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

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

September 2022

Abstract

We run an online experiment to study the origins of algorithm aversion. Participants are either in the role of workers or of 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 either 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. In the baseline treatments, we observe that workers choose the manager more often than the algorithm, and managers also prefer to make the hiring decisions themselves rather than delegate them 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. Providing details on how the algorithm works does not increase the preference for the algorithm, neither for workers nor for managers. Providing feedback to managers about their performance in hiring the best workers increases their preference for the algorithm, as managers are, on average, overconfident.

Acknowledgements:
We would like to thank 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.

Introduction

Algorithms are used by companies for a variety of decisions including forecasts of employees' performance and hiring (Highhouse 2008, Carey and Smith 2016, Chalfin et al. 2016). It appears that firms that use such algorithms perform better than others (Bajari et al. 2019, Camuffo et al. 2020). 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). However, recent evidence suggests that 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 both perspectives – of 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. As such, they will tend to choose the hiring process they believe is more likely to result in the hiring of the best workers, since this generates the highest payoff for them.

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 for 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.

Our experiments are motivated by the recent debates in the EU over the legal requirements for algorithmic decisions. 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. 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). While the GDPR led to some expected effects, e.g., privacy-oriented

consumers opting out of the use of cookies (Aridor et al. 2020), the discussion over the transparency requirements and the constraints on profiling is still ongoing. Recently, the European Parliament came up with the Digital Services Act (DSA), which proposes further increasing the requirements for algorithm disclosure and which explicitly requires providing a profiling-free option to users, together with a complete ban on the profiling of minors.[1] Our first treatment that focuses on the workers aims at identifying whether making the algorithm gender-blind and therefore unable to use gender to discriminate, as advised in the preamble of the GDPR and further strengthened in the proposed DSA, increases its acceptance by the workers. The second treatment is a direct test of the importance of the transparency of the algorithm for the workers. When the algorithm is made transparent in our setup, it becomes evident which gender is favored. This can impact algorithm aversion differently for women and men, for example if workers' preferences are mainly driven by payoff maximization.

The treatments focusing on the managers' preferences aim at understanding why some firms are more reluctant than others to make use of hiring algorithms. One possible explanation for not adopting such algorithms is managerial overconfidence. Overconfidence is a common bias, and its effect on several economic behaviors 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. Managers who believe they make better hiring decisions than they actually do, may prefer to make the hiring decisions themselves. Our paper will provide insights about the effect of overconfidence on the delegation of hiring decisions to algorithms. Similar to the treatments about the preferences of workers, we are also interested in the effect of the transparency of the algorithm on the managers' willingness to delegate the hiring decisions. Disclosing the details of the algorithm can increase the managers' trust in the algorithm.

We run an online experiment in which participants are either in the role of workers or of 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 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 an algorithm. Managers must decide whether they want to delegate the hiring decisions to the

[1] Press release of European Parliament from 23.3.2022. https://www.europarl.europa.eu/news/en/press-room/20220412IPR27111/digital-services-act-agreement-for-a-transparent-and-safe-online-environment

algorithm or decide themselves. If they delegate the decisions, their payoff will depend on the algorithm's hiring decisions instead of their own.

For given task-1 and 2 performances, we observe that female workers perform better in the job task. This could be due to women learning more from previous experience than men, for example. As a result, the algorithm favors female workers over male workers when they have similar task-1 and 2 performances. The managers in our experiment also favor female workers in such cases. Managers observe a random set of 20 workers, which is enough for them to spot the tendency of female workers to outperform male workers given similar task-1 and 2 performances. However, managers favor female workers less than what would have been optimal, while the algorithm favors female workers more than optimally.[2] Overall, the proportion of correct hires by the algorithm is significantly higher than that by the managers (67% vs. 56%). Thus, we refer to the choice of the manager to make the hiring decision as algorithm aversion, both for choices by subjects in the role of managers and of workers. Note, however, that avoiding the algorithm can be rational for some individuals. We analyze the optimality of individual decisions in the results section.

