Dargnies, Marie-Pierre; Hakimov, Rustamdjan; Kübler, Dorothea



# Aversion to Hiring Algorithms: Transparency, Gender Profiling, and Self-Confidence

Management Science

**Provided in Cooperation with:**
WZB Berlin Social Science Center

*Suggested Citation:* Dargnies, Marie-Pierre; Hakimov, Rustamdjan; Kübler, Dorothea (2024) : Aversion to Hiring Algorithms: Transparency, Gender Profiling, and Self-Confidence, Management Science, ISSN 1526-5501, Institute for Operations Research and the Management Sciences (INFORMS), Catonsville, MD, Iss. Ahead of Print, pp. 1-37, https://doi.org/10.1287/mnsc.2022.02774

# Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence

Marie-Pierre Dargnies (University of Paris Dauphine, PSL)
Rustamdjan Hakimov (University of Lausanne)
Dorothea Kübler (WZB Berlin, Technische Universität Berlin & CESIfo)

March 2023

Abstract

We run an online experiment to study the origins of algorithm aversion. Participants are in the role of either workers or managers. Workers perform three real-effort tasks: task 1, task 2, and the job task, which is a combination of tasks 1 and 2. They choose whether the hiring decision between themselves and another worker is made by a participant in the role of a manager or by an algorithm. In a second set of experiments, managers choose whether they want to delegate their hiring decisions to the algorithm. When the algorithm does not use workers' gender to predict their job task performance and workers know this, they choose the algorithm more often than in the baseline treatment where gender is employed. Feedback to the managers about their performance in hiring the best workers increases their preference for the algorithm relative to the baseline without feedback, because managers are, on average, overconfident. Finally, providing details on how the algorithm works does not increase the preference for the algorithm for workers or for managers.

Acknowledgments:
We would like to thank the editor Elena Katok, an associate editor, and three reviewers for the helpful and constructive comments. We are also grateful to Ben Greiner for valuable comments and Jennifer Rontganger for copy editing. Marie-Pierre Dargnies acknowledges financial support from the ANR (ANR JCJC TrustSciTruths), Rustamdjan Hakimov from the Swiss National Science Foundation (project #100018_189152), and Dorothea Kübler from the Deutsche Forschungsgemeinschaft (DFG) through CRC TRR 190.

This is an Accepted Manuscript of an article published by Informs in Management Science on June 19, 2024 (Ahead of Print), available at: https://doi.org/10.1287/mnsc.2022.02774.

## Introduction

Companies use algorithms for a variety of decisions, including forecasts of applicants' performance for hiring decisions (Highhouse 2008, Carey and Smith 2016). Firms that use such algorithms appear to perform better than others (Bajari et al. 2019, Camuffo et al. 2020), for example, because the algorithm increases the hiring of female and minority candidates (Cowgill 2020, Avery et al. 2023). Algorithms are also increasingly used in the realm of public policy, such as jail-or-release decisions (Kleinberg et al. 2018) or credit scoring (Baesens et al. 2003). Recent evidence suggests that while the use of AI can entail important welfare gains (see Chalfin et al. 2016 for the hiring of policemen and tenure decisions of teachers), people are often opposed to the adoption of algorithms, thereby displaying algorithm aversion (Dietvorst et al. 2015, Castelo et al. 2019, Jussupow et al. 2020).

In this paper, we study people's attitudes toward algorithms in the context of hiring decisions. We consider the perspectives of both workers and managers. Workers are directly affected by the hiring decisions, and we therefore expect that self-interest will influence their preference for an algorithmic over a managerial decision. Namely, workers will tend to choose the hiring process they believe is more likely to favor them. Managers are also expected to be self-interested and to care primarily about the efficiency of the hiring decisions. They will tend to choose the hiring process they believe is more likely to result in the hiring of the best workers, because this choice generates the highest payoff for them.[1]

We employ a series of lab experiments to investigate the causes of algorithm aversion of workers and managers in the context of hiring decisions. Lab experiments allow us to tightly control the decision environment. We can measure the performance of workers in a straightforward manner. Moreover, beliefs play an important role in algorithm aversion. The lab setting allows us to elicit the workers' and managers' level of self-confidence and their beliefs about gender differences in performance. Finally, by programming the algorithm ourselves, we can provide participants with complete and truthful information.

Recent debates in the EU over the legal requirements for algorithmic decisions motivate our experiments. Algorithms often use gender, race, and other personal information to predict

[1]In reality, managers might be interested in hiring the cheapest workers rather than the best workers. The insights of the experiment generalize to these cases as long as the managers are interested in the process that gives them the most accurate ranking of the candidates' quality.

behavior, known as profiling and amounting to statistical discrimination. Under which circumstances profiling is acceptable is central to the discussion. On the one hand, the quality of the algorithm's predictions worsens when it uses less information. On the other hand, Paragraph 71 of the preamble to the General Data Protection Regulation (GDPR) requires data controllers to prevent discriminatory effects of algorithms processing sensitive personal data such as ethnic origin and religion.[2] Our experiment aims to understand whether people are opposed to algorithms that make predictions based on gender, which plays a central role in the public discourse even though it is not among the sensitive data mentioned in the preamble of the GDPR.

Furthermore, Articles 13 and 14 of the GDPR state that when profiling takes place, people have the right to "meaningful information about the logic involved" (Goodman and Flaxman 2017).[3] Although the GDPR has led to some expected effects, for example, privacy-oriented consumers opting out of the use of cookies (Aridor et al. 2020), whether and for whom "meaningful information" about the algorithm increases its acceptance are open questions. When the algorithm is made transparent in our setup, the favored gender becomes evident. Thus, transparency can impact algorithm aversion differently for women and men.

We run an online experiment in which participants are in the role of either workers or managers. Workers perform three real-effort tasks: task 1, task 2, and the job task, which is a combination of tasks 1 and 2. Workers are paid for their performance in all three tasks, but only the work task is relevant for the payment of the managers. Both the managers and the algorithm know the applicants' performance in tasks 1 and 2 and their gender, but do not know the performance in the job task.

In the baseline treatment for workers, they choose whether they prefer that a hiring decision between themselves and another worker is made by a participant in the role of a manager or by

[2]For example, the European Parliament passed the Digital Services Act (DSA) to regulate platforms such as Facebook and Google, requiring algorithm disclosure, the provision of a profiling-free option to users, together with a complete ban on the profiling of minors for targeted advertising. The DSA was published on October 27, 2022, came into force on November 16, 2022, and will apply in all EU countries beginning February 17, 2024. See the press release of the European Parliament from March 23, 2022. https://www.europarl.europa.eu/news/en/press-room/20220412IPR27111/digital-services-act-agreement-for-a-transparent-and-safe-online-environment

[3]Transparency is also perceived as normatively desirable in the matching algorithm in the context of school choice and university admissions (Hakimov and Raghavan 2022, Grigoryan 2022).

an algorithm. In the baseline treatment for managers, they decide whether to delegate the hiring to the algorithm or decide themselves. If they delegate the decisions, their payoff will depend on the algorithm's hiring decisions instead of their own.

In the first treatment for workers, we study whether applicants are more likely to accept algorithms that are not allowed to use gender information and are therefore unable to discriminate on the basis of gender.[4] A second series of experiments focuses on the managers' preferences and aims at understanding why some firms are more reluctant than others to use hiring algorithms. One possible explanation for not adopting such algorithms is managerial overconfidence. Overconfidence is a common bias whose effect on economic behavior has been demonstrated (Camerer et al. 1999, Dunning et al. 2004, Malmendier and Tate 2005, Dargnies et al. 2019). In our context, overconfidence is likely to induce managers to delegate the hiring decisions to the algorithm too seldom, because they believe they make better hiring decisions than they actually do. Finally, a third set of experiments examines whether the transparency of the algorithm increases its acceptance among workers and managers. In this version of the experiment, the participants learn how the algorithm weights the available information about the applicants to predict their performance in the job task.

In our experiment, we find that for given task-1 and task-2 performances, female workers perform better. A possible explanation is that women learn more from previous experience than men. As a result, the algorithm favors female workers over male workers when they have similar task-1 and task-2 performances. The managers in our experiment also favor female workers, which suggests their observation of a random set of 20 workers is sufficient to spot this pattern. Overall, the proportion of correct hires by the algorithm is significantly higher than that by the managers (67% vs. 56%).

In the baseline treatment, 47% of workers chose the hiring algorithm. The preference for the algorithm is positively correlated with the workers' confidence in their performance. Also, we observe that choices are motivated by self-interest: female (male) workers who believed that managers favor female (male) workers more than the algorithm were significantly (marginally significantly) more likely to choose managers. When the algorithm does not use the gender of

[4] In general, excluding the use of gender does not rule out the possibility of indirect gender discrimination, for example when other characteristics correlate with gender. However, this is not possible in our experiment, since the gender-blind algorithm only uses the performances of workers.

workers to predict their job task performance, and workers know this, they choose the algorithm significantly more often (59%). Thus, we observe a preference for no gender profiling. On the other hand, providing details about how the algorithm works does not increase the workers' preference for the algorithm. The result suggests the workers' preference for transparency is not very strong or, alternatively, that we did not succeed in conveying how the algorithm works.

Next, we look at algorithm aversion by workers, that is, whether workers avoid hiring by the algorithm despite it being in their best interest. Across all treatments, around half of the workers choose the hiring method that maximizes their earnings. 24.3% of workers are algorithm averse while 25.1% are algorithm loving in that they choose the algorithm although they would have earned more otherwise. In the baseline treatment, the share of algorithm-averse workers is 29%, but when the algorithm does not use the gender of workers, the share of algorithm-averse choices is significantly lower (20.2%). Thus, the guarantee of not using gender profiling increases the preference for algorithmic hiring and reduces algorithm aversion.

For managers, in the baseline treatment, only 34% of managers delegated the hiring decisions to the algorithm, despite the algorithm being, on average, more efficient than the managers. As expected, delegation was negatively correlated with the managers' beliefs about how many workers they hired correctly. Feedback to managers on their ability to hire the best workers significantly increased the delegation to the algorithm (from 34.1% to 49.8%), especially for managers who were overconfident and therefore received feedback that they had made significantly fewer correct hiring decisions than they thought.

We also study algorithm aversion of the managers, that is, managers avoiding delegation to the algorithm despite it being more efficient. The proportion of algorithm-averse managers is 54.8% in the baseline treatment and 53.4% in the treatment with algorithm disclosure, indicating that disclosure did not significantly affect delegation by the managers. However, this proportion is significantly lower at 33.1% in the treatment with feedback on their ability to select the best workers. Thus, we establish a causal effect of overconfidence on algorithm aversion and show overconfidence might be a substantial barrier to the efficient adoption of algorithms.

**Related literature**

Our paper contributes to the literature on preferences for hiring algorithms. Some papers study general attitudes toward algorithmic hiring, whereas recent work specifically considers managers' and workers' attitudes. Among the earlier papers, Lee (2018) manipulates the decision-maker (algorithmic or human) for managerial tasks including hiring, and measures the perceived fairness, trust, and emotional response. She finds people perceive human decisions to be fairer than algorithmic decisions in hiring tasks. Kaibel et al. (2019) had their participants evaluate the selection process of a fictitious company. They observe that participants perceive the organizations as less attractive when a hiring algorithm is used rather than humans. In a recent survey, Will et al. (2022) document that despite the evidence that AI is better than humans in hiring, candidates and recruiters perceive it to be worse. On the other hand, Bigman et al. (2022) show people are less outraged by a discriminating algorithm than by a discriminating human. Cowgill (2020) demonstrates that even when a hiring algorithm is trained with a biased dataset from human decision-makers, it can discriminate less than the underlying training set. Avery et al. (2023) demonstrate integrating AI into the hiring process for STEM workers reduces the gender gap on both the demand and supply sides. Specifically, the study finds women are more inclined to apply for positions when they know they will be evaluated using AI screening. Moreover, when recruiters are informed of an applicant's AI score, they exhibit a greater propensity to choose women than in situations where they are unaware of the score.

