiter, while not changing the perception of the human recruiter. As participants learn about how the algorithm weighs gender and the observed round, I can thus observe how they react to the new information and possibly self-select themselves along these two characteristics. For the potential outcomes of the experiment, I set the following hypotheses: (1) Female (male) participants in the treatment will choose the algorithm to a lesser (higher) degree than in the baseline of the experiment. (2) Treated participants with a relatively slow (high) performance in the observed round will choose the algorithm to a lesser (higher) degree than in the baseline of the treatment. (3) If these two assumptions are found, they should be reflected in the underlying subjective beliefs for the algorithmic recruiter. That is, the difference in believed predictions about both recruiters should be higher after seeing the information-rich treatment and most pronounced along gender and time in the observed round. (4) Analogous to (Yeomans et al., 2017), the treatment will lead to a higher perceived transparency of the algorithmic recruiter.

C. *Methodology*

In the following analyses, I will estimate three OLS regression estimates, using different sets of covariates. To keep the number of regression equations to a minimum, Y_i

[15]Considering for example fairness, I ask participants to agree/disagree to the following statement: "1. The human recruiter is more likely to be a fair evaluator than the algorithmic recruiter: 1 - Strongly disagree to 5 - Strongly agree".

[16]For readers not aquainted with the Five-Factor model, a very brief and inconclusive summary of the five personality traits: Individuals who score high on extraversion are typically comparatively talkative, energetic and assertive. Agreeablenes is associated with being more trustful, good-natured and cooperative. Conscientiousness is associated with more orderly, dependable and self-constraint behavior. Higher levels of neuroticism are typically thought to correlate with being emotionally unstable and are more likely than average to be moody or to experience such feelings as fear, worry or anxiety. The fifth factor, openness, is associated to being intellectual, independent minded and imaginative John et al. (1999)

denotes the different outcome variables of which I will estimate the coefficients in the next sections. First, I use a simple equation to estimate the treatment effect:

$$Y_i = \beta_0 + \beta_1 Trt_i + \varepsilon_i \tag{2}$$

where Trt_i is an indicator of whether participant i received the treatment. In case that both groups are randomized and balanced along both observable and unobservable characteristics, equation (2) thus estimates the average treatment effect (ATE): The change in outcome variables which is due to the information-rich presentation of the algorithmic recruiter. The second type of regression incorporates the self-reported personal characteristics of participants and all interactions of the personal characteristics with the treatment. Specifically, I estimate:

$$\begin{aligned} Y_i = &\beta_0 + \beta_1 Trt_i + \beta_2 Male_i + \beta_4 Age_i + \beta_3 Educ_i + \beta_4 NonWhite_i \\ &+ \beta 5_O bserved_i + \beta_6 (Trt_i \times Male_i) + \beta_7 (Trt_i \times Age_i) + \beta_8 (Trt_i \times Educ_i) \\ &+ \beta_9 (Trt_i \times NonWhite_i) + \beta_{10} (Trt_i \times Observed_i) + \varepsilon_i \end{aligned} \tag{3}$$

This estimation thus shows how the information-rich presentation of the algorithmic recruiter affected participants differently, depending on their personal characteristics. To further refine the treatment effect and to see whether or not results from equation (2) are moderated by personality traits and risk aversion, I estimate:

$$\begin{aligned} Y_i = &\beta_0 + \beta_1 Trt_i + \beta_2 Male_i + \beta_4 Age_i + \beta_3 Educ_i + \beta_4 NonWhite_i \\ &+ \beta 5_O bserved_i + \beta_6 (Trt_i \times Male_i) + \beta_7 (Trt_i \times Age_i) + \beta_8 (Trt_i \times Educ_i) \\ &+ \beta_9 (Trt_i \times NonWhite_i) + \beta_{10} (Trt_i \times Observed_i) \\ &+ \beta_{11} Risk_i + \beta_{12} Open_i + \beta_{13} Consc_i + \beta_{14} Ext_i \\ &+ \beta_{15} Agr_i + \beta_{16} Neuro_i + \beta_{17} (Trt_i \times Risk_i) + \beta_{18} (Trt_i \times Open_i) \\ &+ \beta_{19} (Trt_i \times Consc_i) + \beta_{20} (Trt_i \times Ext_i) + \beta_{21} (Trt_i \times Agr_i) \\ &+ \beta_{22} (Trt_i \times Neuro_i) + \varepsilon_i \end{aligned} \tag{4}$$

where $Risk_i$ is the elicited risk aversion from participant i and $Open_i$, $Consc_i$, Ext_i, Agr_i, $Neuro_i$ are the self-reported personality traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism of participant i.

D. Results

DESCRIPTIVE STATISTICS

66 participants completed the baseline group of the experiment, and 82 participants completed the treatment group, where the functioning of the algorithmic recruiter and its

weights concerning gender and the time in the observed round is communicated.

Table 2—: Sample summary statistics

	Whole Sample		Baseline Group		Treatment Group			
	Mean	SD	Mean	SD	Mean	SD	t-statistic	p-value
Sample size	148		66		82			
% Males	0.47	(0.5)	0.47	(0.5)	0.48	(0.5)	-0.07	0.94
Age (category)	5.78	(2.31)	6.26	(2.59)	5.40	(2.00)	2.21	0.03
Education (category)	4.38	(1.42)	4.29	(1.39)	4.46	(1.45)	-0.75	0.46
Non-White	0.22	(0.42)	0.17	(0.38)	0.27	(0.45)	-1.50	0.13
Time in observed round	17.89	(13.34)	21.92	(17.69)	14.64	(6.94)	3.16	0.00
Average Time	16.56	(8.04)	18.66	(9.31)	14.87	(6.42)	2.81	0.01
Risk Aversion	2.94	(1.58)	3.14	(1.65)	2.78	(1.52)	1.35	0.18
Openness	3.77	(0.89)	3.85	(0.86)	3.70	(0.91)	1.02	0.31
Conscientiousness	4.06	(0.77)	4.00	(0.70)	4.11	(0.82)	-0.87	0.38
Extraversion	2.77	(1.11)	2.64	(1.10)	2.87	(1.12)	-1.23	0.22
Agreeableness	3.81	(0.77)	3.82	(0.77)	3.80	(0.79)	0.17	0.87
Neuroticism	2.74	(1.24)	2.72	(1.22)	2.76	(1.27)	-0.19	0.85

Table 2 shows the sample means of the participants' personal characteristics for the baseline, and treatment group as well as the corresponding two-tailed t-test statistics.[17]. The mean percentages of gender, non-White ethnicity, education level, risk aversion, and the Five-Factor personality traits show no significant difference for the treatment and baseline group. However, the participants' average age, their elapsed time in the observed round and their average time across all exercises show statistically significant differences. Participants in the treatment group were on average younger and performed better in the number-finding exercises, both in the observed round and on average. As the treatment involves communicating the algorithmic recruiter's weights on the observed round, this difference may be troublesome for achieving unbiased estimates of an average treatment effect (ATE). This point will be discussed in more detail in section V.A.

CHOICE OF RECRUITER

My aim in this section is to consider if the treatment changed the ratio of participants who chose the algorithmic/human recruiter and there exist self-selection affects along personal characteristics or personality traits. As the treatment communicated the algorithmic recruiter's weights on gender and the time in the observed round, one would expect that treated participants chose their recruiter partly based on these two variables. Table 3 reports regression estimates explaining the fraction of participants who chose

[17]Note that the variables of age, education, risk aversion, and the Five-Factor model personality traits are ordinally scaled. Thus, metric analysis such as calculating means imposes the assumption that the intervals between the values are of the same length and should be viewed with respective caution. For the detailed description of the measurements of each variable, see appendix A

the algorithmic recruiter. Column (1) shows that there is no significant treatment effect between the sample means of the baseline. In both groups, around 62 % of all participants chose the algorithmic recruiter. Column (2) shows regression estimates including the personal characteristics and their treatment interactions. Thus, it shows if participants with different personal characteristics overtly opted for either recruiter and if this changed due to the treatment.

The key result is that the treatment induces a self-selection effect along gender lines while the same does not hold true for the performance in the observed round. After seeing the detailed description of the algorithm and its gender-bias, male participants choose the algorithmic recruiter on average 34.1 percentage points more compared to females. In absence of the treatment, no significant difference between genders is observed. Second, there exists no observable effect of the treatment controlling for gender, age, education, ethnicity and the performance in the observed round. Third, in contrast to the self-selection along gender lines, there is no observable self-selection of participants along different performances in the observed round. These findings are robust for either a logistic (Table B2) or probit regression specification (Table B3).[18]

As regression estimates only show the relative difference in behavior, it is interesting to know whether the self-selection along gender is driven by females and/or males. Figure 2 shows the mean percentages for choosing the algorithmic/human recruiter for males and females in the baseline and treatment group. Whereas female (male) participants choose the algorithmic recruiter on average 16.7 percentage points less (more) in the baseline, the treatment exacerbates this difference to 55.1 percentage points. Further, it shows that the self-selection is driven by both genders. Treated females choose the algorithmic recruiter on average 17.1 percentages points less compared to females in the baseline while males choose respectively 21.3 percentage points more.

Participants also stated their personality traits along the Five-Factor model and risk aversion. Thus, I can test if psychological traits moderate the choice of algorithmic or human recruiter. Column (1) of Table B4 reports the regression estimates controlling for the standardized five personality traits and risk aversion. The resulting coefficients are remarkably different to Table 3. The previously found self-selection, treated males opting for the algorithmic recruiter and treated females for the human, is not observable. At the same time, out of the five factors and risk aversion, only conscientiousness significantly explains the choice of recruiters, indicating a moderating role. Having a one standard deviation higher score of conscientiousness increases the ratio of choosing the algorithmic recruiter on average by 21.3 percentage points, independent of treatment group. Similar results to the analysis in Table 3 are found for the coefficients of self-reported White and non-White participants. While being non-White correlates with an increased choice of the algorithmic recruiter in the baseline group, this reverses in the treatment group. These findings are robust to the results of a logistic regression specification.[19]

[18]On a ten percent significance level, participants who self-reported themselves as non-White have a preference for the human recruiter in the baseline compared to White participants (-30.4 percentage points) For the treatment group, this trend reverses to a slight preference for the algorithmic recruiter (+ 6 percentage points). Both logistic and probit specifications report this finding as well, albeit also on a significance level of ten percent.