In the baseline treatment, 47% of workers chose the hiring algorithm. We therefore refer to 53% of workers as algorithm-averse. The preference for the algorithm correlated positively with workers' confidence in their performance. Also, we observe that, at least to some extent, choices were driven by self-interest: male (female) workers who believed that managers favor male (female) workers more than the algorithm were more likely to choose managers. When the algorithm does not use the gender of 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. This result supports the new DSA that requires firms to provide profiling-free options for users. On the other hand, we find that providing details about how the algorithm works does not increase the preference for the algorithm by the workers. The result suggests that if there is a preference for transparency by the workers, it is not very strong, at least in our setting.

As for managers, in the baseline treatment only 34% of managers delegated the hiring decisions to the algorithm, which is a clear sign of algorithm aversion, given that the algorithm was, on

[2] The suboptimal performance of the algorithm when making out-of-sample predictions is the limited sample size of the training sample.

average, more efficient in hiring the best workers than 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%). This is especially the case for managers who were overconfident and therefore received feedback informing them that they had made significantly fewer correct hiring decisions than they thought. Notably, the increase in the delegation was driven by those for whom it was optimal to delegate. Thus, we establish a causal effect of self-confidence on delegation to algorithms and show that overconfidence might be a substantial barrier to adopting algorithms. As for transparency, disclosure of the algorithm did not significantly affect delegation by the managers.

Our paper belongs to a literature that is interested in the preference for human decision-making relative to algorithmic decision-making. Several papers document people's reluctance to use algorithms even when they are more efficient than human decision-making. Fildes and Goodwin (2007) found that many professional forecasters do not use algorithms (or do not use them enough) in their forecasting process. Sanders and Manrodt (2003) observe that also firms do not rely on algorithms for forecasting even though firms that did rely on 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) found that 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 found out that the algorithm was imperfect, they became reluctant to use it. Most participants (over 60%) of Chugunova et al. (2022) prefer the algorithm over a human decision maker in the context of redistributive decisions although the decisions made by humans are regarded more favorably than those made by the algorithm. In a recent survey, Will et al. (2022) document that despite the evidence that AI is better than humans in hiring, the candidates and recruiters perceive it to be worse.

Our paper also contributes to the literature investigating ways to mitigate algorithm aversion. Dietvorst et al. (2018) found that being able to slightly modify the forecasts of the algorithm made the participants much 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 that reducing the aversion to moral decision-making by algorithms or machines is not easy, and depends upon making salient the expertise of machines and the ability of humans to override them.

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) where the managers observe a performance measure of the workers that is correlated with their job performance. While 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. We did not expect this, given the previous findings of discrimination in favor of men, and it is likely driven by the fact that we provide more information to the participants in the role of managers than in previous studies. This allows them to learn about the higher job performance of women compared to men with similar past performances. The most important difference to the literature on hiring discrimination is our focus on the preference for algorithms.

Focusing on algorithmic hiring, 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 that 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 an algorithm hires them rather than humans. On the other hand, Bigman et al. (2022) show that people are less outraged by a discriminating algorithm than by a discriminating human.

The ways in which algorithms discriminate between groups of the population, often indirectly by basing the choice on characteristics that are correlated with, e.g., 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 that 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 this is more cost-effective.

Our study relates to the large literature on optimal delegation. 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 trained with more information than the information the manager has access to, and it is programmed to choose the best worker, such that incentives are aligned. However, the manager may not understand or trust the algorithm. More recently, behavioral biases have been accounted for in the literature on delegation, both in theory (see, e.g., Auster and Pavoni 2021) and in experiments (e.g., Danz et al. 2015, Ertac et al. 2020). Algorithm aversion can be understood as a bias that hinders efficient delegation.

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). The latter work also relates to our study in that it deals with building trust in digital services. Aoki (2020) investigates the determinants of trust in AI chatbots which answer questions from the population on behalf of the government. She suggests that 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.