Focusing on the managers' adoption of algorithms, Cowgill et al. (2020a) expose managers to arguments used in AI fairness activism. Emphasizing that algorithms are inevitably biased leads managers to abandon AI in favor of manual review by humans. Managers who tend to abandon discriminating algorithms fear lawsuits and negative PR. On the other hand, exposing decision-makers to claims that although algorithmic bias exists, human-based alternatives could be more biased, encourages the adoption of AI. These motives are absent in our setup, where only performance matters for the managers' payoffs. Instead of fairness issues, our manager treatments study the role of overconfidence and transparency.

Fumagalli et al. (2022) investigate workers' preferences regarding algorithmic or human hiring decisions. Although both types of recruiters receive the same information (results from a test of job performance as well as a set of personal characteristics including gender), they find in a first experiment that subjects believe managers make more errors, whereas algorithms are more meritocratic with respect to performance. Also, lower-performing men have a weak preference

for human recruiters. In another experiment, workers compete against a fictitious worker whose gender is randomized, such that workers can form beliefs about the recruiters' gender bias after observing the recruitment decision. In contrast to this work, we study the relative acceptance of gender-neutral or transparent algorithms. Also, we investigate what drives the manager's decision to delegate the hiring decision to an algorithm.

Our paper also belongs to a literature studying the preference for human decision-making over algorithmic decision-making in contexts other than hiring. Fildes and Goodwin (2007) find many professional forecasters do not use algorithms (or do not use them enough) in their forecasting process. Sanders and Manrodt (2003) observe that many firms do not rely on algorithms for forecasting even though firms that did use them made fewer forecasting errors. In a similar vein, only a minority of clinical psychologists used algorithms when making clinical predictions (Vrieze and Grove 2009). On the other hand, Dietvorst et al. (2015) find people are not always averse to algorithms. Indeed, a majority of participants in their experiment used the algorithm's forecasts rather than their own when they had no information about the algorithm's performance. However, once the participants learned the algorithm was imperfect, they became reluctant to use it. Most participants (over 60%) in Chugunova and Luhan (2022) prefer the algorithm to a human decision-maker in the context of redistributive decisions, although they view the decisions made by humans more favorably than those made by the algorithm. Corgnet (2023) experimentally investigated the effects of human versus algorithmic dismissals on worker productivity. Workers respond more negatively to human dismissals than algorithmic dismissals, exhibiting reduced productivity in subsequent tasks. Jussupow et al. (2020) synthesize evidence of how the characteristics of algorithms and human decision-makers affect algorithm aversion. They find algorithm agency (the algorithm's ability to independently accomplish actions as opposed to having an advisory role), performance, perceived capabilities, and human involvement (particularly in the development and training of the algorithm) strongly influence aversion, along with human agents' expertise and social distance.

A number of papers investigate ways to mitigate algorithm aversion. Dietvorst et al. (2018) find being able to slightly modify the forecasts of the algorithm made the participants more willing to use the algorithm. Rich descriptions and explanations of the algorithm (a recommender system) increased participants' understanding of the recommendation process, which in turn improved their beliefs about the quality of the algorithm's performance (Yeomans et al. 2019). Relatedly, increasing a task's perceived objectivity increases trust in and use of algorithms for

that task (Castelo et al. 2019). Bigman and Gray (2018) suggest reducing aversion to moral decision-making by algorithms or machines is not easy and depends on making salient the expertise of machines and the ability of humans to override them.

Cowgill and Tucker (2020) provide a systematic overview of the sources of unfair algorithms, such as unrepresentative training sets, mislabeling in training sets, or biased programmers (Cowgill et al. 2020b). The ways in which algorithms discriminate between groups of the population, often indirectly by basing the choice on characteristics that are correlated with, for example, gender, is an important field of study (Persson 2016, Barocas and Selbst 2016, Hajian and Domingo-Ferrer 2012). For example, Lambrecht and Tucker (2019) show an algorithm determining who sees an ad on a social network, a process that is supposed to be gender neutral, delivers the ad more often to men, because doing so is more cost effective.

Our work also relates to a large and growing literature on discrimination in hiring (for reviews of the literature, see Charles and Guryan 2011, Lane 2016, Bertrand and Duflo 2017, and Blau and Kahn 2017). We explore ideas similar to Barron et al. (2022), in which the managers observe a performance measure of the workers that is correlated with their job performance. Whereas Barron et al., like most of the literature, find discrimination against women, both statistical and taste based, we find discrimination against men, on average, in the sense that women are favored over men when their past performances are identical. The favoring of women is likely driven by the fact that we provide more information to the managers than previous studies do. This information allows them to learn about the higher job performance of women relative to men with similar past performances. The most important difference between our work and the literature on hiring discrimination is our focus on the choice of algorithms.

In the classic principal-agent framework, a principal can delegate a task to an agent who has superior information but whose incentives are not aligned with those of the principal. In our setup, the algorithm is programmed to choose the best worker, such that incentives are aligned. However, the manager may not understand or trust the algorithm. Behavioral biases influence delegation, as shown in theory (see, e.g., Auster and Pavoni forthcoming) and in experiments (e.g., Danz et al. 2015, Ertac et al. 2020), and algorithm aversion can be understood as such a bias.

Finally, our work is part of a larger literature on human-machine interactions that includes but is not restricted to the role of computerized agents managing supply chains (Kimbrough et al. 2002, Badakhshan et al. 2020), and electronic reputation systems on trading platforms (Bolton et al. 2004). Aoki (2020) investigates the determinants of trust in AI chatbots that answer questions from the population on behalf of the government. She suggests explaining the goals of chatbot use improves trust. Greiner et al. (2022) study the effects of compensation contracts and the framing of algorithms on the reliance on algorithmic advice in a price-estimation task. All of this work relates to our study in that it deals with building trust in digital services.

## Experimental design

We ran an online experiment on the British platform Prolific with participants from the US. We conducted six between-subject treatments. We aimed at having roughly 250 participants (125 men and 125 women) in each treatment, for a total of 750 workers and 750 managers. We ended up collecting data from 744 workers and 754 managers.[5] We conducted experiments between March 21 and May 18, 2022. The experiment lasted an average of nine minutes, with an average payoff of £3.10.

At the start of the experiment, we asked for participants' gender, age, and their consent to participate in the study. Participants were in the role of either workers or managers. We ran a baseline and two treatments for participants in the role of workers, and a baseline and two treatments for participants in the role of managers.

### Treatments for workers

**Baseline treatment for workers (BaselineW)**

Workers first have two minutes to solve 12 real-effort exercises (task 1), and two minutes to solve 12 different real-effort exercises (task 2). They then solve what we call the job task for two minutes, which consists of seven exercises as in task 1 and five exercises as in task 2. They were paid according to their performance in one randomly chosen task among task 1, task 2, and the job task. Task 1 is the standard Raven Matrices test. Task 2 consists of counting zeros

[5]We recruited 260 participants per treatment, because we expected some participants would have to be excluded for clicking through the survey in less than one minute or not entering a decision in one of the main tasks.

in a 6x6 matrix. Workers earn £0.15 (15 pence) for each correct answer in the randomly determined payoff-relevant task.

After working on the tasks, workers are told that an algorithm and participants in the role of managers will have to make hiring decisions between pairs of workers. The algorithm and a manager choose which of the two workers to hire based on the two workers' gender and their task-1 and task-2 performances. We explain that the algorithm is trained to give the best prediction of the highest performer in the job task, based on the data from at least 200 workers, and that it hires the worker with the best predicted performance in the job task. We also state that managers are participants similar to them, but that they observe the task-1, task-2, and job-task performances as well as the gender of a subset of 20 random workers from the workers in the baseline. Managers get £2 if the job-task performance of the worker they chose to hire in one randomly chosen pair is higher than that of the other worker.

Workers must choose whether they prefer that the algorithm or a participant in the role of a manager make the hiring decision. Workers will get an additional payment of 50 pence if they were hired in the pair of workers randomly selected for payment. Note payments to workers are implemented only after the sessions with managers have been run.

We elicit participants' confidence in their relative performance in the job task. Participants can earn an additional 25 pence if they guess what percentage of workers has a lower performance than them, within a margin of error of five percentage points. We also elicit participants' beliefs about the gender composition of workers hired by the managers and by the algorithm. Given the equal number of men and women in the candidate pool, we ask how many of 100 hired workers they believe are men. Participants can get an additional 25 pence if their guess is not further away than five from the correct answer, for both managers and the algorithm making the hiring decisions. Lastly, we elicit beliefs about the gender composition of the best-performing workers, i.e., how many of the 50 best-performing workers will be men. As before, participants earn an additional 25 pence if their guess is no more than five away from the correct answer.

**Gender-blind algorithm treatment for workers (NoGenderW)**

The only difference from the BaselineW treatment is that the algorithm bases its hiring decisions on the task-1 and task-2 performances of the workers but not on their gender. Note managers still learn about the workers' gender.

**Transparency treatment for workers (TranspW)**

Relative to the BaselineW treatment, the only difference is that participants are given details about how the algorithm works before deciding whether they would prefer that the manager or the algorithm make the hiring decisions. More precisely, we disclose the regression equation that the algorithm employs to predict performance in the job task. The wording is the following:

> The algorithm calculates for at least 200 workers it has data on the mean relationship between the task-1 and 2 performances and gender on the one hand and the task-3 performance on the other hand. This relationship is:
>
> Task3 = 0.33*Task1 + 0.39*Task2 – 0.35*Male + 2.6
>
> so that, in order to predict someone's task-3 performance, one must replace, respectively, Task1 and Task2 with the task-1 and 2 performances of the person and deduct 0.35[6] only if the participant is male.

Note that we called the job task "task 3" in the instructions for the participants in order to keep the description as neutral as possible.

## Treatments for managers

**Baseline treatment for managers (BaselineM)**

We conducted the BaselineM treatment after the BaselineW treatment. The managers observe all questions in the three tasks that workers had to solve, but the managers did not have to solve them. The managers also observe the task-1, task-2, and job-task performances as well as the gender of a randomly determined set of 20 workers from the BaselineW and NoGenderW treatments. Hereafter, we refer to this random set of 20 workers as the training set. Every manager observed a different set of 20 workers randomly drawn from all workers of the BaselineW and NoGenderW treatments.

[6]A referee brought to our attention that the coefficient of the Male dummy was 0.34, not 0.35. We unintentionally reported to the participants a slightly higher coefficient than the one the algorithm actually used.

We ask the managers to make 20 hiring decisions from among pairs of workers from the BaselineW treatment. We generated pairs of workers such that every worker was a member of at least one pair. The total performance in task 1 and task 2 of the workers in a pair is similar (the difference does not exceed four for each task). We formed pairs in this manner because we did not want to make the hiring decisions too easy so that we would be able to observe any potential gender bias. In total, we created 600 such pairs. Of them, 10 pairs were randomly chosen and presented to the managers, whereas the other 10 pairs for each manager were selected only from among those pairs of workers whose performance difference did not exceed one and where the two workers were of a different gender. Again, we used this approach to ensure we could identify managers who favored workers of a particular gender, given similar performance. Hereafter, we refer to the set of 20 pairs of workers for which a manager has to make hiring decisions as the prediction set.

After the hiring decisions are made, we elicit the managers' belief regarding how often they chose the better worker in the 20 pairs of the prediction set, that is, the worker with the higher job-task performance. Participants earn an additional 25 pence if their guess is no more than one pair away from the correct answer. Finally, we ask the managers whether they want to delegate the hiring decisions to an algorithm. We tell them the algorithm is a computer program that chooses which of the two workers to hire based on the workers' gender and their performance in task 1 and task 2. We inform them that the algorithm is trained to predict who performs better in the job task, based on the data from at least 200 workers.7 For one randomly chosen hiring decision, the manager earns £2 if the decision is correct, meaning the worker who is hired performs better in the job task than the other worker of the pair. If a manager decided to delegate the hiring decisions to the algorithm, her payoff will depend on one randomly chosen hiring decision made by the algorithm.

**Confidence-feedback treatment for managers (ConfidM)**

In contrast to the BaselineM treatment, the managers in ConfidM receive feedback on the number of correct hires out of their 20 hiring decisions of the prediction set after the belief-elicitation stage and before they decide whether to delegate to the algorithm. Additionally, we inform them of whether they are overconfident (guessed at least two more correct hires than

[7]The algorithm's final training set contains 507 workers; see Table 2 for the exact OLS model and the details.

their actual performance), underconfident (guessed at least two fewer correct hires than their actual performance), or well calibrated (correct hires within an interval of +/-1 from stated).