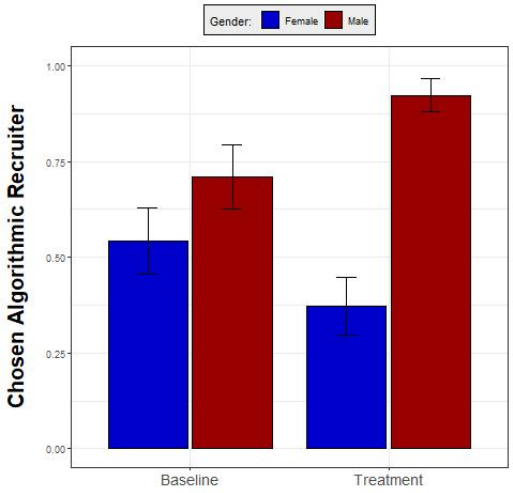
[19]Here, an increase of conscientiousness by one standard deviation implies 5.41 times higher likelihood of choosing

Table 3—: Explaining the choice between recruiters

	Ratio Of Choosing The Algorithmic Recruiter	
	(1)	(2)
Constant	0.621***	0.647*
	(0.060)	(0.283)
Treatment	0.013	-0.494
	(0.080)	(0.344)
Male		0.172
		(0.112)
Education		0.007
		(0.041)
Non-White		-0.304†
		(0.159)
Age		-0.005
		(0.026)
Observed round		-0.003
		(0.003)
Treatment × Gender		0.341*
		(0.143)
Treatment × Education		0.071
		(0.049)
Treatment × Age		-0.010
		(0.035)
Treatment × Non-White		0.361†
		(0.187)
Treatment× Observed round		-0.001
		(0.008)
Observations	148	148
R^2	0.0002	0.261

Note: This table reports OLS regressions where the dependent variable indicates whether participants chose the algorithmic recruiter (1) or the human recruiter (0). Column (1) estimates regression equation (2) and column (2) estimates regression equation (3). Robust standard errors are reported.
Significance indicators: † p<0.1; * p<0.05; ** p<0.01; ** p<0.001

Figure 2. : Mean differences in choosing the algorithmic recruiter



Mean difference across the baseline and treatment group and genders. Means from left to right: 54.3%, 71.0%, 37.2% and 92.3%. Error bars represent standard the error of the mean.

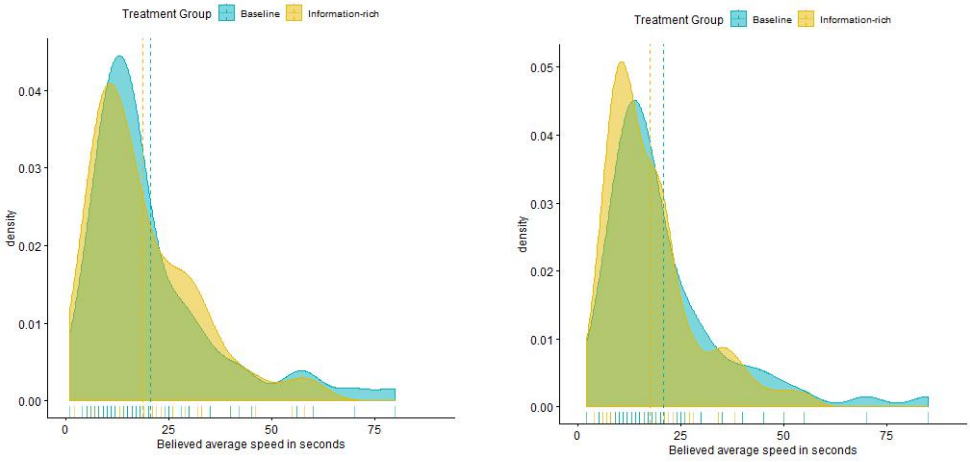
BELIEFS ABOUT RECRUITER PREDICTIONS

One explanation for the choice of recruitment method would be that participants believe to receive different predictions from them and thus choose the recruiter from which they think to receive the higher monetary outcome. The predictions participants believed to receive might thus be an important driver in explaining the choice of recruiter. This section analyzes to which extent participants in the baseline and treatment group differ in their believed predictions from both recruiters. Figures 3a and 3b show the distributions for the two beliefs along the neutral baseline and the information-rich treatment group. The beliefs across both groups are relatively similar, with a average difference of 3.35 seconds for the human and 1.76 seconds for the algorithmic recruiter. The results of a two tailed Kolmogorov-Smirnov test in Table B5 further provides no evidence for a significant difference in the distributions for both recruiters comparing across baseline and treatment group. There seems to however exist a difference between the baseline and treatment group comparing the beliefs about the recruiters directly to each other. This means simply taking the difference between how participants believed to be predicted from both recruitment methods.

These findings are confirmed by the regression estimates in Table 4, which show that the average believed prediction in the treatment group is not significantly different from the baseline group. (around 20.5 seconds). People do not think that either method will give them a significantly different score. Column (3) of Table 4 shows the difference

the algorithmic recruiter. The coefficients of gender and the gender-treatment interaction, which implies self-selection along gender, are not significant on a five percent level, but significant on a ten percent level.

Figure 3. : Distribution of believed predictions from both recruitment methods



(a) Believed predictions from the algorithmic recruiter

(b) Believed predictions from the human recruiter

between both beliefs, where a positive value thus means a better believed prediction for the algorithmic recruiter compared to the human recruiter and vice versa. As neither the constant nor the estimate of the coefficient are significant, there is no indication of an average treatment effect impacting belief-forming by just comparing both groups without taking into account differences in gender, education, age or their time in the observed round.

Table 4—: Average treatment effect on believed predictions

	Believed Prediction From Recruiters		
	Belief_{Human}	$\text{Belief}_{Algorithm}$	$\text{Belief}_{Difference}$
	(1)	(2)	(3)
Constant	20.803***	20.500***	−0.303
	(1.867)	(1.955)	(0.814)
Treatment	−3.352	−1.756	1.596
	(2.217)	(2.412)	(1.018)
Observations	148	148	148
R^2	0.017	0.004	0.017

Note: This Table reports OLS regressions where the dependent variable is the believed prediction from the human recruiter in column (1), the believed prediction from the algorithmic recruiter for column (2) and the difference between both beliefs for column (3). All regressions estimate regression equation (2) Robust standard errors are reported.
Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Next, I want to see if learning about the weights in the treatment induced different believes for different personal characteristics. For example, if males and females in the treatment group expected different predictions by the recruiters compared to the baseline group. Columns (1) and (2) of Table 5 estimate regression equation 3 for the believed predictions of the human and algorithmic recruiter and the relative difference between them. Here, the results are two-fold: Participants seem to base their believed prediction on their performance in the observed round. The beliefs are significantly explained only by the time in the observed round, irrespective of baseline or treatment group. For every additional second in the observed round, the believed prediction from both recruiters increase by 0.67 to 0.62 seconds.[20] Further, the coefficients for the treatment, gender, or gender-treatment interaction do not significantly explain the believed prediction from both recruitment method. Column (3) depicts the same estimated coefficients but for the difference between both believed predictions. Directly comparing the beliefs from both recruiters shows a a difference along gender: Treated males believed to receive a 4.11 seconds lower and thus better prediction from the algorithmic recruiter compared to treated females. Thus, while the specific level of the believed predictions of both recruiters is mostly informed by their time in the observed round, the treatment affects how participants expect both recruiters will predict them relatively.

Column (2) of Table B4 reports estimates including the Five-Factor model and risk aversion. This allows to see if also the expected scores are mediated by, for example, conscientiousness. Controlling for the five personality traits and risk aversion, being male is linked to having a better belief about the algorithmic than the human recruiter in both groups. Concerning the five personality factors, participants who score higher on conscientiousness and neuroticism in the baseline group expect to receive a lower prediction from the algorithm compared to the human recruiter. Within the treatment, the effect of conscientiousness is reversed, with higher levels of conscientiousness implying a lower, thus better, better believed prediction from the human recruiter compared to the algorithmic recruiter. These results are partly contradictory to the analysis of the recruitment choice in column (1) of table B4, where a higher level of conscientiousness implies a higher ratio choosing the algorithmic recruiter.

Belief Confidence and Accuracy

In this section, I will assess whether perceiving more information about the algorithmic recruiter leads people to better understand its prediction. Yeomans et al. (2017) showed that one reason why people are averse to algorithmic recommendations is an insufficient understanding of these. Hence, do treated participants feel more confident in their believed prediction from the algorithmic recruiter and is their belief closer to the actual prediction? Table 6 shows that participants do not feel more confident in stating their believed predictions about either recruiter after learning about the algorithmic recruiter's functioning. This finding is confirmed by controlling for the personal characteristics of participants in Table B6).

[20] As the average time in the observed round was 17.88 seconds this implies also most significant effect in magnitude

Table 5—: Treatment effects on believed predictions including personal characteristics

	Believed Predictions From Recruiters		
	Belief_{Human}	$\text{Belief}_{Algorithm}$	$\text{Belief}_{Difference}$
	(1)	(2)	(3)
Constant	−1.914	6.843	8.757
	(6.127)	(8.888)	(5.459)
Treatment (Treat)	3.873	−2.375	−6.248
	(8.108)	(10.550)	(5.903)
Gender	−1.413	−2.334	−0.921
	(2.413)	(2.647)	(1.314)
Age	0.459	0.415	−0.044
	(0.477)	(0.514)	(0.295)
Education	1.311	−0.295	−1.606†
	(1.146)	(1.499)	(0.868)
Non-White	0.161	−0.936	−1.097
	(4.250)	(4.634)	(1.747)
Observed round	0.678***	0.620***	−0.058
	(0.108)	(0.126)	(0.052)
Treatment ×Gender	−0.971	−5.082	−4.111*
	(3.022)	(3.389)	(1.706)
Treatment × Education	−1.351	0.213	1.564
	(1.418)	(1.807)	(0.994)
Treatment ×Age	0.864	1.456	0.592
	(0.905)	(1.047)	(0.432)
Treatment × Non-White	−4.049	−3.063	0.986
	(4.609)	(5.129)	(2.064)
Treatment ×Observed round	0.053	0.005	−0.048
	(0.231)	(0.262)	(0.093)
Observations	148	148	148
R^2	0.554	0.479	0.189

Note: This table reports OLS regressions where the dependent variable is the believed prediction from the human recruiter for column (1), the believed prediction from the algorithmic recruiter for column (2) and the difference between both beliefs for column (3). All regressions estimate coefficients for regression equation 3 Robust standard errors are reported.
Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table 6—: Average treatment effect on the confidence of beliefs

	Confidence of the believed predictions		
	Confidence_{Alg}	Confidence_{Hum}	Confidence_{Diff}
	(1)	(2)	(3)
Constant	71.758***	69.212***	2.545
	(3.128)	(3.162)	(1.629)
Treatment	−4.660	−2.785	−1.875
	(4.119)	(4.143)	(2.393)
Observations	148	148	148
R^2	0.009	0.003	0.004

Note: This table reports the estimated coefficients for regression equation 2. The dependent variables are: (1) and (2) are the stated confidence in the believed predicted average speed of the algorithmic recruiter and the human recruiter. (3) is the difference between both, where a negative value indicates a less confident belief about the algorithmic recruiter compared to the human recruiter. Robust standard errors are reported. Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