Experimental design

We ran an online experiment on the British platform Prolific with participants from the US. Six treatments in total were conducted between subjects. We aimed at having roughly 250 participants (125 men and 125 women) in each treatment, adding up to a total of 750 workers and 750 managers. We ended up collecting data from 744 workers and 754 managers.[3] 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 and age, and their consent to participate in the study. Participants were either in the role of workers or in the role of managers.

[3] We recruited 260 participants per treatment, as we expected that some participants would have to be excluded because they clicked through the survey in less than one minute, or because they did not make a decision in one of the main tasks..

We run 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), then 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 in a 6x6 matrix. Workers are paid £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 explain 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 the hiring decision to be made by the algorithm or a participant in the role of a manager. Workers will get an additional payment of 50 pence if they were hired in the pair of workers that we randomly selected for payment. Note that 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 percent of workers have a lower performance than themselves, 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. Given the equal number of men and women in a candidate subsample of 100 workers, we ask how many of the 50 best-performing workers they believe 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 to 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 that managers still learn about the workers' gender.

Transparency treatment for workers (TranspW)

Compared to the BaselineW treatment, the only difference is that participants are given details about how the algorithm works before deciding whether they would prefer the hiring decisions to be made by the manager or the algorithm. More precisely, we disclose the regression equation that the algorithm employs to predict performance in the job task. The exact 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 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)

The BaselineM treatment was conducted after the BaselineW treatment. The managers observe all questions in the three tasks that workers had to solve, but 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 treatment.

We ask the managers to make 20 hiring decisions 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, while the other 10 pairs for each manager were selected only among those pairs of workers whose performance difference did not exceed one and where the two workers were of a different gender. Again, this was to ensure that we could identify managers who favored workers of a particular gender, given similar performance.

For one randomly chosen hiring decision, the manager earns £2 if the decision is correct, meaning that the worker who is hired has a better performance in the job task than the other worker of the pair. After the hiring decisions are made, we elicit the participants' belief in how often they had chosen the better worker in the 20 pairs, i.e., 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, the managers are asked whether they want to delegate the hiring decisions to an algorithm. They are told that 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. They are informed that the algorithm is trained to predict who is the better performer in the job task, based on the data from at least 200 workers. If the managers decide to delegate the hiring decisions to the algorithm, their 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 receive feedback on the number of correct hires out of their 20 hiring decisions after the belief elicitation stage and before they decide whether to delegate to the algorithm or not. Additionally, we inform them of whether they are overconfident (guessed at least two more correct hires than their actual performance), underconfident (guessed at least two less 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 to the BaselineM treatment is that managers are provided with information about how the algorithm works before deciding whether to delegate the hiring decisions to the algorithm. The information they receive about the algorithm is the same as in the TranspW treatment.

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.[4]

Treatment differences

Workers:

H1 (role of algorithm using gender for algorithm aversion): A higher share of workers prefers to be hired by the algorithm rather than by managers in NoGenderW than in BaselineW.

Support: Recent debates on potentially discriminative algorithms due to gender and racial profiling suggest that people might have a preference against discrimination based on gender,

[4] 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, it turned out that the managers hire more women than men (even though to a lesser extent than the algorithm) and that 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, H5 and H6 in the manuscript), as they rely on the discrimination against women. We did, however, choose to keep pre-registered hypothesis 4 (now labeled H5 in the manuscript) which predicts that the managers' gender influences which workers they favor, since we can test it with our data.

independent of whether it is advantageous for them or not. An alternative hypothesis would be that the preference for the algorithm or manager depends on the gender of the workers and on their beliefs about the extent of the algorithm and the managers' favoritism toward male workers, pointing to self-serving preferences over algorithms.

H2 (role of transparency for algorithm aversion of workers): A higher share of workers prefers to be hired by the algorithm (rather than by managers) once the algorithm has been explained, i.e., in TranspW compared to BaselineW.