**Transparency treatment for managers (TranspM)**

The only difference from the BaselineM treatment is that we provide managers with information about how the algorithm works before they decide whether to delegate the hiring decisions to the algorithm. The information they receive about the algorithm is the same as in the TranspW treatment.

We summarize the treatments for convenience in Table 1.

| Workers | Managers |
|---|---|
| **BaselineW**: Workers choose between algorithm or manager to make the hiring decisions. | **BaselineM**: Managers make 20 hiring decisions between pairs of workers. Then, they choose whether to delegate the hiring decisions to the algorithm. |
| **NoGenderW**: Same as BaselineW but algorithm only uses the workers' task-1 and task-2 performances, not their gender, to decide whom to hire. | **ConfidM**: Same as BaselineM, but before making the decision whether to delegate the hiring decision to the algorithm, manager is informed about the number of her correct hires and over- or underconfidence or correct guess. |
| **TranspW**: Same as BaselineW, but before choosing between algorithm and manager, workers are informed about the formula the algorithm uses to predict job performance. | **TranspM**: Same as BaselineM, but before deciding whether to delegate the hiring decision to the algorithm, managers are informed about the formula it uses to predict job performance. |

Table 1. Summary of main features of treatments

## Hypotheses

We start by presenting the pre-registered hypotheses concerning treatment differences, and then move to a pre-registered hypothesis that focuses on a correlation of interest.[8]

[8]The experimental design and hypotheses were pre-registered in the AEA RCT Registry, project number AEARCTR-0009068. We deviated from the pre-registered design by abandoning a treatment. The aim of this treatment was to see whether telling the workers that the managers hire fewer women than men (which is what we expected to happen) would increase their preference for the algorithm. However, we found that the managers hire more women than men (though to a lesser extent than the algorithm) and the favoring of women is optimal. For these reasons, we dropped our pre-registered hypotheses 1, 5, and 6 (which do not correspond to hypotheses labeled H1 and H5 in the manuscript), because they rely on the discrimination against women. Another important deviation from the pre-registration is that we now distinguish between the choice of the algorithm or delegation on the one

## Treatment differences

### Workers:

We distinguish between two concepts—a worker's preference for algorithmic hiring and algorithm aversion. A preference for algorithmic hiring corresponds to the choice of the algorithm over managers. Note that workers might opt against algorithmic hiring out of self-interest, not because of algorithm aversion. Algorithm aversion requires a suboptimal choice of managerial hiring.

**H1 (gender profiling):** A higher share of workers prefers to be hired by the algorithm rather than by managers in NoGenderW than in BaselineW. There is less algorithm aversion in NoGenderW than in BaselineW.

Support: Recent debates on potentially discriminatory algorithms due to gender and racial profiling suggest people have a preference against discrimination based on gender, independent of whether the discrimination is advantageous for them.[9] An alternative hypothesis would be that the preference for the algorithm or the managers depends on the workers' beliefs about the algorithm and the managers' favoring their own gender, pointing to self-serving preferences over algorithms.

**H2 (transparency for workers):** A higher share of workers prefers to be hired by the algorithm rather than by managers once the algorithm has been explained, that is, in TranspW relative to BaselineW. There is less algorithm aversion in TranspW than in BaselineW.

Support: Our hypothesis is based on the observation that algorithms, which are perceived as a black box, may be trusted less than transparent algorithms (von Eschenbach 2021, Kizilcec 2016). We therefore hypothesize that the preference for algorithmic hiring is stronger when the algorithm is transparent than when it is not.

---

hand and algorithm aversion on the other. The latter requires a suboptimal choice against the algorithm. Note that we also changed the labels of the hypotheses for clarity and conciseness.

[9] A recent example is the lawsuit and its settlement between the US and the Facebook owner Meta after the US Department of Justice said Meta encouraged housing advertisers to target users based on features such as race, religion, and sex, in violation of the Federal Housing Act. See https://www.reuters.com/legal/transactional/us-meta-settle-lawsuit-over-discrimination-housing-advertising-tool-2022-06-21/ last accessed 6.2.23. Another example are protests against racial profiling, which included white participants, for instance, in Detroit, https://eu.freep.com/story/news/local/michigan/detroit/2021/06/05/racial-profiling-protest-8-mile-detroit/7552406002/ last accessed 6.2.2023.

Note our experimental data show women perform better than men in the job task for equal performances at tasks 1 and 2. As a result, the algorithm favors women for equal task-1 and task-2 performances. Given this finding, an alternative hypothesis could be that women choose the algorithm more often and men less often in the TranspW treatment than in BaselineW, which would be in line with Kizilcec (2016), who observes an interaction between transparency and whether the algorithmic outcome meets the individual's expectations. Yet another possibility is that knowledge about discrimination in TranspW could turn workers away from the algorithm.

### Managers:

Just as for the workers, we distinguish between two concepts—delegation and algorithm aversion of the managers. Delegation is the managers' choice to use the algorithm's recommendation and override their own hiring decisions, whereas algorithm aversion is the decision not to delegate to the algorithm when such a choice would be optimal.

**H3 (managers' self-confidence):** More managers delegate the hiring decisions to the algorithm in ConfidM than in BaselineM. The effect is driven by those managers who are overconfident in the number of correct hires. There is less algorithm aversion in ConfidM than in BaselineM.

Support: Overconfidence is well documented in a variety of contexts (see Möbius et al., 2022, who measure confidence and observe how their participants update it upon receiving information). It is known to affect the market entry of firms (Camerer and Lovallo 1999), health and education decisions, decisions in the workplace (Dunning et al. 2004), and corporate investment (Malmendier and Tate 2005) and to mitigate the unraveling of matching markets (Dargnies et al. 2019). Although underconfidence has been documented as well (see, e.g., Dargnies et al. 2019), we hypothesize that in the context of hiring decisions, subjects will, on average, overestimate the quality of their decisions, which will lead to too little delegation to the algorithm.

**H4 (transparency for managers):** A higher share of managers delegate the hiring decisions to the algorithm when the algorithm is explained, i.e., in TranspM compared to BaselineM. There is less algorithm aversion in TranspM than in BaselineM.

Support: Similar to the workers and based on previous work mentioned in the context of H2, our hypothesis is motivated by the finding that people are reluctant to trust AI in general, especially when they do not understand how it works.

### Additional hypothesis

We pre-registered an additional hypothesis that is not based on treatment differences.

**H5 (discrimination in hiring):** For similar performances of men and women, managers are more likely to hire men than women. The difference is more pronounced for male managers.

Support: We base this hypothesis on an extensive literature in economics on gender discrimination (see Bertrand and Duflo, 2017, and Blau and Kahn, 2017 for reviews of this literature). Additionally, in contexts similar to that of our experiment, recent findings suggest male candidates are favored in hiring experiments. Sarsons et al. (2021) find that male recruiters are less likely to pick female candidates. Barron et al. (2022) observe that participants in the role of managers favor male candidates in a setup similar to ours if male and female candidates have similar performances in closely related tasks.

## Results

We start by investigating the performance of workers. We also present the hiring decisions made by the algorithm and the managers. Based on these findings, we then turn to studying the treatment effects on the workers' and managers' preference for the algorithm.

### Performance of workers and hiring decisions of the algorithm and managers

Unless otherwise stated, we use the data from all treatments in this section. The reason is that the task performance of workers and the hiring decisions of managers are expected to be unaffected by the treatments. The treatments differ only at a later stage, namely, right before the decision is taken regarding whether the manager's hiring decision is implemented or the algorithmic decision is followed.

In task 1 (Raven matrices), men perform better than women (4.04 vs 3.68; a two-sided Mann-Whitney test yields $p$=0.01), and no significant gender difference exists in performance for task 2 (6.19 vs 6.12; $p$=0.43) and the job task (6.14 vs 6.20; $p$=0.96).

Table 2 presents the results of the OLS regression where the job-task performance is the dependent variable. Models (1) and (2) present the results for the full sample. As seen in model (2), conditional on the task-1 and task-2 performances, men perform worse in the job task. Models (3) and (4) present the results for the BaselineW and NoGenderW treatments only, because we used the data from these treatments to generate the hiring algorithm in the BaselineW, TranspW, and NoGenderW treatments.[10] As the significant and negative coefficient of Male implies, the algorithm picks the woman when a man and a woman have the same performance in tasks 1 and 2. Given the same task-2 performance, the algorithm picks a woman even when her performance in task 1 is one point lower than a man's performance, which can be taken from the absolute value of the Male coefficient being larger than that of the task-1 coefficient in Model (4).

| | Job task (1) | Job task (2) | Job task (3) | Job task (4) |
|---|---|---|---|---|
| Task 1 | 0.342*** | 0.350*** | 0.326*** | 0.334*** |
| | (0.034) | (0.034) | (0.043) | (0.043) |
| Task 2 | 0.397*** | 0.395*** | 0.393*** | 0.391*** |
| | (0.025) | (0.025) | (0.031) | (0.031) |
| Male | | -0.211** | | -0.344*** |
| | | (0.104) | | (0.125) |
| Constant | 2.406*** | 2.490*** | 2.419*** | 2.574*** |
| | (0.164) | (0.169) | (0.201) | (0.207) |
| Observations | 744 | 744 | 507 | 507 |
| $R^2$ | 0.441 | 0.444 | 0.427 | 0.436 |
| Sample | All | All | BaselineW and NoGenderW | BaselineW and NoGenderW |

Notes: OLS regression. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Models 3 and 4 display the regression results used in the NoGenderW, TranspW, and TranspM treatments.

Table 2. Correlates of job-task performance

We used the following equation for the hiring decisions of the algorithm in the BaselineW and TranspW treatments:

$$\text{Jobtask} = 0.33*\text{Task1} + 0.39*\text{Task2} - 0.34*\text{Male} + 2.6.$$

For the NoGenderW treatment, we used the following equation to make the hiring decisions:

$$\text{Jobtask} = 0.33*\text{Task1} + 0.39*\text{Task2} + 2.4.$$

[10] We did not use the data from the TranspW treatment to develop the algorithm, because we needed to have the algorithm ready before this treatment to be able to disclose it to participants.

Among the pairs of workers that we created, 62.6% of the best-performing candidates are women. The algorithm ends up hiring 84.9% of women.[11] Managers hire 56.1% of women (54.1% for male managers and 57.7% for female managers; a two-sided Mann-Whitney test yields $p<0.01$, meaning female managers hire significantly more female workers than male managers). Whereas the managers hire a proportion of female workers (56.1%) that is closer to the proportion of best-performing candidates who are female (62.6%) than the hiring rate of women by the algorithm (84.9%), the algorithm is more efficient in hiring the best-performing workers: 66.9% of hiring decisions made by the algorithm are correct, whereas only 55.9% of the managers' hiring decisions are correct (a two-sided Mann-Whitney test yields $p<0.01$).

The mean task-1 performances of men and women hired by the managers are, respectively, 4.02 and 3.87 (a two-sided Mann-Whitney test yields, $p<0.01$). The mean task-2 performances of men and women hired by the managers are, respectively, 6.17 and 6.00 (a two-sided Mann-Whitney test yields, $p<0.01$). Thus, managers require higher task-1 and -2 performances from male workers than from female workers. Managers act in line with the fact that the women's performance in the job task tends to be higher than that of male workers for given task-1 and task-2 performances.