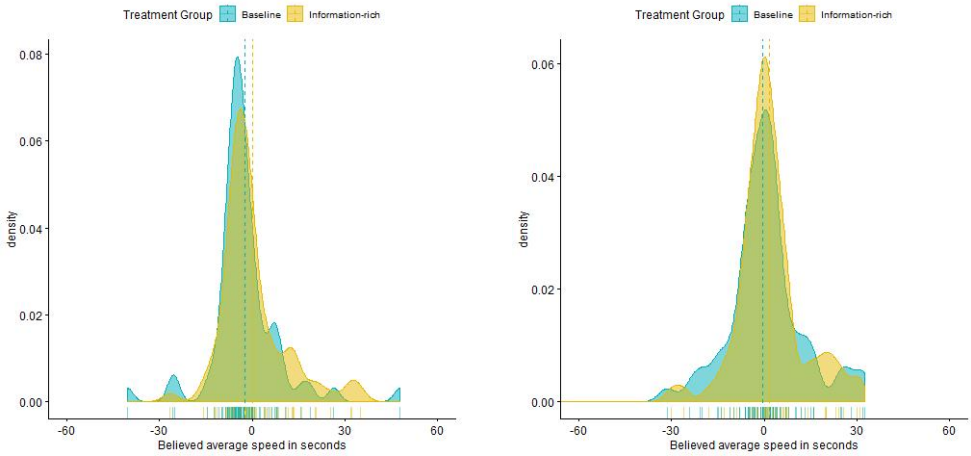
Algorithmic predictions have been shown to be more accurate than their human counterparts (Mullainathan and Spiess, 2017; Yeomans et al., 2017). My experimental design enables a brief analysis whether knowing more about the algorithmic recruiter leads participants to expect its prediction more accurately. To see how accurately participants formed their beliefs and if the accuracy is affected by the treatment, I take the difference between the stated beliefs of participants and the actual predictions made by both recruitment methods. As the treatment communicated the functioning and two weights of the algorithmic recruiter, one may expect that the treated participants form their beliefs more accurately. However, figures 4b and 4a show that the distributions of the accuracy for both recruiters center around zero with no clear deviations concerning treatment. A two-sample Kolmogorov-Smirnov test (see Table B7) further does not reject the null hypothesis that the accuracy of beliefs come from the same distribution. Further, Column (1) and (2) of Table B8 show that there is no significant difference in accuracy between recruitment methods and no observable treatment effect.

Rationality of Participants Behavior

This section analyzes if the treatment leads participants to choose the payoff-maximizing recruitment method or had an effect on how participants acted upon their believed predictions. If participants were maximizing their expected payoff, believing to receive a lower prediction from either the algorithmic or the human recruiter should lead participants to choose this method. Indeed, if the recruiter's prediction and thus the resulting

for the believed predictions

Figure 4. : Distribution of the accuracy of believed predictions



(a) Accuracy for the algorithmic recruiter

(b) Accuracy for the human recruiter

payoff would be the only relevant category, then the choice of participants should only be related to a dichotomous variable indicating which recruiter is believed to predict the lower score. The level of the beliefs itself should then be irrelevant.

Table 7 reports an estimate of

$$(5) \quad Alg_i = \beta_0 + \beta_1 D_{Belief_i} + \beta_2 Belief_{Hum_i} + \beta_3 Belief_{Alg_i} + \beta_4 Trt_i + \beta_5 Trt_i \times D_{Belief_i} + \beta_6 Trt_i \times Belief_{Alg_i} + \beta_7 Trt_i \times Belief_{Hum_i} + \varepsilon_i$$

where Alg_i indicates if participant i chose the algorithmic recruiter, D_{Belief_i} indicates if participant i thought that the algorithmic recruiter would assign her a better score[21] and $Belief_{Hum_i}$ and $Belief_{Alg_i}$ are the respective average times participant i believed the human and algorithmic recruiter would predict for her. Table 7 shows that, contrary to pure payoff-maximizing behavior, the choice of recruiter is not guided by the higher believed prediction but by their levels. A one-second increase in the believed prediction of the algorithmic (human) recruiter correlates with a 2 percentage point higher frequency to choose the algorithmic (human) recruiter. The treatment does not affect this relationship.[22] These results are robust for a logistic regression specification, see Table B9.

To what extent do participants choose the recruitment method which yields the highest

[21] There are no instances within the data when the human and algorithmic prediction are identical.

[22] On a significance level of ten percent, treated males are more likely to act upon the better believed recruiter compared to treated females. As the corresponding t-statistic (t = 1.84) is close to the critical value of 1.96, the effect can be seen at least as suggestive, as the study may be underpowered to show all significant treatment effects, see section V.B for a full discussion of the matter

Table 7—: Relation between choice of recruitment method and believed predictions

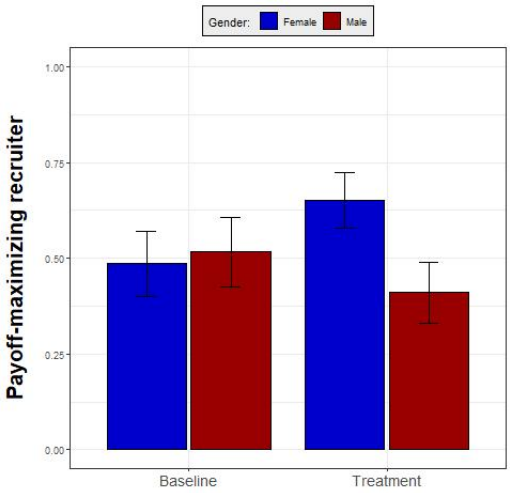
	Chosen Recruitment Method
	Algorithmic Recruiter
Constant	0.803^{***}
	(0.164)
D-Belief	-0.190
	(0.13)
Belief_{Human}	0.020^{*}
	(0.0081)
$\text{Belief}_{Algorithm}$	-0.024^{**}
	(0.0075)
Treatment (Trt)	−0.125
	(0.217)
Treatment × D-Belief	$0.330^{\dagger}$
	(0.1775)
Treatment × $\text{Belief}_{Algorithm}$	−0.006
	(0.0185)
Treatment × Belief_{Human}	0.006
	(0.0217)
Observations	148
R^2	0.166
Adjusted R^2	0.124
Residual Std. Error	0.454 (df = 140)
F Statistic	3.979^{***} (df = 7; 140)
Note:	*p<0.1; **p<0.05; ***p<0.01

Note: This table reports OLS regressions where the dependent variable is an indicator for the participant to choose the algorithmic recruiter (being 1) or the human recruiter (being 0). D-Belief indicates that the participants believed to receiver a higher prediction from the algorithm than from the human recruiter. Robust standard errors are reported.
Significance indicators: † p<0.1; * p<0.05; ** p<0.01; ** p<0.001

payoff or choose consistent with their believed predictions? Transparent information about algorithms are often advocated as enabling people to better decide which prediction to use (Dwork et al., 2012). Based on my experimental design, I can calculate the ratio of participants who chose the recruiter yielding them the higher payoff by comparing the actual predictions made by both recruiters. Further, a participant is defined to act consistent with one's belief if she chooses the recruitment method which she believed to assign the better prediction. Column (1) of Table 8 shows that 50 % of participants choose the payoff maximizing recruiter in the baseline, with no significant difference between treatment and baseline group. Column (2) show that for acting consistent with one's belief, the treatment increases the percentage of participants who chose the higher-believed recruiter from 47% to 64.7 %.

Do both indicators also exhibit differences in treated genders, echoing the self-selection of participants? Figures 5 and 6 show the mean ratios for choosing the payoff maximizing recruiter and acting consistent with one's belief. The treatment predominantly affected how females act upon their beliefs while showing smaller effects for males. Without taking into account further controls, the graphs show that treated females increasingly choose the recruiter giving them a higher payoff (+16.5 percentage points) and also show a higher ratio of choosing the recruiter believed to predict the better prediction (+18 percentage points) compared to females who didn't receive the information-rich description. Treated males show a small average decline for the payoff-maximizing recruiter (-10.5 percentage points) and slight average increase in choosing the recruiter believed to assign the better prediction.

Figure 5. : Mean Differences in choosing the payoff-maximising recruiter



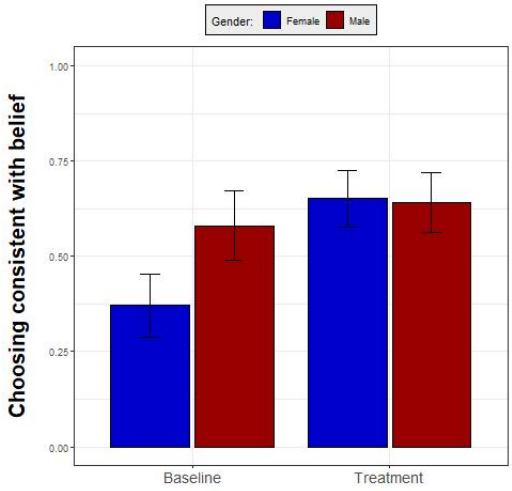
Choosing the payoff maximizing recruiter: Mean differences across the Treatment/ Baseline group and gender. Means from left to right: 48.6 %, 51.6 %, 65.1 % and 41 % (error bars represent standard error of the mean)

Table 8—: Choosing the payoff-maximising recruiter and acting belief-consistent

	Ratio of rational behavior			
	Payoff-max	Belief-consistent	Payoff-max	Belief-consistent
	(1)	(2)	(3)	(4)
Constant	0.5***	0.47***	0.394	0.555*
	(0.062)	(0.062)	(0.321)	(0.289)
Treatment	0.037	0.177*	0.742*	0.583
	(0.084)	(0.082)	(0.392)	(0.381)
Gender			0.020	0.231†
			(0.124)	(0.121)
Age			0.010	−0.049*
			(0.023)	(0.024)
Education			−0.004	0.013
			(0.048)	(0.044)
NonWhite			0.284†	−0.289†
			(0.170)	(0.170)
Observed round			0.0003	0.005
			(0.004)	(0.004)
Treatment × Gender			−0.278†	−0.326†
			(0.163)	(0.163)
Treatment × Age			−0.025	0.074*
			(0.032)	(0.036)
Treatment ×Treatment			−0.057	−0.053
			(0.059)	(0.058)
Treatment ×Non-White			−0.456*	0.292
			(0.204)	(0.207)
Treatment × Observed round:			−0.006	−0.032***
			(0.009)	(0.009)
Observations	148	148	148	148
R^2	0.034	0.051	0.091	0.165
Adjusted R^2	0.014	0.032	0.017	0.097

Note: This table reports OLS regressions where the dependent variable is in column (1) and (3) an indicator for the participant to choose the payoff-maximizing recruitment method (1 if yes, 0 if not). Column (2) and (4) estimates the treatment and gender effects for a dependent variable, indicating if participants chose the recruitment method, which they believed to be superior (1 if yes, 0 if not). Robust standard errors are reported. Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Figure 6. : Mean Differences in acting belief-consistent



Choosing the recruiter consistent with one's belief: Mean differences across the Treatment/ Baseline group and gender. Means from left to right: 37.1 %, 58.0 %, 65.1 % and 64.1 % (error bars represent standard error of the mean)

Columns (3) and (4) of table 8 then control for personal characteristics and their treatment interactions. Choosing the payoff maximizing recruiter is estimated to increase significantly by 74.2 percentage points when the functioning of the algorithm and its weights is explained. Second, this effect is considerably lowered for participants who identify themselves as non-White. Similarly controlling for belief-consistency reveals that the positive treatment effect of column (2) is moderated by age and the treatment interaction with the time in observed round. Older participants in the baseline group show a lower ratio of belief-consistency compared to younger ones. Within the treatment group, this trend reverses as being older correlates with a higher ratio of choosing the better believed recruiter.[23] More consistent with intuition, a one second increase in the observed round leads treated participants to a 3.2. percentage points lower ratio of acting consistent with one's belief. In other words, a worse observable performance correlates with not acting on one's believed predictions. However, little robust evidence for the findings in (3) and (4) is found when controlling for personality traits and risk aversion, see table B10.[24]

[23]The significant treatment-interaction of age is surprising, as the treatment does not reveal any information about how the algorithmic and human recruiter handles age or ethnicity. Section V.A discusses this finding in more detail.