Support: Our hypothesis is motivated by the criticism that algorithms are not transparent. Non-transparent algorithms may be suspected of being faulty or discriminatory which could contribute to algorithm aversion. We therefore hypothesize that, for workers, transparent algorithms are more attractive relative to managers regarding hiring decisions, compared to non-transparent algorithms.

Managers:

H3 (role of managers' self-confidence on delegation to algorithms): A higher share of managers delegates the hiring decisions to the algorithm in ConfidM than in BaselineM. The effect is driven by the majority of managers being overconfident of their number of correct hires.

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

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

Support: Similar to workers, our hypothesis is motivated by the common criticism that algorithms are not transparent.

Additional hypothesis

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

H5 (stereotypes 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 recent findings suggesting the favoring of male candidates in experimental hiring experiments. Sarsons et al. (2021) find in a hiring experiment 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 an experimental hiring 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 can then turn to studying the effect of our treatments on 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, since the treatments differ only at a later stage, namely right before the decision is taken of whether the manager's hiring decision is implemented or the algorithmic decision is followed.

In task 1 (Raven matrices), men have a higher performance than women (3.67 vs 4.03; a two-sided Mann-Whitney test yields p=0.01), and there is no significant gender difference in performance for task 2 (6.12 vs 6.18; p=0.43) and the job task (6.19 vs 6.13; p=0.96). Table 1 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 have a worse performance in the job task. Models (3) and (4) present the results for the BaselineW and NoGenderW treatments only, as the data from these treatments were used to generate the hiring algorithm in the BaselineW, TranspW and NoGenderW treatments.[5] As can be taken from the significant and negative coefficient of Male, the algorithm picks the woman when a man and a woman have the same performance in tasks 1 and 2. Moreover, 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 as the absolute value of the Male coefficient is larger than that of the Task 1 coefficient.

	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 * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1. 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:

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

[5] The data from the TranspW treatment were not used to develop the algorithm, since we needed to have the algorithm ready before this treatment to be able to disclose it to participants.

In 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.[6] Managers hire 56.1% of women (54.1% for male managers and 57.7% for female managers; a two-sided Mann-Whitney test yields $p<0.01$). While 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, and this is the case for only 55.9% of the managers' hiring decisions ($p<0.01$).

The mean task-1 performances of men and women hired by the managers are, respectively, 4.02 and 3.87 ($p<0.01$). The mean task-2 performances of men and women hired by the managers are, respectively, 6.17 and 6.00 ($p<0.01$). 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 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 2 presents the results of probit regressions.[7] In models (1) and (2), we regress a dummy equal to one 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,[8] 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 of whether the algorithm hires the first worker of the pair. Thus, model (3) presents the

[6] The high percentage of women being hired is due to the fact that we match workers with similar performances, which even leads to the hiring of too many women compared to the optimum. If we randomly matched workers into pairs, the algorithm would hire 53.1% of women (based on 1,000 simulations for each worker in the sample).
[7] For this and all other probit regressions in the paper, we also document the OLS regressions in Appendix A (online). All results are qualitatively the same.
[8] This dummy is equal to 1 if the first worker is male and the second worker 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 coefficient of this variable can be interpreted as female workers being favored in the hiring decisions.

coefficients for optimal hiring, and we use it as a benchmark to judge the optimality of the managers and algorithm's hiring decisions.

Managers favor female candidates, as can be seen from the negative coefficient of "1st worker male minus 2nd worker male." It confirms that observing the task-1, task-2, and job-task performances as well as the gender of a subset of workers allows the managers to learn that female workers have a higher job-task performance than male workers for given task-1 and 2 performances.

	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*** (0.006)	-0.086*** (0.009)	-0.147*** (0.036)	-0.111*** (0.01)
Task1 of 1st worker minus 2nd worker	0.211*** (0.009)	0.211*** (0.009)	0.102*** (0.02)	0.078*** (0.007)
Task2 of 1st worker minus 2nd worker	0.163*** (0.008)	0.162*** (0.008)	0.076*** (0.025)	0.098*** (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 2. Determinants of hiring by managers and by the algorithm.