We investigate further what drives the managers' decisions when choosing between workers in each pair. More specifically, we are interested in how gender and performance differences affect the hiring decisions. Table 3 presents the results of probit regressions.[12] In models (1) and (2), we regress a dummy equal to 1 if the manager chose to hire the first worker of the pair (the first and second worker of the pair is randomly determined) on a variable equal to the difference between the male dummies of the two workers of the pair,[13] the difference in performance between the two workers for task 1 and task 2, and in model (2), the interaction between the first variable and the gender of the manager. Models (3) and (4) use the same variables as model (1), but the dependent variables are, respectively, a dummy indicating whether the first worker of the pair is the one with the higher job-task performance and a dummy

[11] The high percentage of women being hired is due to the fact that we match workers with similar performances, which even leads to more women being hired than is optimal. If we randomly matched workers into pairs, 53.1% of the algorithm's hires would be women (based on 1,000 simulations for each worker in the sample).
[12] For this and all other probit regressions in the paper, we also document the OLS regressions in Appendix A (online). The signs and level of significance of all variables are the same in both specifications.
[13] This dummy is equal to 1 if the first worker is male and the second is female, 0 if both workers are of the same gender, and -1 if the first worker is female and the second is male. Therefore, a negative marginal effect of this variable can be interpreted as female workers being favored in the hiring decisions.

for whether the algorithm hires the first worker of the pair. Thus, model (3) presents the marginal effects for optimal hiring.

Given the task-1 and task-2 performances, female workers are favored over male workers in optimal decisions, see model (3). We observe this for both the managers and the algorithm, as can be seen from the negative marginal effect of "1st worker male minus 2nd worker male" in models (1) and (4). While this is not surprising for the algorithm, for the managers it suggests that they learn that female workers have a higher job-task performance than male workers for given task-1 and task-2 performances from observing the performances of the training set of 20 workers.

| | 1st worker of the pair hired by manager (1) | 1st worker of the pair hired by manager (2) | 1st worker of the pair is the correct hire (3) | 1st worker of the pair is hired by algorithm (4) |
|---|---|---|---|---|
| 1st worker male minus 2nd worker male | -0.063*** | -0.086*** | -0.147*** | -0.111*** |
| | (0.006) | (0.009) | (0.036) | (0.01) |
| Task1 of 1st worker minus 2nd worker | 0.211*** | 0.211*** | 0.102*** | 0.078*** |
| | (0.009) | (0.009) | (0.02) | (0.007) |
| Task2 of 1st worker minus 2nd worker | 0.163*** | 0.162*** | 0.076*** | 0.098*** |
| | (0.008) | (0.008) | (0.025) | (0.009) |
| 1st worker male minus 2nd worker male * male manager | | 0.049*** | | |
| | | (0.013) | | |
| Observations | 15080 | 15080 | 15080 | 15080 |
| Clustered errors | Manager | Manager | Pair | Pair |
| Sample | All | All | All | All |

Notes: Marginal effects of probit regression of dummy for hiring the first worker of a pair by the manager or by the algorithm. "1st worker male minus 2nd worker male" is the difference between the Male dummies corresponding to each worker of the pair. "Task1 of 1st worker minus 2nd worker" is the difference in performance between the two workers for task 1. "Task2 of 1st worker minus 2nd worker task 2" is the difference in performance between the two workers for task 2. "1st worker male minus 2nd worker male * male manager" is the interaction between "1st worker male minus 2nd worker male" and the gender of the manager. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3. Determinants of hiring by managers and by the algorithm

Hypothesis 5, which states that managers favor male workers, is not validated. However, male managers favor female workers to a lesser extent than female managers. This finding can be taken from the positive marginal effect of the interaction term of model (2). Our data therefore support the second part of Hypothesis 5 according to which male managers favor male workers relatively more than female managers. Note that not favoring female candidates enough leads

to somewhat fewer correct hires by male managers (55% for male managers and 56.7% for female managers; the difference is marginally significant, $p$=0.053, based on the marginal effect in a regression with clustered errors at the subject level).

We sum up the main findings of this section:

***Result 1 (Workers' performance and hiring decisions):*** *Conditional on task-1 and task-2 performances, female workers outperform male workers at the job task. Managers favor female candidates, conditional on the workers' performance in tasks 1 and 2. Male managers tend to give a significantly smaller premium to female candidates than female managers, which results in a smaller percentage of correct hires (marginally significant). The proportion of correct hires by the algorithm is significantly higher than that by the managers.*

Next, we present the main results of the experiments. We start with the workers and then move on to the managers.

## Workers: Choice of hiring algorithm

We investigate the workers' choice of the algorithm over the managers and the drivers of this choice. We distinguish between two concepts—a preference for algorithmic hiring and algorithm aversion. Note workers might choose managerial hiring out of self-interest, not because of algorithm aversion. Algorithm aversion requires a suboptimal choice of managerial hiring.[14] The main focus of this section is on treatment differences in the preference for algorithmic hiring. We present the results for algorithm aversion at the end of the section.

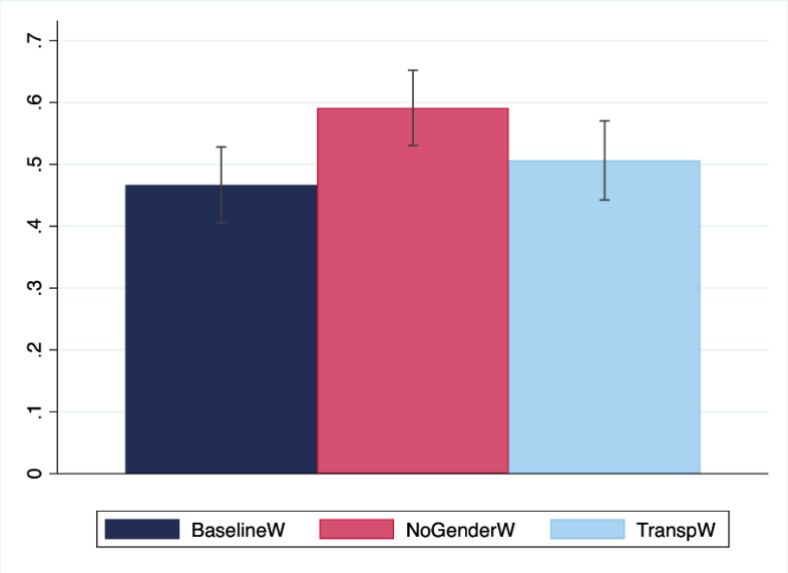
Figure 1 presents the proportion of workers who chose the algorithm by treatments. In the baseline, 46.67% of workers prefer that the algorithm rather than managers make the hiring decisions. The proportion of workers choosing the algorithm is significantly higher in NoGenderW than in BaselineW (59.13% vs. 46.67%; a two-sided Mann-Whitney test yields $p$<0.01), which indicates a reluctance of workers to be subject to an algorithm that bases its hiring decisions on gender. For TranspW, the proportion of workers choosing the algorithm is not significantly different from BaselineW (50.63% vs 46.67%, a two-sided Fisher's exact test

[14] A stronger test of algorithm aversion would require that workers opt against the algorithm despite believing that the algorithm gives them a higher expected payoff. We cannot study this because we did not elicit the beliefs about the relative chances of being hired by managers and the algorithm.

yields $p$=0.42) and is marginally lower than in NoGenderW ($p$=0.07, two-sided Fisher's exact test). Thus, we find support for H1 and no support for H2.

The absence of a strong impact of transparency on the preference for the algorithm is puzzling and may be due to several factors. First, it may be indicative of the challenge to convey information about how algorithms work. In our experiment, we nonetheless find the belief in the proportion of men among the workers hired by the algorithm is slightly lower (a two-sided Mann-Whitney test yields $p$=0.02) in TranspW (50.6) than in BaselineW (53). Therefore, some workers seem to have been able to understand from the formula that the algorithm discriminates against men.

Second, because women in our dataset improve their performance between tasks 1 and 2 and the job task more than men, the algorithm favors women for given task-1 and -2 performances. Participants in the role of workers may find this favoring of women unacceptable. Further research is needed to investigate whether transparency has an effect on the preference for the algorithm if the algorithm does not discriminate by gender.




Notes: Black lines correspond to 95% confidence intervals.

Figure 1: Proportion of workers who chose the algorithm by treatments

Beyond the overall treatment effects, we investigate treatment differences depending on the gender of workers, their confidence in their relative performance, and their beliefs about which gender each hiring process favors. Table 4 presents the marginal effects of probit regressions of the choice of the algorithm – the dummy being equal to 1 if the worker prefers that the

algorithm, rather than managers, make the hiring decisions. In model (1), we regress the choice of the algorithm on the two treatment dummies. Models (2) to (5) add additional controls. These regressions serve to determine whether the choice of the algorithm correlates with the workers' performance, confidence, and beliefs about which of the hiring processes is more favorable to male workers. The workers' age and performances in tasks 1 and 2 do not correlate with the choice of the algorithm. In model (4), we add the variable "Confidence," which measures the belief regarding the proportion of workers who have a lower sum of task-1 and 2 performances than oneself. A significant correlation exists between the workers' confidence in their performance and their choice of the algorithm.[15] The better the workers think they performed, the more likely they are to choose the algorithm.

| | Choice of algorithm (1) | Choice of algorithm (2) | Choice of algorithm (3) | Choice of algorithm (4) | Choice of algorithm (5) |
|---|---|---|---|---|---|
| NoGenderW | 0.125*** | 0.126*** | 0.121*** | 0.132*** | 0.141*** |
| | (0.044) | (0.044) | (0.044) | (0.044) | (0.043) |
| TranspW | 0.040 | 0.044 | 0.040 | 0.040 | 0.054 |
| | (0.045) | (0.045) | (0.045) | (0.045) | (0.044) |
| Age | | -0.001 | -0.000 | -0.000 | -0.000 |
| | | (0.001) | (0.001) | (0.001) | (0.001) |
| Male | | 0.057 | 0.053 | 0.030 | 0.076** |
| | | (0.036) | (0.036) | (0.037) | (0.038) |
| Task 1 | | | 0.010 | 0.002 | -0.000 |
| | | | (0.012) | (0.013) | (0.012) |
| Task 2 | | | 0.006 | 0.001 | 0.003 |
| | | | (0.009) | (0.009) | (0.009) |
| Confidence | | | | 0.003*** | 0.003*** |
| | | | | (0.001) | (0.001) |
| DiffBeliefAlgo Manager | | | | | -0.008*** |
| | | | | | (0.002) |
| Male* DiffBeliefAlgo Manager | | | | | 0.012*** |
| | | | | | (0.003) |
| Observations | 744 | 744 | 744 | 744 | 743 |
| Sample | All | All | All | All | All |

Notes: Marginal effects of probit regression of choosing the algorithm by workers. "Confidence" is the belief regarding how many workers out of 100 have lower task-1 and -2 performances than oneself. "DiffBeliefAlgoManager" equals the difference between beliefs regarding how many men are hired by the

[15] Self-confidence is significantly correlated with actual performance (rho=0.37). Men are more confident than women about their task-1 and task-2 performances. On average, men and women respectively believe 53.4% and 44.8% (a two-sided Mann-Whitney test yields $p<0.01$) of workers have a lower sum of task-1 and task-2 performances than themselves.

algorithm minus how many men are hired by managers. "Male*DiffBeliefAlgoManager" is the interaction of DiffBeliefAlgoManager and the dummy for the worker being male. Standard errors in parentheses, and * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 4. Determinants of the workers' choice of the algorithm

We also collected the beliefs of workers concerning the proportion of men who would be hired by the managers and by the algorithm, respectively, from a gender-balanced pool of workers. Workers believe that managers will favor men more than the algorithm: on average, workers believe 55.7% of managers' hires will be men and that 51.8% of the algorithm's hires will be men (a t-test yields p<0.01). . Women, compared to men, hold a stronger belief that both managers and the algorithm have a preference towards hiring men. On average, men estimate that 54.8% of the hires made by managers will be men, while women estimate this proportion to be 56.7%; a two-sided Mann-Whitney test yields p<0.01. Regarding the belief about the proportion of male hires made by the algorithm, it was found that men, on average, estimated it to be 51.1%, while women estimated it to be 52.5%; a two-sided Mann-Whitney test yields p=0.0496. Model (5) controls for how many more men a worker believes will be hired by the algorithm as opposed to the managers, with the variable "DiffBeliefAlgoManager" and with an interaction of this variable with the Male dummy. Female workers are all the more inclined to choose the algorithm when they perceive it to favor them more than their managers do. Male workers have the analogous tendency: they are more likely to choose the algorithm when they believe it favors male workers more strongly than the managers do; although this difference is only marginally significant (a Wald test of significance of "DiffBeliefAlgoManager"+"Male*DiffBeliefAlgoManager" yields $p$=0.08). Thus, we observe a weaker preference for the algorithm of those workers who believe the algorithm discriminates against their gender or that the managers favor their gender. Note that when controlling for the beliefs about gender-related hiring by the algorithm compared with the managers, the male dummy becomes significant, pointing to a higher tendency of male workers to choose the algorithm when they believe the managers and the algorithm favor male workers equally.[16]

Table A1 in the Appendix presents analyses of treatment differences by gender. Model (1) replicates model (5) of Table 4. Model (2) includes interactions of the Male dummy and the treatment dummies. None of the treatment differences remain significant at the 5% level,

[16] Including triple interactions of DiffBeliefAlgoManager, a male dummy, and each of the treatment dummies yields non-significant marginal effects. Thus, we do not find stronger treatment effects for workers who believe that the algorithm discriminates against their gender.

potentially due to lack of power. However, the marginal effects of NoGenderW and NoGenderW*Male are relatively large, suggesting the treatment effect is larger for male workers. Indeed, by splitting the sample into only male (model (3)) and only female workers (model (4)), we confirm male workers are primarily responsible for the increase in the preference for algorithmic hiring when gender is removed. This finding is consistent with male workers expecting the algorithm to favor women. For the TranspW treatment, the gender differences in treatment effects are small. This finding is puzzling because male workers could be expected to choose the algorithm less often because of its explicit favoring of female workers. One possible explanation is that the algorithmic formula reveals the importance task-1 and 2 performances. Given that the male workers are significantly more confident than female workers (a t-test yields $p<0.01$), the transparency regarding the important role of performance in the algorithm might outweigh the algorithm's favoring the female workers.