[24]Robust evidence is found for the interaction between performance in the observed round and acting consistent with ones' belief. On the other hand, this specification shows that participants who for example identify their ethnicity as non-White are 47.4 percentage points more likely to choose the payoff maximizing recruiter compared to participants who identify themselves as White in the baseline. This reverses for treated participants, as then non-White participants are 19 percentage points less likely to be payoff-maximizing.

QUALITATIVE PERCEPTIONS ABOUT RECRUITER CHARACTERISTICS

The previous results suggest that participants do not only act upon (believed) monetary outcomes and raise the question which other factors explain how participants choose between the two recruitment methods. Previous literature shows that the usage or choice of algorithmic decision systems is as well influences by qualitative perceptions. For example, Yeomans et al. (2017) find that the acceptance of algorithmic recommendations is increased if participants feel more familiar with it or think that they can understand it.

One possible explanation of the observed self-selection apart from monetary outcomes is thus that treated females chose the human recruiter not only because they expect to receive a smaller payoff from the algorithm, but also because they dislike the fact that they are being discriminated against. As participants also submitted ten qualitative questions on how they perceived both recruitment methods relative to each other, I can analyze this along several qualitative dimensions. Table 9 shows that treated females perceived the algorithm on average as less fair, more likely to discriminate, less familiar, more likely to be prone to error and less likely to care about their performance compared to treated males. A noteworthy exception is the consistency of how transparent both genders perceived the recruiters. An inspection for the same qualitative beliefs between treatment and control group in Table B11 reveals that the treatment did on average not lead to a higher perceived transparency.

Table 9—: Comparing qualitative beliefs across gender in the treatment group

	Female	Male	t-statistic	p-value
Fair	2.84	3.64	-3.20	0.00
Transparent	3.79	3.90	-0.45	0.65
Simpler	3.49	3.74	-1.00	0.32
Discriminative	3.12	1.85	4.58	0.00
Familiar	2.23	3.00	-3.05	0.00
Care about Characteristics	2.12	2.59	-1.92	0.06
Prone to Error	2.91	2.36	2.25	0.03
Care about Performance	2.67	3.33	-2.57	0.01
Decide Quickly	4.47	4.18	1.86	0.07
Other	2.21	2.15	0.23	0.82

Note: This table reports the sample means and two-tailed t-test for the ten qualitative beliefs for male and female participants in the treatment group. Specifically, participants were asked: "The algorithm is more likely to be [] than the human recruiter" where [] stands for one of the ten items. The answers are recorded on a Likert-scale ranging from 1 (totally disagree) to 7 (totally agree).

To what extent can these differences in the gender's perception be viewed as causal? Table B12 depicts the same means and two-tailed t statistic for the baseline group. Here, no item shows a statistically significant difference, indicating that the described gender differences in Table 9 result from learning that the algorithm assigns better scores to men and worse to women. However, as the treatment and control group are not balanced across all characteristics and had to be sampled at two different dates, the causal interpre-

tation of a difference in means as an average treatment effect (ATE) needs more caution (Angrist and Pischke, 2008).

Controlling for the personal characteristics of participants, Table 10 shows OLS regression estimates for the three standardized qualitative beliefs thought to be most affected by the treatment: Perceived discrimination, fairness and transparency. The results corroborate two out of the three found differences in means. Treated females find the algorithmic recruiter comparatively around one standard deviation more discriminatory than treated males.[25] Further, Tables B13 and B14 show that this gender difference in perceived discrimination is solely driven by females: While males do not show any differences for both fairness and perceived discrimination, females in the treatment feel significantly more discriminated. Column (3) of Table 10 further confirms that the treatment had no effect on the perceived transparency of the algorithmic recruiter compared to the human recruiter, even though half of the information-rich representation of the treatment focused on explaining the functioning of the algorithm.

The question arises whether beliefs about monetary outcomes and qualitative beliefs about participant's perceived fairness and discrimination are two sides of the same medal. Do treated participants who perceive the algorithmic recruiter in a less favorable light than the human recruiter (more discriminatory, less fair, less caring about their other performance and more prone to error) also expect a lower prediction from the algorithmic recruiter and are thus less likely to choose this recruitment method? The Anova/F-test in Table B15 tests if the inclusion of additional qualitative beliefs helps in explaining the variance of the choice of recruitment method. Including the qualitative beliefs of fairness and discrimination within the regression of Table 7 significantly improves the fit. All further beliefs do not significantly improve the fit of the model. Further evidence that the qualitative beliefs are not interchangeable with the beliefs about the assigned scores can be found in Table B16. In a regression of the incentivized beliefs on the qualitative beliefs fairness, discrimination and transparency, no estimated coefficient is statistically significant. Beliefs about the recruiters' predictions and perceived qualitative beliefs about both recruiters seem to explain two different drivers in choosing the recruiter.

[25] On a ten percent significant level, treated females find the algorithmic recruiter also less fair, around half a standard deviation. Other observed effects are that treated participants receive the algorithmic recruiter as less disriminatory compared to the human recruiter for each higher level of achieved education. Considering fairness, older participants are shown to perceive the algorithm as less fair, all other variables held equal.

Table 10—: Treatment effects on qualitative beliefs including personal characteristics

	The algorithmic recruiter is more likely to ... than the human recruiter		
	Discriminate	Be fair	Be Transparent
	(1)	(2)	(3)
Constant	−0.789	0.925	1.224**
	(0.527)	(0.632)	(0.470)
Gender	0.102	0.070	−0.303
	(0.187)	(0.190)	(0.215)
Age	−0.011	−0.083*	−0.115**
	(0.039)	(0.040)	(0.039)
Education	0.097	−0.034	−0.090
	(0.075)	(0.092)	(0.082)
NonWhite	0.254	0.055	0.091
	(0.357)	(0.291)	(0.279)
Observed round	0.003	−0.002	0.001
	(0.007)	(0.011)	(0.007)
Treatment	1.619*	−0.953	−1.117
	(0.747)	(0.830)	(0.751)
Treatment × Gender	−0.924***	0.561†	0.396
	(0.299)	(0.300)	(0.327)
Treatment × Age	0.054	−0.076	0.043
	(0.072)	(0.064)	(0.078)
Treatment × Education	−0.278**	0.112	0.127
	(0.098)	(0.116)	(0.121)
Treatment × NonWhite	−0.334	−0.038	−0.282
	(0.433)	(0.405)	(0.420)
Treatment × Observed round	0.023	0.004	0.008
	(0.018)	(0.019)	(0.023)
Observations	148	148	148
R^2	0.274	0.208	0.068
Adjusted R^2	0.215	0.144	−0.007

Note: This table reports OLS regressions where the dependent variable is a ordinal variable displaying how the participants thought about both recruitment methods in three dimensions and range from 1 (strongly disagree) to 7 (strongly agree). Column (1) depicts fairness, column (2) depicts transparency and column (3) depicts discrimination. All regressions estimate coefficients for regression equation 3. Robust standard errors are reported.
Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

V. Discussion

A. General Discussion

The key results of the experiment are: (1) Different degrees of transparency with which an algorithm is presented, lead to stark differences in behavior. Learning about the algorithmic recruiter discriminating against females while favoring males induces large self-selection effects from both genders. The communicated weight of the once observable performance leads, however, to no significant self-selection. (2) Underlying beliefs and perceptions about a recruitment method matter. The choice of recruitment method is partly explained by the predictions participants believe to receive, while the treatment leads participants to put more emphasis on which recruiter is believed to predict them better. Second, the self-selection in (1) is as well driven by female participants perceiving the algorithmic recruiter as more discriminatory after learning about its weights. This effect is not found for males, who perceive the algorithmic recruiter as equally fair and discriminatory in both groups. (3) Contrary to the results of Yeomans et al. (2017), the explanation of the algorithmic recruiter and communicating two of its weights does not significantly increase its perceived transparency or familiarity. There is also no compelling evidence that the treatment led to a better understanding of the algorithmic recruiter, more accurate beliefs, higher confidence or a higher ratio of choosing the payoff-maximizing recruiter.

Explaining (1) is fairly straightforward. If males and females anticipated a gender-neutral algorithm, then adapting one's behavior according to the new information and choosing the recruiter which assigns a better prediction is the rational thing to do. This reasoning might also explain why participants did not react strongly to the information that the algorithmic recruiter predicts a higher average time for a higher time in the observed round: They already anticipated that the algorithm would give much of its weight to the observed round and thus did not update their behavior to this information. Evident for this interpretation is that for every additional second in the observed round, participants in both the treatment and control group believe to receive around 0.65 seconds higher predictions by the algorithmic recruiters, irrespective of their treatment group.

Concerning (2), the level of believed predictions robustly explains a part of the choice of recruitment. The finding that participants base their choice mostly on the levels of beliefs and only in about half the cases on the relatively better-believed recruiter is consistent with previous findings about the interplay of choice and monetary beliefs, see for example (Costa-Gomes and Weizsäcker, 2008). At the same time, I show that qualitative perceptions about fairness and justice were also affected by the communicated weights and explain the choice of recruitment method. Women who learn that the algorithmic recruiter assigns them a comparatively lower prediction find the same method less fair and more discriminatory. Given the evidence, it further seems that both types of beliefs are complementary, echoing research comparing economic and psychological preferences Becker et al. (2012). I can not fully explain why males' perceived fairness and discrimination stayed constant. One explanation is that only the group that is adversely affected by algorithmic discrimination perceives it as negative, while the group that is positively

affected does not perceive it as more or less fair. If so, the positive self-selection of males would best explainable by treated males expecting a higher monetary payoff due to knowing about their positive discrimination, while the negative self-selection of females' behavior is driven by both moral and monetary considerations.