Hypothesis 5, which states that managers favor male workers, is not validated. However, the comparison of models (1) and (3) with respect to the coefficient of the first regressor shows that managers do not favor women as much as would have been optimal. Furthermore, male managers favor female workers to a lesser extent than female managers. This can be taken from the positive coefficient 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 coefficient in a regression with clustered errors at the subject level).

Lastly, model (4) presents the results of the hiring decisions made by the algorithm. The decisions of the algorithm are closer to the optimum with respect to favoring women than the decisions by the managers, since the coefficient for "1st worker male-2nd worker male" is closer to that of model (3). Note that the coefficient is still smaller than the one corresponding to an optimal decision. Interestingly, the algorithm gives a weight that is less than optimal to the performance in tasks 1 and 2, while managers overweigh the task-1 and task-2 performances.

We sum up the main findings of this section:

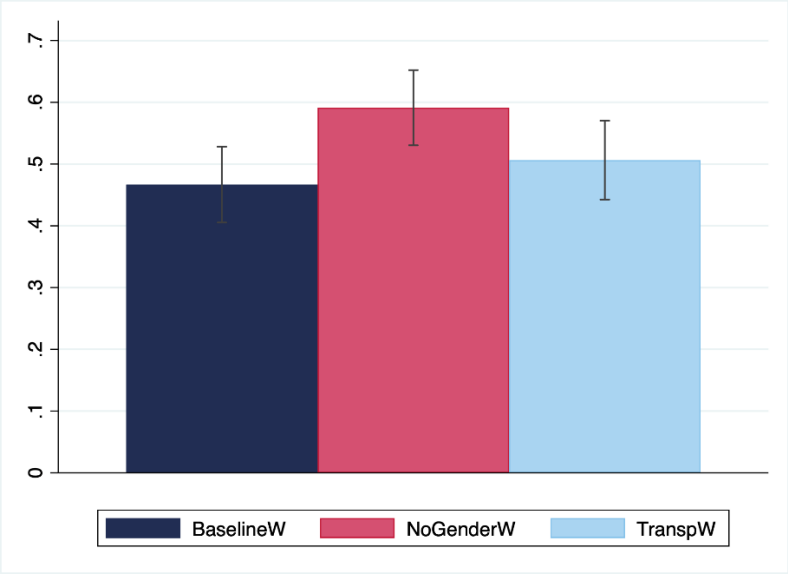
Result 1 (Workers' performance and hiring decisions): *Conditional on task-1 and task-2 performances, female workers outperform male workers at the job task. Managers expect this and favor female candidates, conditional on the workers' performance in tasks 1 and 2. Male managers tend to give a significantly 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. Figure 1 presents the proportion of workers that 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 compared to 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. As for TranspW, the proportion of workers choosing the algorithm is not significantly different from BaselineW (50,63% vs 46,67%, a two-sided Mann-Whitney test yields p=0.38) and is marginally lower than in NoGenderW (p=0.06). Thus, we find support for H1 and no support for H2.



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 is favored by each hiring process. Table 3 presents the marginal effects of probit regressions of the choice of the algorithm – the dummy being equal to one if the worker prefers 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" that measures the proportion of workers one believes to have a lower sum of task-1 and 2 performances than oneself. There is a significant correlation between the workers' confidence in their performance and their choice of the algorithm.[9] The better the workers think they performed, the more likely they are to choose the algorithm.

[9] Note that self-confidence is significantly correlated with actual performance (rho=0.37). Men are more confident than women about their task-1 and 2 performances. On average, men and women believe that, respectively, 53.4% and 44.8% (p<0.01) of workers have lower task-1 and 2 performances than themselves.