Next, we analyze algorithm aversion. We ask how close the workers' choices are to the empirical optimum; that is, whether they take payoff-maximizing decisions at the individual level. Whereas the algorithm hires more efficiently, the workers might prefer managers to make the hiring decisions if they think the managers are more likely than the algorithm to hire them. To analyze the optimality of the workers' decisions, we simulate hiring by managers based on model (1) of Table 3 and based on the algorithm. For each worker, we simulate pairs of workers from the entire pool of workers. If a given worker is hired more often by the algorithm than by the managers, we say choosing the algorithm is the optimal decision for the worker.[17]

Overall, only 50.5% of workers (54% of women and 47% of men) make the optimal, that is, individual-payoff-maximizing, choice between the algorithm and the managers. The gender difference is driven by BaselineW, where the proportion of optimal choices is significantly smaller for male than for female workers (a two-sided Fisher's exact test yields $p=0.02$). The difference is not surprising, because male workers choose the algorithm as often as female workers, but the algorithm favors the latter. In both NoGenderW and TranspW, there is no gender difference in the optimality of hiring choices (two-sided Fisher's exact tests yield $p=1.00$ and $p=0.44$, respectively). In NoGenderW, the female workers cannot be favored, thus

[17]The simulation of hiring decisions based on the managers' model is warranted because we observe only few, and sometimes one single, hiring decision(s) per worker. The hiring outcome depending only on one or very few manager and worker matches should not determine the optimality of the preference of one of the two hiring processes ex ante.

removing the advantage of the algorithm for female workers. In TranspW, although the algorithm still favors female workers, male workers have the opportunity to learn this and make better choices between the managers and the algorithm.

Figure 2 presents the workers' decisions to choose the algorithm depending on whether or not doing so was optimal. Among all workers, 24.3% are algorithm averse, because they chose managerial hiring despite being more likely to be hired by the algorithm (light blue bars in Figure 2). At the same time, 25.1% of workers are algorithm loving because they chose the algorithm despite being more likely to be hired by the managers (dark gray bars in Figure 2). Thus, we document algorithm aversion of about one quarter of workers and investigate whether this share is affected by the two treatments. Table A2 in the Appendix presents the marginal effects of probit regressions of algorithm aversion – the dummy being equal to 1 if the worker prefers the manager despite a higher chance of being hired by the algorithm. Workers are significantly less algorithm averse in NoGenderW than in the baseline, whereas no significant difference exists between TranspW and BaselineW, controlling for other variables. These relationships are illustrated in Figure 2.

The reduction of algorithm aversion in NoGenderW could either mean an increase in optimal choices or an increase in the proportion of algorithm-loving workers. The former is the case: the proportion of optimal choices is marginally significantly higher in NoGenderW than in BaselineW (53.5% vs. 45.4%, a two-sided Fisher exact test yields $p$=0.08). Note that men are significantly less algorithm averse than women when pooling all treatments (a two-sided Fisher exact test yields $p$<0.01). This is in part driven by the fact that the algorithm favors women, which makes choosing the algorithm less often optimal for men (the algorithm is the optimal choice for 62% of women and 40.6% of men; a two-sided Fisher exact test yields $p$<0.01).

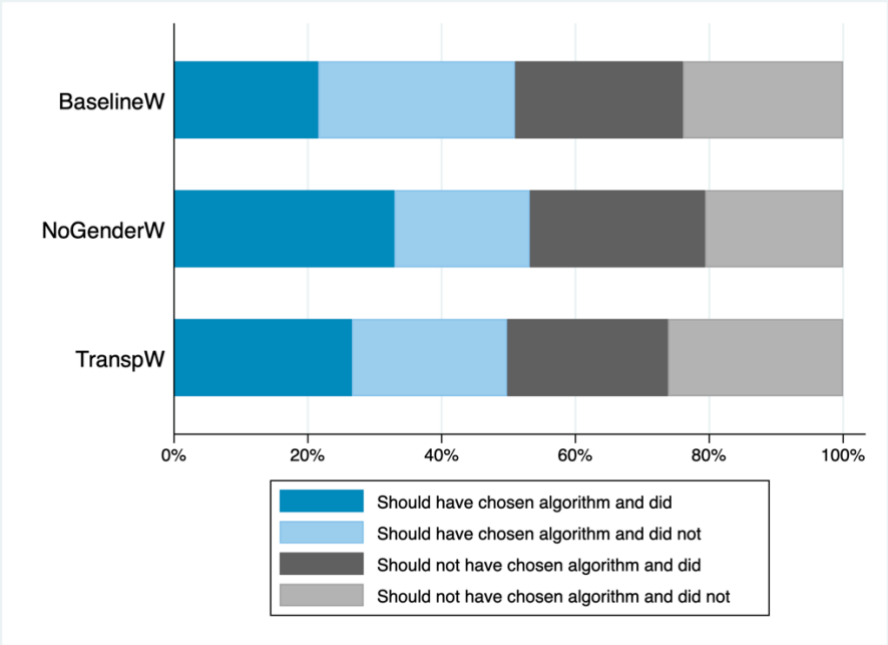



Figure 2. Workers' choice of algorithm depending on whether choosing the algorithm was optimal or not

We sum up the main results concerning the workers' choice of the algorithm and algorithm aversion:

***Result 2 (Workers' preference for the algorithm and algorithm aversion):***
*Workers are more likely to choose the algorithm that is gender blind and are less algorithm averse toward a gender-blind algorithm than an algorithm using gender. Disclosure of the details of the algorithm neither affects the preference for algorithmic hiring nor algorithm aversion.*

## Managers: Delegation to the algorithm

Our main interest is again in the treatment differences, now regarding the managers' decisions of whether to delegate the hiring decision to the algorithm. We first present relevant descriptive statistics of the managers' hiring decisions. On average, managers made 11.2 correct hiring decisions out of the 20 in their prediction set (11.0 for men and 11.3 for women; a two-sided Mann-Whitney test yields $p$=0.056). On average, managers believe they made 11.5 correct hiring decisions out of 20 (11.7 for men and 11.3 for women; a two-sided Mann-Whitney test yields $p$=0.02). Thus, we observe significant overconfidence of men (a t-test yields p<0.01) but

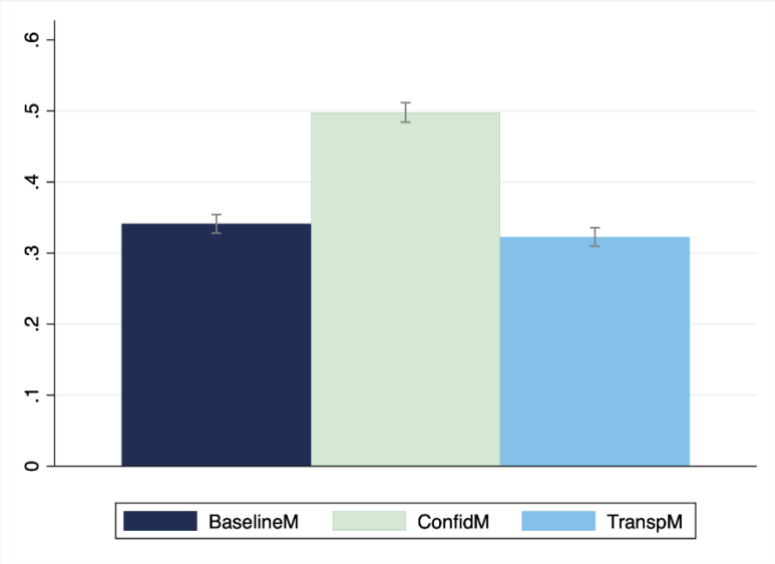
not of women (a t-test yields $p$=0.66), where overconfidence is defined as the difference between the believed and the actual number of correct hiring decisions.[18] In contrast, we will refer to confidence as the managers' belief in how many hiring decisions were correct.

We next investigate the managers' choice to delegate to the algorithm. We distinguish between two concepts—delegation and algorithm aversion by managers.[19] Delegation is the managers' choice to use the algorithm's recommendation and override their own hiring decisions, whereas algorithm aversion is the decision not to delegate to the algorithm when such a choice would be optimal. We start with delegation. Figure 3 presents the proportion of managers delegating the hiring decision to the algorithm, separately for each treatment.

In the baseline, 34.1% of managers chose to delegate the hiring decisions to the algorithm. The proportion of delegating managers is significantly higher in ConfidM at 49.8% (a two-sided Mann-Whitney test yields $p$<0.01), which demonstrates a causal effect of correcting the managers' beliefs regarding their hiring decisions on their delegation decision. For TranspM, the proportion of managers delegating to the algorithm is 32.2%, which is not significantly different from BaselineM (a two-sided Fisher's exact test yields $p$=0.71) and significantly lower than in ConfidM (a two-sided Fisher's exact test yields $p$<0.01). Thus, we find support for H3 and no support for H4.

[18]Overconfidence in the psychology and behavioral economics literature refers to various different phenomena. In terms of the terminology suggested by Moore and Healy (2008), overconfidence in our context is an overestimation of one's ability which they call overoptimism.

[19] As in the case of workers, a stronger version of the test for algorithm aversion would require managers to opt against the algorithm despite believing that the algorithm is better at selecting workers. We cannot study this because we did not elicit the beliefs about the number of correct hires by the algorithm.




Notes: Black lines correspond to 95% confidence intervals.