Interpreting finding (3) is more difficult. While I do not replicate the findings of Yeomans et al. (2017), my experimental treatment considerably differs from theirs. First, the participants in my design are *evaluated* by both recruitment methods, while the participants in Yeomans et al. (2017) decide between *using* the advice from either system. One explanation is thus that people act and perceive systems differently depending on if they *use* these systems or are *evaluated* by them. Second, compared to their design, my treatment also includes information about the weights of the algorithmic recruiter, communicating its gender-biased weight. One could well argue that this inclusion makes my treatment more transparent than the one used by Yeomans et al. (2017), whose text does not say anything about the weights of the algorithm used. However, both male and females do not perceive the treatment of this research as more transparent. Further, even though the treatment was designed to communicate that the algorithmic recruiter also weighs positively on the performance in the observed round, treated females perceived that the algorithmic recruiter cared less about their performance. One possible explanation for this is participants selectively discounting communicated weights of the algorithmic recruiter based on fairness considerations. While neutral information about an algorithmic prediction, as in Yeomans et al. (2017), is perceived as more transparent, information that includes discriminatory or biased weights do not increase perceived transparency or might even lower it. In short, a described weight of an algorithm might not be perceived as increasingly transparent, if its content is discriminatory. Future experimental studies should disentangle neutral and biased information to confirm this hypothesis. Further, more information about an algorithmic decision system does not necessarily lead to better choices of the people being judged by them. These findings are important in the ongoing discussion about algorithmic discrimination and transparency (Datta et al., 2016).

My results suggest further that the personality trait of conscientiousness might play a role in moderating the choice of human or algorithmic recruiter. Table B4 shows that more conscientious participants are more likely to choose the algorithmic recruiter and believe to receive a better prediction from it, irrespective of treatment group. The inclusion of the Five-Factor model also rendered the previous significant gender-treatment interaction insignificant. Within the psychological literature, conscientiousness refers to differences in determination, impulse control, desire to achieve and self-organization (Costa and MacCrae, 1992). Out of the Five-Factor personality traits, conscientiousness also has the strongest and most robust relationship with job performance (Hogan and Ones, 1997). One explanation of its moderating effect on choosing the human or algorithmic recruiter is thus that more conscientious people also adapt more to information about how a recruiter treats for example gender and react accordingly. The only paper of which I am aware of looking at the interplay between conscientiousness and technology, Devaraj et al. (2008), finds that conscientiousness moderates the relationship between

perceived usefulness and intention to use technology and the relationship between subjective norms and the intention to use the technology. In the context of this experiment, a more conscientious individual would thus react on both the monetary outcome and the qualitative beliefs.

B. *Internal Validity*

Conducting an experiment on Amazon's Mechanical Turk comes with some innate disadvantages concerning the internal validity of the experiment. I do not know to what extent participants paid close attention to their choices, or whether they were in the midst of other activities while filling out the online experiment.[26] In general, it seems that the participants in the main experiment well understood the experimental design and consequences of their decisions. All 148 participants had at least two correct answers in each set of understanding checks for either the recruitment choice or belief elicitation. Second, as an additional measure of attention, only one out of the 149 participants who completed all stages of the experiment entered nonsensical answers for the open text questions asking the reasons behind choice of recruiter and belief-elicitation. Most other answers gave comprehensible explanations of their choices centering around innate preferences for the algorithmic/human recruiter, efficiency, monetary outcomes and discrimination.[27]

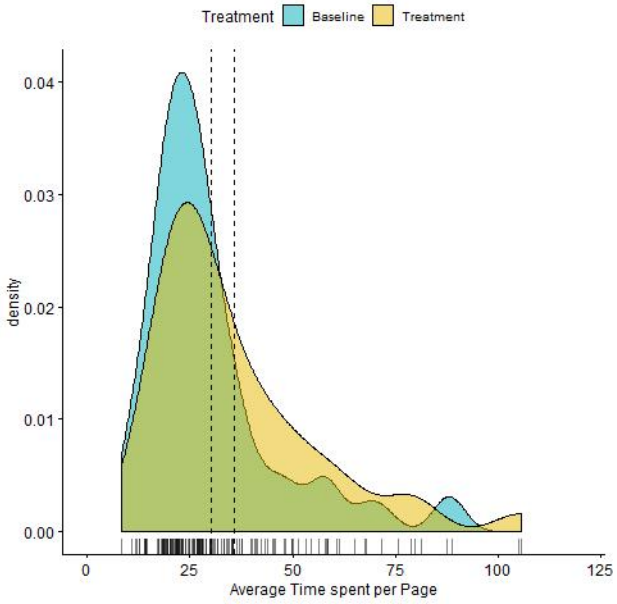
On the other hand, through data provided by MTurk, I observe that some participants spend either very little or a lot of time on specific pages. 58 (39) participants in the treatment (baseline) spend only one or two seconds on some pages of the experiment. Further, 9 (5) participants in the treatment (baseline) took longer than 20 minutes on some pages. Comparing the groups, Figure 7 shows that the distribution of the average time spent per page is shifted more to the right for the treatment. Table B17 shows that this also holds true for important decisions within the experiment: Participants in the treatment spent less time reading the task instructions and more time on decision of recruitment choice and belief-elicitation. This may partly due to the fact that participants had to read more text due to the detailed description of the algorithmic recruiter.

These differences between the two treatment groups link to a more general problem. Table 2 showed that there exist further significant differences between the baseline and treatment group. Treated participants are comparatively older and perform considerably better in the task. The latter is profound: Treated participants needed on average around seven seconds less to solve the observed round, roughly a third of the average time of the baseline group. Here it is to note again that due to logistical problems, the two groups had to be recruited separately and that the recruitment happened around three weeks apart. This poses two problems: First, it might imply that treatment and control group are not similar on observed and unobserved characteristics except the treatment which then might bias all treatment effects. As all variables included in the experiment other

[26] Also in laboratory experiments, it can not be known for sure to which degree participants think about each choice or action. While their allotted time can be controlled more closely, the mental attention/effort that participants exert within a lab stays private, for smaller samples using neurological techniques such as fMRI's, see Camerer et al. (2005)

[27] The specific answer which was excluded was "na". A sheet with all answers to the open ended-questions asking for the participants' reasoning behind their choice of recruitment method and belief can be found here

Figure 7. : Average time participants spent per page of the experiment



than the ones in table 2 are potentially affected by the treatment, I can not rule out this problem with the data at hand. Thus future replicating or similar studies which show a more balanced recruitment process between both groups are needed to make the reported findings robust.

To what extent are the found effects likely to be biased due to the differences in observed task performance? To answer this question, I assume that the treatment groups are balanced across unobservables so that only the differences in age and performance remain. First, both age and the observed time were shown to not possess significant treatment interactions with the choice of recruiter when controlling for all other balanced variables. The most probable result affected by the unequal distribution of performance is belief formation. As the beliefs for both recruiters correlates very similarly with the observed performance (0.678 and 0.62) Table 5, this would mean that participants in the treatment group had on average better beliefs on both recruiters' predictions than the baseline group, even if no treatment would have occurred. In my understanding this should not exert great bias to the overall results, as both recruitment methods would be equally affected. Further, the results of the qualitative beliefs seem not to be dependent on changes in the believed predictions of both recruiters, see Table 10.

One puzzle which remains is the significant interaction of the treatment with variables which were not affected by the treatment. Table 3 and B3 show that the treatment affects the recruitment choice of participants depending if they identify as White. Further, Table 8 and B10 show that the treatment has different effects given how old participants are. However, the treatment only adds information on how the algorithm treats gender and

performance in the observed round. While I can not rule out that these effects are induced by the treatment itself, another more likely explanation is differences between the baseline and the treatment group which then lead to different behavior. Participants in the treatment group are on average 4.1 years younger than in the baseline group. Second, between both groups exists a difference in the ratio of self reported non-Whites of about 10 percentage points. Further, the small sub-samples of ethnicity[28] are likely not high enough to balance out random or erratic behavior.

Last, the experiment is likely to not have picked up all significant treatment effects due to its relatively low sample size. The used sample size of 150 participants is mostly due to this paper being written in the scope of my master thesis and resulting logistical and financial constraints. For expected smaller effect sizes,[29] power tests performed before the experiment suggested a total sample size of between 300 and 900 participants, which is comparable to existing behavioral studies on the choice between algorithmic and human decision systems (Yeomans et al., 2017; Dietvorst et al., 2015; Logg et al., 2019). Areas which would benefit most from a higher sample are in my opinion the disentanglement of monetary and qualitative beliefs and being able to assess treatment effects within genders more closely. Concerning the latter, larger sample sizes would allow to test if women who select the algorithmic recruiter despite the communicated gender bias 'act male', that is, if such females are comparatively more similar to males in personal characteristics and personality traits.[30] Such behavior has for example been shown for people of color 'acting White' in the contexts of schooling and employment (Austen-Smith and Fryer Jr, 2005; Fryer Jr and Torelli, 2010).

C. *External Validity*

To what extent are the reported findings transferable to real labor market scenarios? Most importantly, the participants within this experiment did not apply for a real job, but were aware of being in an experiment. This reveals obvious differences regarding the aspirations and stakes to a real job interview. As a small defense, it is to note that people on Amazon's Mechanical Turk do so to earn a (small) part of their income and that performing well within the experiment is also incentivized through the quality ratings within MTurk.[31] As an online labor market, MTurk thus is more similar to a 'real' job market than a classic laboratory experiment.[32] Another obvious difference to a job interview is the role of the human recruiter. In this resesarch, the human recruiter is mostly a neutral counterfactual to the differently described algorithmic recruiter. In a real hiring setting, job applicants can often interact with the human recruiter, making

[28]Within the baseline group, there are 11 participants who don't report themselves as White. In the treatment group this increases to 22

[29]These include the effects on accuracy and confidence, as well as the difference in believed precidictions

[30]This analysis was not feasible within this paper, as the number of treated females who chose the algorithmic recruiter was only 16. Coupled with a rather low variance in personality traits, this is not enough for a meaningful analysis of personality differences.

[31]Many higher income tasks on MTurk require quality ratings from previous tasks of around 95 % (Horton et al., 2011)

[32]The experimental design is neither a classic laboratory study or a field study, as participants on MTurk were aware of being in an experiment declassifying it as a field-study. Following the taxonomy of Charness et al. (2013) it can be described as an "extra-lab experiment",

use of private information or interpersonal skills. Within the experiment, having the ability to interact with a real human recruiter might affect the participants' choice and the corresponding belief about the assigned prediction. While it would be possible to change the experimental design into a 2x2 setting where job applicants can interact live with the human recruiter, I see this as a fruitful extension for further studies on this topic.

Comparing the sample of MTurk participants of this experiment to the U.S. population (Current Population Survey (CPS), 2018), the former is on average younger, has a higher ratio of achieving/achieved undergraduate degree and has an equal percentage of Whites and people of other ethnicities. The sample recruited on MTurk is still much more representative to the overall working population than standard student samples. Also along characteristics not elicited in this experiment, Paolacci et al. (2010) find that the participant pool on MTurk is "at least as representative of the U.S. population as traditional subject pools".

While field studies have the advantage of real stakes and situations, the experimental approach within this paper allows to ask detailed questions about the participants' beliefs, confidence and personality. If the main results of this study, can be replicated by field studies within real labor market and job applications is a question that future studies have to answer. The presented findings are however a useful starting ground to discuss the far-reaching implications of the increased use of algorithmic screening - and hiring technologies, especially when the training data or constructed algorithms have adverse effects.

VI. Conclusion

In this paper, I analyze the decision process of people being evaluated by either a human or algorithmic prediction, in the specific context of job applicants. I introduce a novel experimental approach, which consists of setting up a real algorithmic and human recruiter and where participants consequently act upon different degrees of information and state their underlying expected predictions. This allows me to study how job applicants adapt their behavior when learning about the functioning of the algorithmic recruiter, and to see which beliefs drive their decisions. My results show that job applicants have to be treated as rational players: They self-select themselves along gender lines after learning about the algorithmic recruiter discriminating women. The second piece of information, how the algorithm weighs in on performance does, however, not seem to entice similar behavior. Participants select a recruitment method depending on monetary expectations and perceptions of justice. In particular, they seem to discount the transparency of communicated weights based on how discriminatory they perceive them.

My results suggest important implications for the design of algorithmic decision systems. First, my results show that the behavior and decision process of the people being evaluated by an algorithm (recruiter) have to be considered and understood in order to fully understand the resulting implications. Second, faced with people valuing both monetary outcomes and non-discriminating algorithms, the policy makers' goal has to also cater for both dimensions. Current legislation mostly aims at restricting algorithms on

the basis of discrimination and biases against specific subgroups of the population. From the view of the job applicants, it is, however, also important to look into which levels of transparency best allows to select the more beneficial recruiter or decision system. Last, more information about an algorithmic recruiter does not necessarily lead to higher perceived transparency. Within the current discussion about algorithmic transparency, one often-heard claim is to explain its weights to the people being judged by them. This research therefore hints towards a potential discrepancy between the intended effect of such measures and their perceived effect. Due to partly unbalanced treatment groups and a comparatively small sample size, these findings should be taking with some caution. However, as this is the first paper which explores the realm of how job applicants or individuals behave facing (discriminatory) algorithmic screening, I believe that they point towards a multiple promising topics for further research.

Within the scope of this paper, some open questions remain. As the treatment only manipulates the information about the algorithmic recruiter, it is unclear whether the observed effects are easily transferable to human recruiters or are to some extent innate to algorithms. The same holds true for the observed gender differences in perceiving discrimination. In this study, the algorithm discriminated against females while giving men a more favorable outcome. Consequently, women perceived it as more discriminatory and less fair, while men showed no such differences. Do men and women perceive algorithmic discrimination equally? Future studies could therefore also include a treatment with a male discriminating algorithm. Another fruitful extension of this experimental design would be a disentanglement of communicating the functioning and the weights of the algorithmic recruiter, thus allowing a more detailed conclusion about perceived transparency. To get more robust findings, one could also replicate this experiment with a more balanced and controllable pool of participants. Apart from experiments, the most promising avenue for further research, in my opinion, is transferring the experimental designs to real labor markets. There, field studies monitoring the decision process of job applicant or subjects of algorithmic evaluation using real stakes and firms would greatly enhance the current debate surrounding algorithmic job recommendation.

REFERENCES

Ajunwa, I. and 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*, pages 61–91. Emerald Publishing Limited.

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., and Rieke, A. (2019). Discrimination through optimization: How facebook's ad delivery can lead to skewed outcomes. *arXiv preprint arXiv:1904.02095*.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arechar, A. A., Gächter, S., and Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, 21(1):99–131.

Austen-Smith, D. and Fryer Jr, R. G. (2005). An economic analysis of acting white. *The Quarterly Journal of Economics*, 120(2):551–583.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104:671.

Becker, A., Deckers, T., Dohmen, T., Falk, A., and Kosse, F. (2012). The relationship between economic preferences and psychological personality measures. *Annual Review Economics*, 4(1):453–478.

Camerer, C., Loewenstein, G., and Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of economic Literature*, 43(1):9–64.

Charness, G., Gneezy, U., and Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, 149:74–87.

Charness, G., Gneezy, U., and Kuhn, M. A. (2013). Experimental methods: Extra-laboratory experiments-extending the reach of experimental economics. *Journal of Economic Behavior & Organization*, 91:93–100.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree - an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM.

Costa, P. T. and MacCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI): Professional manual*. Psychological Assessment Resources, Incorporated.

Costa-Gomes, M. A. and Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3):729–762.

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

Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on Security and Privacy (SP)*, pages 598–617. IEEE.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674.

Devaraj, S., Easley, R. F., and Crant, J. M. (2008). Research notehow does personality matter? relating the five-factor model to technology acceptance and use. *Information systems research*, 19(1):93–105.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM.

Eckel, C. C. and Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, 23(4):281–295.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM.

Fryer Jr, R. G. and Torelli, P. (2010). An empirical analysis of acting white. *Journal of Public Economics*, 94(5-6):380–396.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

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

Heyman, J. and Ariely, D. (2004). Effort for payment: A tale of two markets. *Psychological Science*, 15(11):787–793.

Hoffman, M., Kahn, L. B., and Li, D. (2017). Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800.

Hogan, J. and Ones, D. S. (1997). Conscientiousness and integrity at work. In *Handbook of personality psychology*, pages 849–870. Elsevier.

Holzmeister, F. (2017). otree: Ready-made apps for risk preference elicitation methods. *Journal of Behavioral and Experimental Finance*, 16:33–38.

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

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

John, O. P., Srivastava, S., et al. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and research*, 2(1999):102–138.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018a). Algorithmic fairness. In *AEA Papers and Proceedings*, volume 108, pages 22–27.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. (2018b). Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10.

Kuhn, P. and Mansour, H. (2014). Is internet job search still ineffective? *The Economic Journal*, 124(581):1213–1233.

Kuhn, P. and Skuterud, M. (2004). Internet job search and unemployment durations. *American Economic Review*, 94(1):218–232.

Lambrecht, A. and Tucker, C. (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*.

Lang, F. R., John, D., Lüdtke, O., Schupp, J., and Wagner, G. G. (2011). Short assessment of the big five: Robust across survey methods except telephone interviewing. *Behavior Research Methods*, 43(2):548–567.

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

Mason, W. and Suri, S. (2012). Conducting behavioral research on amazons mechanical turk. *Behavior Research Methods*, 44(1):1–23.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Paolacci, G. and Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

Prahl, A. and Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. (2017). Making sense of recommendations. *Journal of Behavioral Decision Making*.

DESCRIPTION OF ALL USED VARIABLES

Main outcome variables:

- Choosing the algorithmic recruiter: A dichotomous variable indicating if participants chose the algorithmic recruiter (0) or the human recruiter
- Belief$_{Human}$: Believed predicted average time from the human recruiter in seconds. Specific question within the experiment: "What average time do you think the human recruiter assigns to you?"
- Belief$_{Algorithm}$: Believed predicted average time from the algorithmic recruiter in seconds. Specific question within the experiment: "What average time do you think the algorithmic recruiter assigns to you?"
- Belief$_{Difference}$: Difference between the believed scores: Belief$_{Algorithm}$ - Belief$_{Human}$.
- Confidence$_{Hum}$: Confidence of participants with which they stated their beliefs. "How likely is it that your estimate is within 10 seconds of the actual human prediction? (0 means not likely at all, 100 means certainty"
- Confidence$_{Alg}$: Confidence of participants with which they stated their beliefs. "How likely is it that your estimate is within 10 seconds of the actual algorithmic prediction? (0 means not likely at all, 100 means certainty"
- Confidence$_{Diff}$: Difference between both confidence levels: Confidence$_{Alg}$ - Confidence$_{Hum}$
- Accuracy$_{Alg}$: Accuracy of the stated belief about the algorithmic recruiter compared to its actual prediction.
- Accuracy$_{Alg}$: Accuracy of the stated belief about the human recruiter compared to its actual prediction.
- Accuracy$_{Diff}$: Difference between both accuracy levels: Accuracy$_{Alg}$- Accuracy$_{Hum}$.
- Payoff-max: A dichotomous variable indicating if participants chose the recruitment method which would have earned them the higher income.
- Belief-consistent: A dichotomous variable indicating if participants chose the recruitment method which they believed to assign them the higher score.

Variables used for explanation and as an outcome:

- Discriminate: Participants stated to what extent the algorithmic recruiter discriminates more than the human recruiter on a 1-5 Likert scale, with 1: Strongly disagree; 2: Disagree; 3: Neither agree nor disagree; 4: Agree; 5: Strongly agree
- Be Fair: Participants stated to what extent the algorithmic recruiter is more likely to be fair than the human recruiter. Same Likert scale used as above.
- Be Transparent: Participants stated to what extent the algorithmic recruiter is more likely to be transparent than the human recruiter. Same Likert scale used as above.

Variables used only as explanatory variables: Personal characteristics

- Gender: Dichotomous variable indicating if a participants is male (1) or female (0). Self-reported
- Age: Ordinal variable in 5-year intervals, ranging from "less than 15" (1) to "more than 65" (12). Self-reported
- Education: Ordinal variable for the highest achieved education level. Possible answers are: some secondary education (high school) (1), completed secondary education (graduated high school) (2), trade/technical/vocational training (3), some undergraduate education (college or university) (4), completed undergraduate education(5); some postgraduate education (6); completed postgraduate education (masters or doctorate) (7). Self-reported.
- NonWhite: A dichotomous variable indicating if participants self-reported themselves as "White" (0) or any other ethnicity (1).

Variables used only as explanatory variables: Five factor model and risk aversion. All five variables are ordinal variables constructed out of three self-describing sentences where participants agreed/disagreed on a 1-5 point Likert scale.

1) Openness: ”I see myself as someone who is original, comes up with new ideas”, ”I see myself as someone who values artistic, aesthetic experiences”, ”I see myself as someone who has an active imaginations”.

2) Extraversion: ”I see myself as someone who is talkative”, ”I see myself as someone who is outgoing, sociable”, ”I see myself as someone who is reserved”

3) Neuroticism: ”I see myself as someone who worries a lot”, ”I see myself as someone who gets nervous easily”, ”I see myself as someone who is relaxed, handles stress well”

4) Conscientiousness: ”I see myself as someone who does a thorough job”, ”I see myself as someone who tends to be lazy”, ”I see myself as someone who does things efficiently”

5) Agreeableness: ”I see myself as someone who is sometimes rude to others”, ”I see myself as someone who has a forgiving nature”, ”I see myself as someone who is considerate and kind to almost everyone”

6) Risk Aversion: Ordinal variable ranging from 1-5, depicting each one lottery which the participants could choose and which increased in volatility. For more details, see (Holzmeister, 2017)

Qualitative Beliefs:

– Simpler: Participants stated to what extent the algorithmic recruiter is more likely to decide simple than the human recruiter on a 1-5 Likert scale, with 1: Strongly disagree; 2: Disagree; 3: Neither agree nor disagree; 4: Agree; 5: Strongly agree

– Familiar: Participants stated to what extent the algorithmic recruiter is more familiar to them than the human recruiter. Same Likert scale used as above.

– Care about characteristics: Participants stated to what extent the algorithmic recruiter is more likely to care about their characteristics than the human recruiter. Same Likert scale used as above.

– Prone to error: Participants stated to what extent the algorithmic recruiter is more likely to make errors than the human recruiter. Same Likert scale used as above.

– : Participants stated to what extent the algorithmic recruiter is more likely to care about their performance than the human recruiter. Same Likert scale used as above.