Figure 3: Proportion of managers who delegate the hiring decision to the algorithm by treatment

Table 5 presents the marginal effects of probit regressions of the delegation decisions. Overall, the results of the regressions confirm that providing feedback on their performance increases the managers' delegation of the hiring decisions to the algorithm. In BaselineM and TranspM, overconfidence is negatively correlated with delegation, as seen in model (3). This finding indicates managers who overestimate their hiring success are less likely to delegate the decision to the algorithm. The interaction term of model (3) shows higher overconfidence is associated with a significantly stronger treatment effect of ConfidM. The marginal effects are robust to controlling for the actual number of correct hires by the managers (see model (4)), which ensures the worst or best performers, who are more likely to be over- or underconfident, respectively, do not mechanically drive the effect of overconfidence. Finally, analogous to the transparency treatment for the workers, the regressions confirm providing details about how the algorithm works does not increase the managers' delegation to the algorithm.

| | Delegation (1) | Delegation (2) | Delegation (3) | Delegation (4) |
|---|---|---|---|---|
| ConfidM | 0.157*** | 0.155*** | 0.130*** | 0.137*** |
| | (0.043) | (0.043) | (0.044) | (0.044) |
| TranspM | -0.019 | -0.020 | -0.045 | -0.043 |
| | (0.042) | (0.042) | (0.043) | (0.043) |
| Age | | -0.000 | -0.000 | -0.000 |
| | | (0.000) | (0.000) | (0.000) |
| Male | | 0.043 | 0.044 | 0.045 |

| | | | | |
|---|---|---|---|---|
| | | (0.035) | (0.035) | (0.035) |
| Overconfidence | | | -0.018*** | -0.016*** |
| | | | (0.006) | (0.006) |
| ConfidM*Overconfidence | | | 0.034*** | 0.034*** |
| | | | (0.009) | (0.009) |
| Number of correct hires | | | | 0.007 |
| | | | | (0.009) |
| Observations | 754 | 754 | 752 | 752 |
| Sample | All | All | All | All |

Notes: Marginal effects of probit regression of delegation to algorithm. "Overconfidence" is the difference between the belief regarding how many hires were correct and the actual number of correct hires. "ConfidM*Overconfid" is the interaction of Overconfid and the dummy for treatment ConfidM. "Number of correct hires" is the number of pairs in which the manager hired the worker with the higher task-3 performance. Standard errors in parentheses, and * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 5. Determinants of delegation to the algorithm by managers

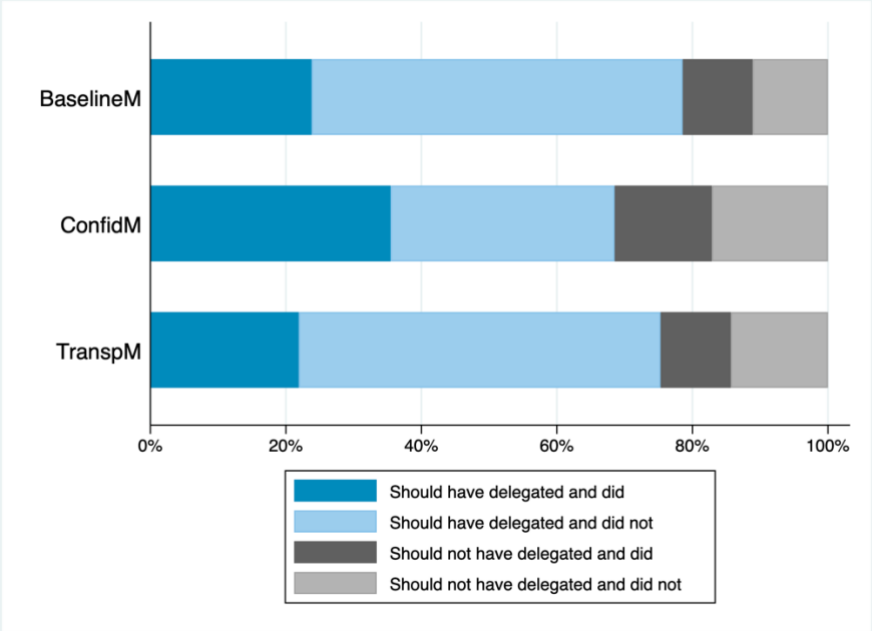



Figure 4. Managers' delegation decisions depending on whether delegating was optimal or not

We next study algorithm aversion by the managers. Because managers differ in their ability to hire the better worker and in their confidence level, we first ask whether managers optimally sort into delegation and whether our treatments impact the optimality of delegation decisions. Figure 4 shows the breakdown of the managers' decisions to delegate depending on whether or not doing so was optimal. The proportion of optimal delegation decisions by treatment (i.e., the sum of the left-most and right-most parts of the bars in Figure 4) is the following: 34.9%, 52.6%, and 36.3% in BaselineM, ConfidM, and TranspM, respectively. The proportion in ConfidM is significantly higher than in Baseline and TranspM (a two-sided Mann-Whitney test yields

p<0.01 for both comparisons). More managers in ConfidM than in BaselineM should have delegated and did so (a two-sided Fisher's exact test yields $p<0.01$), and should not have delegated and did not do so (a two-sided Fisher's exact test yields $p=0.06$), as indicated by the longer dark blue and light gray bars for ConfidM in Figure 4. Finally, the managers who should have delegated and did not (light blue bar) are algorithm averse. The proportions of algorithm-averse managers are 54.8% in BaselineM, 33.1% in ConfidM, and 53.4% in TranspM. The difference between ConfidM and BaselineM is significant (a two-sided Fisher's exact test yields p<0.01). Thus, providing feedback on managers' over- or underconfidence significantly increases the optimality of delegation decisions and significantly decreases algorithm aversion. Finally, none of the differences between TranspM and BaselineM are significant.

We sum up the main results concerning the managers' choices:

***Result 3 (Managers' delegation to the algorithm and algorithm aversion):*** *Managers delegate to the algorithm more often andmore often when it is optimal, and are less algorithm averse when they receive feedback on their performance, an effect that increases in strength with the manager's confidence. Disclosure of the algorithm does not increase delegation to the algorithm nor the optimality of the delegation decisions.*

## Conclusion

We designed an online experiment to shed light on the determinants of preferences for algorithmic hiring decisions from two distinct perspectives—workers and managers. For workers, we find a substantial increase in the preference for algorithmic hiring when the algorithm is gender blind. This result points to a preference for no discrimination (advantageous or disadvantageous) based on gender. We interpret it as direct support for regulations that would make illegal any profiling by ethnicity, gender, or other group attributes.

For managers, we replicate the finding in the literature that managers delegate decisions to the algorithm too rarely. Overconfidence in their ability to hire the better worker causes managers to make this costly mistake. Providing managers with feedback on the quality of their hiring decisions increases both the frequency and optimality of delegation decisions. The former finding is in line with Glaeser et al. (2021), who show that inspectors, deciding on which

restaurant to inspect, use the recommendations of the algorithm only about half of the time, despite a substantial potential gain in efficiency. Mandating the adoption of the algorithm is one way to increase efficiency (as suggested by Glaeser et al., 2021) but we show feedback on past performance can increase efficiency through the voluntary adoption of the algorithm.

Interestingly, we find no effect of transparency in the form of disclosure of the algorithm on its adoption. This lack of a finding suggests regulating the transparency of algorithms alone is unlikely to affect the preference for algorithmic decision-making. However, we do not claim our findings speak against transparency per se, because its goal can, for example, be to monitor compliance with a no-profiling requirement. Moreover, our algorithm is straightforward, which might dilute any positive effect of transparency. Participants could perceive it as too simple and therefore believe it is inefficient. Furthermore, our data are peculiar because women improve their performance between tasks 1 and 2 and the job task more than men. Therefore, the algorithm favors women for given task-1 and -2 performances. By making this explicit in the transparency treatment, two effects may cancel each other out: some participants in the role of workers and managers may find the discrimination against men unacceptable, whereas others, especially women, may perceive it as acceptable. Finally, understanding the presentation of the algorithm in our transparency treatments may not be straightforward for some participants. We display the formula used by the algorithm, and we also put it into words. Experimenting, for example, with graphical representations of algorithms, or providing information about the performance of the algorithm could be an avenue for future research.

A final caveat relates to the stylized character of the tasks employed in the experiments. Our environment is simple in the sense that the job task is a combination of two tasks for which performance can be perfectly observed ex ante (before hiring a worker) and in that performance in the job task is perfectly observable ex post. The simplicity of the environment may overstate the advantages of hiring algorithms relative to environments where the demands on workers are more complex, for example. However, we note that although the observed relative performance of algorithms compared with managers may not be externally valid, also due to inexperienced subjects in the role of managers, we expect the treatment effects regarding attitudes toward algorithms to be independent of these levels and therefore informative.

## References


Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. Government Information Quarterly 37.4, 101490.

Aridor, G., Che, Y.-K., & Salz, T. (2020). The economic consequences of data privacy regulation: Empirical evidence from GDPR. Cambridge, MA, USA: National Bureau of Economic Research.

Avery, M., Leibbrandt, A., & Vecci, J. (2023). Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech. Mimeo

Auster, S. & Pavoni, N. (forthcoming). Optimal delegation and information transmission under limited awareness. Theoretical Economics.

Badakhshan, E., Humphreys, P., Maguire, L., & McIvor, R. (2020). Using simulation-based system dynamics and genetic algorithms to reduce the cash flow bullwhip in the supply chain. International Journal of Production Research, 58(17), 5253-5279.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society 54(6), 627-635.

Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2019). The impact of big data on firm performance: An empirical investigation. AEA Papers and Proceedings 109, 33-37.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 671-732.

Barron, K., Ditlmann, R., Gehrig, S., & Schweighofer-Kodritsch, S. (2022). Explicit and implicit belief-based gender discrimination: A hiring experiment. Mimeo.

Bertrand, M. & Duflo, E. (2017). Field experiments on discrimination. In A. Banerjee and E. Duflo (Eds.), Handbook of Field Experiments. North Holland.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. Cognition 181, 21-34.

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2022). Algorithmic discrimination causes less moral outrage than human discrimination. Journal of Experimental Psychology: General. Advance online publication. https://doi.org/10.1037/xge0001250

Blau, F. D. & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. Journal of Economic Literature 55(3), 789–865.

Bolton, G.E., Katok, E. & Ockenfels, A. (2004). How effective are online reputation mechanisms? An experimental study. Management Science 50(11), 1587-1602.

Camerer, C., & Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. American Economic Review 89(1), 306-318.

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

Camuffo, A., Cordova, A., Gambardella, A. & Spina, C. (2020). A scientific approach to entrepreneurial decision making. Management Science 66 (2), 564–586.

Castelo, N., M. W. Bos & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. Journal of Marketing Research 56(5), 809-825.

Chalfin, Aaron, Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J. & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. American Economic Review 106.5, 124-27.

Charles, K. K. & Guryan, J. (2011). Studying discrimination: Fundamental challenges and recent progress. Annual Review of Economics 3(1), 479–511.

Chugunova, M. & Luhan, W. J. (2022). Ruled by robots: Preference for algorithmic decision makers and perceptions of their choices. Max Planck Institute for Innovation & Competition Research Paper 22-04.

Corgnet, B. (2023) An experimental test of algorithmic dismissals. Mimeo

Cowgill, B. (2020). Bias and productivity in humans and algorithms: Theory and evidence from resume screening. Columbia Business School, Columbia University, *29*.

Cowgill, B., Dell'Acqua, D., & Matz, S. (2020a). The Managerial Effects of Algorithmic Fairness Activism. AEA Papers and Proceedings 110, 85-90.

Cowgill, B. Dell'Acqua, F., Deng, S., Hsu, D., Verma, N., & Chaintreau, A. (2020b). Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics. In Proceedings of the 21st ACM Conference on Economics and Computation, 679-681.

Cowgill, B., & Tucker, C. (2020). Algorithmic Fairness and Economics. Mimeo.

Danz, D., Kübler, D., Mechtenberg, L., & Schmid, J. (2015). On the failure of hindsight-biased principals to delegate optimally. Management Science 61 (8), 1938-1958.

Dargnies, M. P., Hakimov, R., & Kübler, D. (2019). Self-confidence and unraveling in matching markets. Management Science 65(12), 5603-5618.

Dietvorst, B. J, Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General 144.1, 114-126.

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

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. Psychological Science in the Public Interest 5(3), 69-106.

Ertac, S., Gumren, M., & Gurdal, M. Y. (2020). Demand for decision autonomy and the desire to avoid responsibility in risky environments: Experimental evidence. Journal of Economic Psychology, 77, 102200.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. Interfaces 37(6), 570-576.

Fumagalli, E., Rezaei, S., & Salomons, A. (2022). OK computer: Worker perceptions of algorithmic recruitment. Research Policy, 51(2), 104420.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI magazine 38(3), 50-57.

Glaeser, E. L., Hillis, A., Kim, H., Kominers, S. D. & Luca, M. (2021). Decision authority and the returns to algorithms. Harvard Business School Working Paper.

Greiner, Ben, Philipp Grünwald, Thomas Lindner, Georg Lintner, & Martin Wiernsperger (2022). Incentives, Framing, and Trust in AI: An experimental study. Mimeo.

Grigoryan, Aram. (2022) Transparency in Allocation Problems. Mimeo

Hajian, S., & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. IEEE transactions on knowledge and data engineering, 25(7), 1445-1459.

Hakimov, R., & Raghavan, M. (2023). Improving Transparency and Verifiability in School Admissions: Theory and Experiment. Mimeo

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. Industrial and Organizational Psychology 1(3), 333-342.