– : Participants stated to what extent the algorithmic recruiter is more likely to decide quickly than the human recruiter. Same Likert scale used as above.

– : Participants stated to what extent the algorithmic recruiter is more likely care about other things not mentioned than the human recruiter. Same Likert scale used as above.

APPENDIX FOR THE MAIN EXPERIMENT

Table B1—: Logistic regression: Estimated coefficients for the discrete choice between the human recruiter (0) and the algorithmic recruiter (1)

		Confidence Interval			
	Odds-ratio	2.5 %	97.5 %	z-score	p-value
Constant	1.95	0.14	27.33	0.50	0.62
Treatment	0.04	0.00	2.11	-1.58	0.11
Gender	2.24	0.74	6.81	1.42	0.16
Education	1.03	0.69	1.54	0.14	0.89
Non-White	0.26†	0.06	1.19	-1.74	0.08
Age	0.98	0.79	1.21	-0.22	0.82
Observed round	0.99	0.96	1.02	- 0.69	0.49
Treatment × Gender	11.08**	1.71	71.79	2.52	0.01
Treatment × Education	1.81†	0.96	3.41	1.83	0.07
Treatment × Age	0.92	0.65	1.32	-0.43	0.67
Treatment × Non-White	6.28†	0.83	47.6	1.78	0.08
Treatment × Observed round	0.99	0.90	1.08	-0.28	0.78

Note: This table reports estimates of a logistic regression where the dependent variable is an indicator of the participant choosing the algorithmic recruiter (being 1) or the human recruiter (being 0). The column of "Odds-ratio" shows the exponentiated regression coefficient. As an example of interpreting the odds-ratio for eg. the treatment - gender interaction: Holding all other characteristics and treatment interactions constant, treated males have a 1008% higher probability of selecting the algorithmic recruiter than treated females.
Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table B2—: Logistic regression: Estimated coefficients for the discrete choice between the human recruiter (0) and the algorithmic recruiter (1) including personality traits and risk aversion

		Confidence Interval			
	Odds-ratio	2.5 %	97.5 %	z-score	p-value
Constant	0.01	0.00	12.00	-1.23	0.22
Gender	3.63 †	0.91	14.47	1.82	0.07
Age	0.96	0.72	1.28	-0.30	0.77
Education	1.01	0.59	1.73	0.03	0.97
NonWhite	0.10 *	0.01	0.96	-2.00	0.05
Observed round	0.99	0.95	1.02	-0.74	0.46
Treatment	0.05	0.00	2406.24	-0.56	0.58
Risk aversion	0.85	0.57	1.28	-0.77	0.44
Openness	0.66	0.27	1.61	-0.92	0.36
Conscientiousness	5.41**	1.46	20.00	2.53	0.01
Extraversion	0.72	0.36	1.46	-0.90	0.37
Agreeableness	0.94	0.31	2.80	-0.12	0.90
Neuroticism	1.85	0.80	4.27	1.44	0.15
Treatment $\times$Gender	11.35†	0.96	134.69	1.93	0.05
Treatment $\times$Age	0.85	0.54	1.34	-0.68	0.49
Treatment $\times$Education	1.90	0.87	4.18	1.60	0.11
Treatment $\times$NonWhite	17.69 *	1.14	273.87	2.06	0.04
Treatment $\times$Observed round	1.01	0.91	1.12	0.16	0.87
Treatment $\times$Risk aversion	1.08	0.60	1.94	0.25	0.80
TTreatment $\times$Openness	0.92	0.25	3.38	-0.12	0.90
Treatment $\times$Conscientiousness	0.43	0.09	2.17	-1.02	0.31
Treatment $\times$Extraversion	1.95	0.72	5.31	1.31	0.19
Treatment $\times$Agreeableness	2.00	0.48	8.40	0.95	0.34
Treatment $\times$Neuroticism	0.55	0.19	1.62	-1.08	0.28

Note: This table reports estimates of a logistic regression where the dependent variable is an indicator of the participant choosing the algorithmic recruiter (being 1) or the human recruiter (being 0). The column of "Odds-ratio" shows the exponentiated regression coefficient. As an example of interpreting the odds-ratio for eg. conscientiousness: Holding all other characteristics and treatment interactions constant, a one-standard deviation higher level of conscientiousness implies a 5.41 higher probability to choose the algorithmic recruiter. Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table B4—: Full regression specification including personal covariates, personality traits, risk aversion and interaction effects for the discrete choice and difference of beliefs

	Chosen recruitment method:	
	Algorithmic Recruiter	Belief$_{Difference}$
	(1)	(2)
Constant	0.744***	6.938*
	(0.287)	(3.680)
Gender	0.245†	$-$2.491*
	(0.136)	(1.202)
Age	0.0001	$-$0.141

	(0.036)	(0.254)
Education	−0.00001	−0.892†
	(0.042)	(0.486)
NonWhite	−0.374*	0.512
	(0.181)	(2.171)
Observed round	−0.003	−0.004
	(0.004)	(0.056)
Treatment	−0.454	−4.184
	(0.376)	(4.621)
Risk aversion	−0.030	−0.544
	(0.039)	(0.617)
Openness	−0.045	−0.669
	(0.070)	(0.690)
Conscientiousness	0.213***	−4.193**
	(0.067)	(1.681)
Extraversion	−0.071	−0.172
	(0.073)	(0.634)
Agreeableness	0.002	−1.012
	(0.067)	(1.243)
Neuroticism	0.144†	−3.680*
	(0.087)	(1.747)
Treatment × Gender	0.261	−3.239*
	(0.171)	(1.879)
Treatment × Age	−0.029	0.606
	(0.044)	(0.392)
Treatment × Education	0.068	0.754
	(0.050)	(0.710)
Treatment × NonWhite	0.446*	−0.576
	(0.209)	(2.420)
Treatment × Observed round	0.002	−0.111
	(0.008)	(0.095)
Treatment × Risk aversion	0.010	0.905
	(0.050)	(0.718)
Treatment × Openness	0.0002	0.073
	(0.083)	(0.912)
Treatment × Conscientiousness	−0.136	4.892**
	(0.086)	(1.750)

Treatment × Extraversion	0.113	−0.073
	(0.084)	(0.842)
Treatment × Agreeableness	0.062	0.699
	(0.083)	(1.334)
Treatment × Neuroticism	−0.147	$3.403^{\dagger}$
	(0.109)	(1.873)
Observations	148	148
R^2	0.381	0.379
Adjusted R^2	0.266	0.264

Note: This table reports the estimated coefficients for regression equation 4. The dependent variables are (1) an indicator of choosing the algorithmic recruiter and (2) the difference between the believed predicted average times of both recruitment methods. If (2) is negative, it indicates a better believed prediction from the algorithmic recruiter, if positive a better believed prediction from the human recruiter. The personality variables of the five factor model are standardized to a mean of zero and a standard deviation of 1. Robust standard errors are reported.
Significance indicators: † $p<0.1$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$

Table B3—: Probit regression: Estimated coefficients for the discrete choice between the human recruiter (0) and the algorithmic recruiter (1)

	Ratio of choosing the algorithm
Constant	0.376
	(0.812)
Treatment	$-1.938^{\dagger}$
	(1.165)
Gender	0.489
	(0.341)
Education	0.024
	(0.124)
NonWhite	$-0.839^{\dagger}$
	(0.471)
Age	−0.014
	(0.067)
Observed round	−0.007
	(0.010)
Treatment × Gender	1.378^{**}
	(0.530)
Treatment × Education	$0.353^{\dagger}$
	(0.187)
Treatment × Age	−0.030
	(0.108)
Treatment × NonWhite	$1.065^{\dagger}$
	(0.618)
Treatment × observed round	−0.010
	(0.027)
Observations	148
Log Likelihood	−74.758
Akaike Inf. Crit.	173.516
Note:	*p<0.1; **p<0.05; ***p<0.01

Note: This table reports estimates of a probit regression where the dependent variable is an indicator of the participant choosing the algorithmic recruiter (being 1) or the human recruiter (being 0). Normal standard errors are supported Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table B5—: Kolmogorov-Smirnov Test: Comparing if the distributions from the believed predicted scores come from the same common distribution

Test	Distributions	KS test statistic	p-value
Two tailed Kolmogorov-Smirnov (KS)	$Belief_{AlgorithmicRecruiter}$	0.096	0.888
Two tailed Kolmogorov-Smirnov (KS)	$Belief_{HumanRecruiter}$	0.1378	0.49
Two tailed Kolmogorov-Smirnov (KS)	Difference between both recruiters	0.24797	0.022

Table B10—: Belief-Choice Interaction 3: Do participants choose the recruiter who is believed to assign better predictions?

	Ratio of "rational" behavior	
	Payoff-maximizing recruiter	Belief-consistent
	(1)	(2)
Constant	0.310	0.509
	(0.341)	(0.349)
Gender	−0.074	0.182
	(0.138)	(0.148)
Age	0.009	−0.059*
	(0.029)	(0.023)
Education	0.032	0.048
	(0.062)	(0.061)
NonWhite	0.474**	−0.222
	(0.178)	(0.171)
Observed round	0.002	0.006
	(0.004)	(0.004)
Treatment	0.775†	0.630
	(0.452)	(0.436)
Risk aversion	−0.026	−0.027
	(0.044)	(0.046)
Openness	0.011	0.038
	(0.088)	(0.076)
Conscientiousness	−0.031	−0.021
	(0.089)	(0.085)
Extraversion	0.048	−0.124
	(0.080)	(0.082)
Agreeableness	−0.186*	−0.127
	(0.080)	(0.083)

Neuroticism	−0.134	−0.151
	(0.097)	(0.106)
Treatment × Gender	−0.250	−0.284
	(0.189)	(0.193)
Treatment × Age	−0.019	0.085*
	(0.043)	(0.033)
Treatment × Education	−0.086	−0.084
	(0.073)	(0.070)
Treatment × NonWhite	−0.664***	0.199
	(0.213)	(0.202)
Treatment × observed round	−0.008	−0.033***
	(0.010)	(0.009)
Treatment × Risk aversion	0.045	0.019
	(0.058)	(0.057)
Treatment × Openness	−0.055	−0.081
	(0.106)	(0.095)
Treatment × Conscientiousness	0.002	0.072
	(0.110)	(0.104)
Treatment × Openness	−0.088	0.019
	(0.100)	(0.100)
Treatment × Agreeableness	0.108	0.176†
	(0.097)	(0.100)
Treatment × Neuroticism	0.108	0.185
	(0.122)	(0.131)
Observations	148	148
R^2	0.169	0.251
Adjusted R^2	0.014	0.112

Note: This table reports the estimated coefficients for regression equation 4. The dependent variables are (1) an indicator for choosing recruiter that yields the higher monetary outcome and (2) an indicator for choosing the recruiter which is believed to assign the higher score. The psychological character traits of the five factor model were standardized to display a mean of zero and standard deviation of one. Robust standard errors are reported.
Significance indicators: † $p<0.1$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$

Table B6—: Confidence of stated beliefs: OLS regression results for the algorithmic recruiter (1), the human recruiter (2) and the difference between both (3)