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

Kaibel, C., Koch-Bayram, I., Biemann, T., & Mühlenbock, M. (2019). Applicant perceptions of hiring algorithms - uniqueness and discrimination experiences as moderators. In Academy of Management Proceedings 2019(1), 18172.

Kimbrough, S. O., Wu, D. J., & Zhong, F. (2002). Computers play the beer game: can artificial agents manage supply chains? Decision support systems 33 (3), 323-333.

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2390-2395.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The Quarterly Journal of Economics 133(1), 237-293.

Lambrecht, A. & Tucker, C.E. (2019). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. Management Science 65(7), 2966-2981.

Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. European Economic Review 90, 375–402.

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

Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment. The Journal of Finance 60(6), 2661-2700.

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. Management Science 68(11), 7793-8514.

Persson, A. (2016, August). Implicit bias in predictive data profiling within recruitments. In IFIP international summer school on privacy and identity management, 212-230.

Sanders, N. R. & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. Omega 31(6), 511-522.

Sarsons, H., Gërxhani, K., Reuben, E. & Schram, A. (2021). Gender Differences in Recognition for Group Work. Journal of Political Economy 129(1), 101-147.

Von Eschenbach, W.J. (2021). Transparency and the black box problem: Why we do not trust AI. Philos. Technol. 34, 1607–1622.

Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. Professional Psychology: Research and Practice 40(5), 525.

Will, P., Krpan, D. & Lordan, G. (2022). People versus machines: introducing the HIRE framework. Artificial Intelligence Review 1-30.

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. Journal of Behavioral Decision Making 32(4), 403-414.

## Appendix

## Additional Tables

| | Choice of algorithm (1) | Choice of algorithm (2) | Choice of algorithm (3) | Choice of algorithm (4) |
|---|---|---|---|---|
| NoGenderW | 0.141*** | 0.104* | 0.181*** | 0.098 |
| | (0.043) | (0.062) | (0.060) | (0.062) |
| TranspW | 0.054 | 0.051 | 0.043 | 0.053 |
| | (0.044) | (0.062) | (0.064) | (0.061) |
| Age | -0.000 | -0.000 | 0.001 | -0.001 |
| | (0.001) | (0.001) | (0.002) | (0.002) |
| Male | 0.076** | 0.049 | | |
| | (0.038) | (0.062) | | |
| Task 1 | -0.000 | 0.000 | 0.008 | -0.015 |
| | (0.012) | (0.012) | (0.016) | (0.019) |
| Task 2 | 0.003 | 0.003 | -0.008 | 0.013 |
| | (0.009) | (0.009) | (0.012) | (0.013) |
| Confidence | 0.003*** | 0.003*** | 0.005*** | 0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| DiffBeliefAlgo Manager | -0.008*** | -0.008*** | 0.003* | -0.008*** |
| | (0.002) | (0.002) | (0.002) | (0.002) |
| Male* DiffBeliefAlgo Manager | 0.012*** | 0.011*** | | |
| | (0.003) | (0.003) | | |
| NoGenderW*Male | | 0.074 | | |
| | | (0.087) | | |
| TranspW*Male | | 0.004 | | |
| | | (0.088) | | |
| Observations | 743 | 743 | 366 | 377 |
| $R^2$ | | | | |
| Adjusted $R^2$ | | | | |
| errors clustered | | | | |
| controls | | | | |
| sample | All | All | Male only | Female only |

Notes: Marginal effects of probit regression of choosing the algorithm by workers. "Confidence" is the belief regarding how many workers out of 100 have lower task-1 and -2 performances than oneself. "DiffBeliefAlgoManager" equals the difference between the belief regarding how many men are hired by algorithm minus how many men are hired by managers. "Male*DiffBeliefAlgoManager" is the interaction of DiffBeliefAlgoManager and the dummy for the worker being male. Standard errors in parentheses, and * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A1. Determinants of choice of algorithm by workers

| | Algorithm-averse choice | Algorithm-averse choice | Algorithm-averse choice | Algorithm-averse choice | Algorithm-averse choice |
|---|---|---|---|---|---|

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| NoGenderW | -0.092** | -0.097** | -0.081** | -0.088*** | -0.091*** |
| | (0.038) | (0.038) | (0.033) | (0.033) | (0.033) |
| TranspW | -0.062 | -0.077* | -0.037 | -0.039 | -0.049 |
| | (0.040) | (0.039) | (0.036) | (0.036) | (0.035) |
| Age | | 0.003** | 0.002 | 0.002 | 0.002 |
| | | (0.001) | (0.001) | (0.001) | (0.001) |
| Male | | -0.100*** | -0.066** | -0.054* | -0.081*** |
| | | (0.031) | (0.027) | (0.028) | (0.029) |
| Task 1 | | | -0.116*** | -0.111*** | -0.111*** |
| | | | (0.009) | (0.009) | (0.009) |
| Task 2 | | | 0.069*** | 0.072*** | 0.069*** |
| | | | (0.007) | (0.007) | (0.007) |
| Confidence | | | | -0.002** | -0.001* |
| | | | | (0.001) | (0.001) |
| DiffBeliefAlgo Manager | | | | | 0.004*** |
| | | | | | (0.001) |
| Male* DiffBeliefAlgo Manager | | | | | -0.007*** |
| | | | | | (0.002) |
| Observations | 744 | 744 | 744 | 744 | 743 |
| Sample | All | All | All | All | All |

Notes: Marginal effects of probit regression of algorithm aversion. "Confidence" is the belief regarding how many workers out of 100 have lower task-1 and -2 performances than oneself. "DiffBeliefAlgoManager" equals the difference between the belief regarding how many men are hired by the algorithm minus how many men are hired by managers. "Male*DiffBeliefAlgoManager" is the interaction of DiffBeliefAlgoManager and the dummy for the worker being male. Standard errors in parentheses, and * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2. Determinants of algorithm aversion by workers

**Online Appendix A**
**Alternative specifications: OLS regressions**

| | 1st worker of the pair hired by manager (1) | 1st worker of the pair hired by manager (2) | 1st worker of the pair is the correct hire (3) | 1st worker of the pair is hired by algorithm (4) |
|---|---|---|---|---|
| 1st worker male minus 2nd worker male | -0.070*** | -0.094*** | -0.157*** | -0.443*** |
| | (0.007) | (0.010) | (0.042) | (0.011) |
| Task1 of 1st worker minus 2nd worker | 0.161*** | 0.161*** | 0.097*** | 0.153*** |
| | (0.004) | (0.004) | (0.016) | (0.015) |
| Task2 of 1st worker minus 2nd worker | 0.113*** | 0.113*** | 0.071*** | 0.190*** |
| | (0.005) | (0.005) | (0.022) | (0.017) |
| 1st worker male minus 2nd worker male * male manager | | 0.050*** | | |
| | | (0.014) | | |
| Observations | 15080 | 15080 | 15080 | 15080 |
| $R^2$ | 0.160 | 0.162 | 0.136 | 0.794 |
| Clustered errors | Manager | Manager | Pair | Pair |
| Sample | All | All | All | All |

Notes: OLS regression of dummy for hiring the first worker of a pair by the manager or by the algorithm. "1st worker male minus 2nd worker male" is the difference between the Male dummies corresponding to each worker of the pair. "Task1 of 1st worker minus 2nd worker" is the difference in performance between the two workers for task 1. "Task2 of 1st worker minus 2nd worker task 2" is the difference in performance between the two workers for task 2. "1st worker male minus 2nd worker male * male manager" is the interaction between "1st worker male minus 2nd worker male" and the gender of the manager. Standard errors in parentheses, and * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3. Determinants of hiring by managers and by the algorithm

| | Choice of algorithm (1) | Choice of algorithm (2) | Choice of algorithm (3) | Choice of algorithm (4) | Choice of algorithm (5) |
|---|---|---|---|---|---|
| NoGenderW | 0.125*** | 0.126*** | 0.121*** | 0.132*** | 0.142*** |
| | (0.044) | (0.044) | (0.044) | (0.044) | (0.044) |
| TranspW | 0.040 | 0.044 | 0.040 | 0.040 | 0.054 |
| | (0.045) | (0.045) | (0.045) | (0.045) | (0.045) |
| Age | | -0.001 | -0.000 | -0.000 | -0.000 |
| | | (0.001) | (0.001) | (0.001) | (0.001) |
| Male | | 0.057 | 0.053 | 0.030 | 0.076** |
| | | (0.037) | (0.037) | (0.037) | (0.038) |
| Task 1 | | | 0.010 | 0.002 | -0.000 |
| | | | (0.012) | (0.013) | (0.012) |
| Task 2 | | | 0.006 | 0.001 | 0.003 |
| | | | (0.009) | (0.009) | (0.009) |
| Confidence | | | | 0.003*** | 0.003*** |
| | | | | (0.001) | (0.001) |
| DiffBeliefAlgo Manager | | | | | -0.008*** |
| | | | | | (0.002) |
| Male* DiffBeliefAlgo Manager | | | | | 0.012*** |
| | | | | | (0.003) |
| Observations | 744 | 744 | 744 | 744 | 743 |
| $R^2$ | 0.011 | 0.015 | 0.017 | 0.028 | 0.059 |
| Sample | All | All | All | All | All |

Notes: OLS regression of choosing the algorithm by workers. "Confidence" is the belief regarding how many workers out of 100 have lower task-1 and -2 performances than oneself. "DiffBeliefAlgoManager" equals the difference between the belief regarding how many men are hired by the algorithm minus how many men are hired by managers. "Male*DiffBeliefAlgoManager" is the interaction of DiffBeliefAlgoManager and the dummy for the worker being male. Standard errors in parentheses, and * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4. Determinants of choice of the algorithm by workers

| | Delegation (1) | Delegation (2) | Delegation (3) | Delegation (4) |
|---|---|---|---|---|
| ConfidM | 0.157*** | 0.156*** | 0.131*** | 0.136*** |
| | (0.043) | (0.043) | (0.043) | (0.044) |
| TranspM | -0.019 | -0.020 | -0.045 | -0.044 |
| | (0.043) | (0.043) | (0.044) | (0.044) |
| Age | | -0.000 | -0.000 | -0.000 |
| | | (0.000) | (0.000) | (0.000) |
| Male | | 0.043 | 0.044 | 0.045 |
| | | (0.035) | (0.035) | (0.035) |
| Overconfidence | | | -0.017*** | -0.015*** |
| | | | (0.006) | (0.006) |
| ConfidM*Overconfidence | | | 0.034*** | 0.034*** |
| | | | (0.010) | (0.010) |
| Number of correct hires | | | | 0.007 |
| | | | | (0.009) |
| Observations | 754 | 754 | 752 | 752 |
| $R^2$ | 0.026 | 0.029 | 0.047 | 0.048 |
| Sample | All | All | All | All |

Notes: OLS regression of delegation to algorithm. "Overconfidence" is the difference between the belief regarding how many hires were correct and the actual number of correct hires. "ConfidM*Overconfid" is the interaction of Overconfid and dummy for treatment ConfidM. "Number of correct hires" is the number of pairs where the manager hired a worker with higher task-3 performance. Standard errors in parentheses, and * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5. Determinants of delegation to the algorithm by managers

## Appendix B (online)

## Instructions

Below, we document the instructions that the participants received on their screens. In square brackets, we indicate the purpose of the screens, but this information was not visible to the participants.

**[Start of the survey (common to all treatments)]**

**Screen 1. [Consent page]**

You are invited to take part in a research study. The study is administered by researchers at the University of Lausanne, University Paris-Dauphine, and Technical University of Berlin. You will receive 1£ for participation, and will be able to earn up to 3.05£ in addition depending on your answers. Total duration of the study is 8 to 10 minutes. Please note that participation in this study is entirely voluntary and that you may discontinue participation at any time. In this case, you will not be

compensated. All data will be treated confidentially. Data will be used in anonymized way for academic research only. Anonymized data will be made available to other researchers for replication purposes

- I understand the conditions and consent to participate in this study
- I reject participation

**Screen 2. [Data page]**

Gender What is your gender?

- Male
- Female

Age How old are you?