	Confidence of the believed predictions		
	Confidence_{Alg}	Confidence_{Hum}	Confidence_{Diff}
	(1)	(2)	(3)
Constant	85.414***	92.662***	−7.247
	(15.469)	(15.938)	(8.175)
Treatment	−24.484	−28.074	3.590
	(19.129)	(20.342)	(13.019)
Gender	5.918	2.390	3.528
	(6.120)	(6.212)	(3.676)
Age	−1.397	−1.957†	0.560
	(1.034)	(1.059)	(0.474)
Education	−0.810	−1.773	0.962
	(2.699)	(2.781)	(1.358)
Observed round	−0.152	−0.168	0.016
	(0.154)	(0.172)	(0.064)
Non-White	−5.317	−6.294	0.977
	(8.869)	(9.028)	(3.887)
Treatment × :Gender	7.806	8.960	−1.154
	(8.150)	(8.278)	(5.225)
Treatment × Age	3.732*	3.399*	0.332
	(1.451)	(1.673)	(1.155)
Treatment × Education	0.384	1.318	−0.934
	(3.275)	(3.420)	(2.039)
Treatment × Observed round	−0.518	−0.361	−0.156
	(0.372)	(0.382)	(0.241)
Treatment × Non-White	0.609	0.419	0.190
	(10.383)	(10.669)	(5.112)
Observations	148	148	148
R^2	0.123	0.095	0.028

Note: This table reports the estimated coefficients for regression equation 2. The dependent variables are: (1) and (2) are the stated confidence in the believed predicted average speed of the algorithmic recruiter and the human recruiter. (3) is the difference between both, where a negative value indicates a less confident belief about the algorithmic recruiter compared to the human recruiter. Robust standard errors are reported.
Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table B7—: Kolmogorov-Smirnov Test: Comparing if the distributions from the belief accuracies come from the same common distribution

Test	Distributions	KS test statistic	p-value
Two tailed Kolmogorov-Smirnov (KS)	$\text{Accuracy}_{AlogrithmicRecruiter}$	0.17	0.2412
Two tailed Kolmogorov-Smirnov (KS)	$Accuracy_{HumanRecruiter}$	0.0983	0.8716

Table B8—: Accuracy of stated beliefs: OLS regression results for the algorithmic recruiter (1) and the human recruiter (2)

	Accuracy of the believed predictions	
	Accuracy_{Alg}	Accuracy_{Hum}
	(1)	(2)
Constant	1.444	−6.383
	(8.888)	(6.837)
Gender	1.765	$-5.931^{\dagger}$
	(2.646)	(3.172)
Age	−0.335	−0.365
	(0.514)	(0.549)
Education	−1.005	1.001
	(1.499)	(1.394)
Non-White	−1.965	−7.729
	(4.634)	(8.081)
Observed round	0.090	0.365^{**}
	(0.126)	(0.125)
Treatment	−2.376	8.394
	(10.549)	(9.026)
Treatment × Gender	−5.082	2.621
	(3.389)	(3.760)
Treatment × Education	0.213	−2.198
	(1.807)	(1.684)
Treatment × Age	1.456	1.454
	(1.047)	(0.956)
Treatment × Non-White	−3.064	3.164
	(5.129)	(8.416)
Treatment × Observed round	0.005	−0.232
	(0.262)	(0.264)
Observations	148	148
R^2	0.117	0.207

Note:

Note: This table reports the estimated coefficients for regression equation 2. The dependent variables are: (1) and (2) are the how accurate participants stated their believed predicted average scores, compared to the actually predicted scores from both recruitment methods. Robust standard errors are reported.
Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table B9—: Logistic regression: Estimated coefficients for the discrete choice between the human recruiter (0) and the algorithmic recruiter (1)

		Confidence Interval			
	Odds-ratio	2.5 %	97.5 %	z-score	p-value
Constant	3.9† 0	0.91	16.72	1.83	0.07
D-Belief	0.40	0.11	1.42	-1.42	0.16
Belief$_{Human}$	1.15 †	1.00	1.32	1.96	0.05
Belief$_{Algorithm}$	0.85 *	0.74	0.98	-2.20	0.03
Treatment	0.58	0.08	4.24	-0.54	0.59
Treatment $\times$ D-Belief	5.00 †	0.85	29.43	1.78	0.08
Treatment $\times$ Belief$_{Algorithm}$	0.98	0.80	1.18	-0.25	0.80
Treatment $\times$ Belief$_{Human}$	1.03	0.84	1.26	0.28	0.78

Note: This table reports estimates of a logistic regression where the dependent variable is an indicator for the participant to choose the algorithmic recruiter (being 1) or the human recruiter (being 0). The column of "Odds-ratio" shows the exponentiated regression coefficient. Interpration of the coefficient of $Belief_{Human}$: Holding all other characteristics and treatment interactions constant, an increased (worse) believed prediction from the human recruiter increases the odds of choosing the algorithmic reccruiter by 15 %. Standard errors are reported. Significance indicators: † p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table B11—: Comparing the recruiters: Average Qualitative Beliefs for both genders across the treatment and control group

	Baseline	Treatment	T-statistic
Fairness	3.73	3.22	2.92
Transparent	3.85	3.84	0.05
Simpler	3.83	3.61	1.31
Discriminate	1.86	2.51	-3.26
Familiar	2.52	2.60	-0.44
Care about Characteristics	2.02	2.34	15.24
Prone to Error	2.33	2.65	18.72
Performance	3.52	2.99	2.56
Quickly	4.26	4.33	-0.56
Other	2.05	2.18	-0.84

Table B12—: Comparing the recruiters: Average Qualitative Beliefs across genders in the control group

	Female	Male	T-statistic
Fairness	3.66	3.81	-0.67
Transparent	3.94	3.74	0.96
Simpler	3.63	4.06	-1.94
Discriminate	1.83	1.90	-0.30
Familiar	2.46	2.58	-0.46
Care about Characteristics	2.00	2.03	-0.12
Prone to Error	2.40	2.26	0.57
Performance	3.40	3.65	-0.77
Quickly	4.23	4.29	-0.30
Other	2.11	1.97	0.64

Table B13—: Comparing the recruiters: Average Qualitative Beliefs of females in the baseline and control group

	Baseline	Treatment	T-statistic
Fairness	3.66	2.84	3.38
Transparent	3.94	3.79	0.81
Simpler	3.63	3.49	0.61
Discriminate	1.83	3.12	-4.54
Familiar	2.46	2.23	0.91
Care about Characteristics	2.00	2.12	-0.47
Prone to Error	2.40	2.91	-1.95
Performance	3.40	2.67	2.57
Quickly	4.23	4.47	-1.48
Other	2.11	2.21	-0.42

Table B14—: Comparing the recruiters: Average qualitative beliefs of males in the baseline and control

	Baseline	Treatment	T-statistic
Fairness	3.81	3.64	0.71
Transparent	3.74	3.90	-0.61
Simpler	4.06	3.74	1.27
Discriminate	1.90	1.85	0.24
Familiar	2.58	3.00	-1.52
Care about Characteristics	2.03	2.59	-2.10
Prone to Error	2.26	2.36	-0.44
Performance	3.65	3.33	1.06
Quickly	4.29	4.18	0.54
Other	1.97	2.15	-0.78

Table B15—: ANOVA Analysis wether including Qualitative Beliefs to the incentivized Beliefs on the assigned scores helps in explaining the Choice of Recruitment Method

	Res.Df	RSS	DF	Sum of Sq.	F-stat	Pr(>F
Incentivized Beliefs	140	28.826	8	64.174	38.96	2.26e-16 ***
Perceived Fairness	138	25.233	2	3.593	9.824	0.0001***
Perceived Discrimination	136	24.104	2	3.593	3.1862	0.444*
Perceived Performance	134	23.582	2	0.522	1.842	0.231
All further seven beliefs	126 1	23.276	8	0.306	0.207	0.989

Note: The table depicts an ANOVA Analysis (equivalent to an F-test), testing if the addition of further qualitative beliefs helps in explaining the variance of choosing the algorithmic/human recruiter. "Incentivized Beliefs" means a model as in table 7. The next three rows show the inclusion of the qualitative belief about fairness, discrimination and caring about the participants' performance. The last row shows the result for including all further seven beliefs within a regression explaining the choice of algorithmic/human recruiter. Significance indicators: Significance indicators: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table B16—: Interplay of Beliefs: To what extent do the qualitative beliefs explain the incentivized beliefs about the payoff?

	Believed predicted scores by the recruiters	
	Belief_{Alg}	Belief_{Diff}
	(1)	(2)
Constant	21.569***	0.286
	(3.283)	(1.038)
Fair	−0.854	0.875
	(4.152)	(1.501)
Discriminate	2.188	2.112
	(3.640)	(2.134)
Performance	−1.091	−0.972
	(2.084)	(0.817)
Transparent	2.799	0.763
	(2.338)	(0.925)
Treatment	−3.778	0.390
	(3.536)	(1.143)
Treatment × Fair	−0.031	−1.968
	(4.399)	(1.673)
Treatment × Discriminate	0.692	−0.717
	(4.184)	(2.258)
Treatment × Performance	0.407	0.530
	(2.741)	(0.960)
Treatment × Transparent	−1.839	0.179
	(2.716)	(1.071)
Observations	148	148
R^2	0.062	0.155
Adjusted R^2	0.001	0.100
Observations	148	148
R^2	0.062	0.155
Adjusted R^2	0.001	0.100

Note: This table reports OLS regressions where the dependent variable is in Column (1) the incentivized belief of participants about the predicted prediction from the algorithmic recruiter and (2) is a continuous variable indicating the difference between both scores. A negative values thereby shows a better believed prediction from the algorithmic recruiter, while a positive prediction signifies a better believed prediction by the human recruiter. The qualitative beliefs of perceived fairness, discrimination, performance and transparency are standardized to possess a mean of zero and a standard deviation of one. Robust standard errors are reported.
Significance indicators: *p<0.1; **p<0.05; ***p<0.01

Table B17—: Time Participants needed for each Page of the Experiment

Page Name	Baseline	Treatment
Introduction	131.53	280.73
Questionaire	34.38	31.63
Start	96.15	73.95
training	27.71	20.20
training2	16.20	30.55
training3	7.58	10.21
q1	17.85	13.40
q2	23.05	15.29
q3	22.02	17.56
q4	11.62	11.50
q5	16.95	12.40
q6	20.98	19.06
q7	15.32	11.38
q8	23.09	17.43
q9	21.98	14.68
q10	13.68	16.07
Results	10.44	9.80
Payoff2	10.02	11.24
Introduction	51.65	109.16
Introduction2	31.47	35.98
Discrete	11.77	11.05
Belief	47.73	37.88
Belief3	45.23	74.95
Results	18.91	18.22
Introduction	94.45	102.80
Likert	51.17	52.33
Instructions	25.77	32.89
Decision	24.94	23.27
Results	7.18	6.60
Introduction	3.23	3.01
Questionaire	43.82	43.98
Results	10.08	10.98
PaymentInfo	3.41	3.06