ProlificID

**[BaselineW]**

**Screen 3. Task 1 out of 3**
In the next screen you will have 12 questions and 90 seconds to answer the questions. If this task will be randomly selected for the payment, you will earn 0.15£ for each correct answer.

**Screen 4. [**Task 1]
12 Raven matrices

**Screen 5. Task 2 out of 3**
In the next screen you will have 12 questions and 90 seconds to answer the questions. If this task will be randomly selected for the payment, you will earn 0.15£ for each correct answer.

**Screen 6. [**Task 2]
12 counting zeros

**Screen 7. Task 3 out of 3**
In the next screen you will have 12 questions and 90 seconds to answer the questions. If this task will be randomly selected for the payment, you will earn 0.15£ for each correct answer.

**Screen 8. [**Task3]
5 Raven matrices and 7 counting zeros

**Screen 9. AI introduction screen**

In the next block, you will have to make several decisions that can bring you an additional bonus. It is important to provide you with some context.

**Artificial intelligence (AI)** in hiring **involves the use of technology to automate aspects of the hiring process.** Advances in artificial intelligence, such as the advent of machine learning and the growth of big data, enable AI to be utilized to recruit, screen, and predict the success of applicants.

How is AI used for hiring? AI-powered preselection software **uses predictive analytics to calculate a candidate's likelihood to succeed in a role.** This allows recruiters and hiring managers to make data-driven hiring decisions rather than decisions based on their gut feeling.

**Screen 10.** [Choice screen]
In this task, you might earn additionally £0.5 if you are hired in subsequent experiments.

The hiring might be done either by participants like you who will play the role of **managers** or by **artificial intelligence (AI)**. The manager or AI will choose which of two workers to hire based on three pieces of information:
1) both workers' **genders**
2) number of correct answers in **task 1**
3) number of correct answers in **task 2.**

**AI is trained t**o give the best prediction of the performance in task 3, **based on the gender** and **tasks 1 and 2 performance** from 200 workers. No other objectives or information is available to AI. **There was no human supervision to "correct" or change the algorithm due to any objectives. Artificial intelligence hires the worker from the pair for whom it predicts the highest task 3 performance.**

The **managers** know all the questions in tasks 1, 2, and 3 and all correct answers to the questions. Before hiring decisions they **go through training** where they see performances of 20 workers in all 3 tasks, together with the gender of the workers. They also know the proportion of questions in task 3 which are similar to tasks 1 and 2 respectively. For one random pair of workers for whom they have made a hiring decision, **a manager will get 2£ if they decided to hire the worker with the highest performance in task 3.**

Do you want to be hired by a manager or by the AI? Your decision will be implemented, and if you are hired in one random pair, you will additionally receive £0.5.

- I want that my hiring decision is taken by **manager** (1)
- I want that my hiring decision is taken by **artificial intelligence** (2)

**Screen 11. [**Confidence]

Think about your performance in tasks 1 and 2. Out of random 100 participants, how many do you think have a total number of correct answers at tasks 1 and 2 lower than you? If your answer is within 5 from correct answer, you will additionally earn 0.25 pounds.

| | 0 – you have the worst score 0 10 20 30 40 50 60 70 80 90 100 100 – you have the best score |
|---|---|
| Mover slider to determine how many participants have lower score than yours? () |  |

**Screen 12.**

BeliefManager

Imagine **managers' decisions** about whom to hire based on gender and performance in tasks 1 and 2. There is an equal number of men and women candidates. **Out of 100 hired workers, how many will be men?**

You will earn 0.25 pounds if your guess will be within 5 from correct answer.

| | Only women are hired 0 10 20 30 40 50 60 70 80 90 100 Only men are hired |
|---|---|
| Out of 100 hired workers, there will be men () |  |

Belief algorithm

Imagine decisions **by artificial intelligence** about whom to hire based on gender and performance in tasks 1 and 2. There is an equal number of men and women candidates. **Out of 100 hired workers, how many will be men?**

You will earn 0.25 pounds if your guess will be within 5 from correct answer.

| | Only women are hired 0 10 20 30 40 50 60 70 80 90 100 Only men are hired |
|---|---|
| Out of 100 hired workers, there will be men () |  |

**Screen 13.** Belief gender performance

Imagine performance of participants in task 3. There is an equal number of men and women workers. Out of 100 workers, **how many will be men among 50 best performers?**

You will earn 0.25 pounds if your guess will be within 5 from correct answer.

| | All best performers areall best performers are women men<br>0 5 10 15 20 25 30 35 40 45 50 |
|---|---|
| Out of 100 hired workers, there will be men () |  |

**[NoGenderW]**

[All as in **BaselineW** except for the **Choice screen]**

In this task, you might earn additionally £0.5 if you are hired in subsequent experiments.

The hiring might be done either by participants like you who will play the role of **managers** or by **artificial intelligence (AI)**. The manager will choose which of two workers to hire based on three pieces of information:

1) both workers' **genders**
2) number of correct answers in **task 1**
3) number of correct answers in **task 2.**

**The AI has no access to the gender of candidates, only to their performance in tasks 1 and 2.**

**AI is trained t**o give the best prediction of the performance in task 3, **based on tasks 1 and 2 performance** from at least 200 workers. No other objectives or information is available to AI. **There was no human supervision to "correct" or change the algorithm due to any objectives. Artificial intelligence hires the worker from the pair for whom it predicts the highest task 3 performance.**

The **managers** know all the questions in tasks 1, 2, and 3 and all correct answers to the questions. Before hiring decisions they **go through training** where they see performances of 20 workers in all 3 tasks, together with the gender of the workers. They also know the proportion of questions in task 3 which are similar to tasks 1 and 2 respectively. For one random pair of workers for whom they have made a hiring decision, **a manager will get 2£ if they decided to hire the worker with the highest performance in task 3.**

Do you want to be hired by a manager or by the AI? Your decision will be implemented, and if you are hired in one random pair, you will additionally receive £0.5.

- I want that my hiring decision is taken by **manager** (1)
- I want that my hiring decision is taken by **artificial intelligence** (2)

**[TranspW]**

[All as in **BaselineW** except **Choice screen]**

In this task, you might earn additionally £0.5 if you are hired in subsequent experiments.

The hiring might be done either by participants like you who will play the role of **managers** or by **artificial intelligence (AI)**. The manager or AI will choose which of two workers to hire based on three pieces of information:
1) both workers' **genders**
2) number of correct answers in **task 1**
3) number of correct answers in **task 2.**

**The AI is trained t**o give the best prediction of the performance in task 3, **based on the gender** and **tasks 1 and 2 performance** from at least 200 workers. No other objectives or information is available to the AI. **There was no human supervision to "correct" or change the algorithm due to any objectives. The Artificial intelligence hires the worker from the pair for whom it predicts the highest task 3 performance.**
**The algorithm calculates for the at least 200 workers it has data on the mean relationship between the task 1 and 2 performances and gender on the one hand and the task 3 performance on the other hand.** This relationship is:
**Task3=0.33*Task1+0.39*Task2-0.35*Male+2.6**
so that, in order to predict someone's task 3 performance, one must replace respectively Task1 and Task2 by the tasks 1 and 2 performances of the person and **deduct 0.35 only if the participant is male.**

The **managers** know all the questions in tasks 1, 2, and 3 and all correct answers to the questions. Before hiring decisions they **go through training** where they see the performances of 20 workers in all 3 tasks, together with the gender of the workers. They also know the proportion of questions in task 3 which are similar to tasks 1 and 2 respectively. For one random pair of workers for whom they have made a hiring decision, **a manager will get 2£ if they decided to hire the worker with the highest performance in task 3.**

Do you want to be hired by a manager or by the AI? Your decision will be implemented, and if you are hired in one random pair, you will additionally receive £0.5.

**[BaselineM]**

**Screen 3. [Intro]**
In this survey, you will make hiring decisions. Participants in the role of workers had to perform three tasks. In the next block, you will see precisely the tasks workers performed.

You will be asked whom to hire among 20 pairs of workers. You will know the workers' performance in tasks 1 and 2 and the gender of each worker.

Your goal will be to hire the worker of each pair with the best performance in task 3. Note that your decisions will also matter for workers, as hired workers might earn additional payoff.

Note that the task 3 consists of 7 questions of the type of task 1 and 5 questions of the type of task 2.

**Screen 4. [Task 1]**
Next 12 questions represent task 1. Workers had 1.5 minutes to answer as many questions as possible. You have up to 30 seconds to get familiar with the questions of this task.

**Screen 5. [Task 2]**
Next 12 questions represent task 2. Workers had 1.5 minutes to answer as many questions as possible. You have up to 30 seconds to get familiar with the questions of this task.

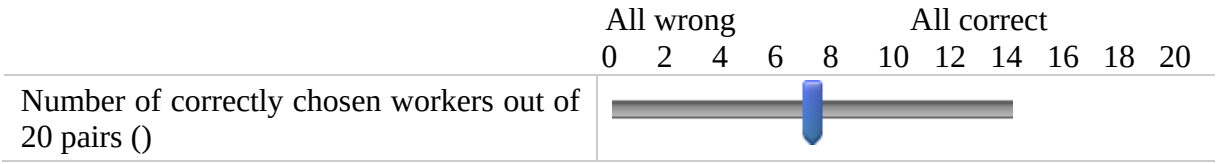
**Screen 6. [Task 3]**
Next 12 questions represent task 3. Workers had 1.5 minutes to answer as many questions as possible. You have up to 30 seconds to get familiar with the questions of this task.

**Screen 7. [Training example]**
Before you start the hiring decisions, we present you with 20 random workers and their gender and tasks 1, 2, and 3 performance, so you can learn which workers you want to hire and which characteristics matter for task 3 performance. You will have up to 2 minutes to observe the information before moving to the hiring task.

**Screen 8. [Choice]**
In the next task you will have to make 20 hiring decisions. In each decision there will be two candidates. You will know the number of correct answers of each candidate in tasks 1 and 2, and also the gender of each candidate. Your task is to hire the worker of each pair with the best performance in task 3. We will select one random decision of you, and if you hire the candidate with indeed higher number of correct answers in Task 3 you will receive £2.

**Screens 9-29 Hiring between pairs**

**Screen 30 [Confidence]**
Think about the hiring decisions you just made. Out of these 20 decisions, how many do you think are correct, i.e the chosen worker indeed had a better Task 3 performance? If your answer is within 1 from correct answer, you will additionally earn 0.25 pounds.




**Screen 31. [Delegation]**
We have developed an algorithm that is trained to predict the performance of workers in task 3 based on their performance in task 1, task 2, and their gender. The algorithm is trained on 200

workers. The algorithm always hires the worker from the pair for whom it predicts the highest task 3 performance.
Now you have a chance to delegate your decisions to the algorithm. If you decide so, then instead of your hiring decisions, we will use the algorithm choices, and these will be the ones relevant for your payoff in the hiring decision task. What do you choose?

- Keep my hiring decisions
- Override my hiring decisions with those of the algorithm

**[TranspM]**

[All as in BaselineM except delegation screen]

We have developed an algorithm that is trained to predict the performance of workers in task 3 based on their performance in task 1, task 2, and their gender. The algorithm is trained on 200 workers. The algorithm always hires the worker from the pair for whom it predicts the highest task 3 performance.

**The algorithm calculates for the at least 200 workers it has data on the mean relationship between the task 1 and 2 performances and gender on the one hand and the task 3 performance on the other hand.** This relationship is:
**Task3=0.33*Task1+0.39*Task2-0.35*Male+2.6**
so that, in order to predict someone's task 3 performance, one must replace respectively Task1 and Task2 by the tasks 1 and 2 performances of the person and **deduct 0.35 only if the participant is male.**

Now you have a chance to delegate your decisions to the algorithm. If you decide so, then instead of your hiring decisions, we will use the algorithm choices, and these will be the ones relevant for your payoff in the hiring decision task. What do you choose?

- Keep my hiring decisions
- Override my hiring decisions with those of the algorithm

**[ConfM]**

[All as in BaselineM but an extra screen between Confidence screen right and delegation screen]

You think that you have correctly hired XXX workers out of 20 pairs.

In fact, you hired correctly YYY workers.

Thus, you are **overconfident/underconfident/close to correct answer.**