	Choice of algorithm (1)	Choice of algorithm (2)	Choice of algorithm (3)	Choice of algorithm (4)	Choice of algorithm (5)	Choice of algorithm (6)
NoGenderW	0.125***	0.126***	0.121***	0.132***	0.141***	0.139***
	(0.044)	(0.044)	(0.044)	(0.044)	(0.043)	(0.043)
TranspW	0.040	0.044	0.040	0.040	0.054	0.050
	(0.045)	(0.045)	(0.045)	(0.045)	(0.044)	(0.044)
Age		-0.001	-0.000	-0.000	-0.000	-0.000
		(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Male		0.057	0.053	0.030	0.076**	-0.154
		(0.036)	(0.036)	(0.037)	(0.038)	(0.141)
Task 1			0.010	0.002	-0.000	0.001
			(0.012)	(0.013)	(0.012)	(0.012)
Task 2			0.006	0.001	0.003	0.002
			(0.009)	(0.009)	(0.009)	(0.009)
Confidence				0.003***	0.003***	0.003***
				(0.001)	(0.001)	(0.001)
DiffBeliefAlgo Manager					-0.008***	-0.008***
					(0.002)	(0.002)
Male* DiffBeliefAlgo Manager					0.012***	0.011***
					(0.003)	(0.003)
BeliefMenTop50						-0.006*
						(0.003)
Male* BeliefMenTop50						0.008*
						(0.005)
Observations	744	744	744	744	743	743
Sample	All	All	All	All	All	All

Notes: Marginal effects of probit regression of choosing the algorithm by workers. "Confidence" is the belief of how many workers out of 100 have lower task-1 and 2 performance than oneself. "DiffBeliefAlgoManager" equals the difference between belief of 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 manager being male. "BeliefMenTop50" is the belief participants have of how many of the top 50 workers in terms of performance out of 100 are men. "Male* BeliefMenTop50" is the interaction of BeliefMenTop50and the dummy for the manager being male. Standard errors in parentheses, and * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 3. Determinants of the workers' choice of the algorithm

We also collected the beliefs of workers concerning the proportion of men that would be hired by the managers and by the algorithm, respectively, from a gender-balanced pool of workers. On average, workers believe that managers will hire 55.7% of men (men answer 54.8% on average, women 56.7%; a two-sided Mann-Whitney test yields $p<0.01$) and that the algorithm will hire 51.8% of men (men answer 51.1% on average, women 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 compared to the managers with the variable

"DiffBeliefAlgoManager" and with an interaction of this variable with the Male dummy. Female workers are more likely to choose the algorithm the more they believe that the algorithm favors female workers more strongly than the managers. 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. Thus, we observe a weaker preference for the algorithm of those workers who believe that the algorithm discriminates against their gender. Note that when controlling for the beliefs about gender-based hiring by the algorithm compared to the managers, the male dummy becomes significant, pointing to a higher tendency of male workers to choose the algorithm when believing that the managers and the algorithm favor male workers equally.[10]

Finally, model (6) controls for participants' beliefs about which gender is more productive. When female participants believe that more men are in the top 50 most-productive workers, they are less likely to choose the algorithm (see the coefficient for BeliefMenTop50), and the opposite is true for male participants. Both coefficients are only marginally significant at the 10% level, but the evidence suggests that both genders perceive the algorithm to be more meritocratic than the managers' hiring decisions.

How close are the workers' choices to the empirical optimum from an individual perspective, i.e., do they take payoff-maximizing decisions at the individual level? While the algorithm hires more efficiently, the workers might prefer managers to make the hiring decisions if they think that the managers are more likely to hire them than the algorithm. To analyze the optimality of the workers' decisions, we simulate hiring by managers based on model (1) of Table 2 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 it is the optimal decision for the worker to choose the algorithm.

Overall, only 50.5% of workers (54% of women and 47% of men) make the optimal, i.e., individual-payoff-maximizing, choice between the algorithm and the managers. The gender

[10] Including triple interactions of DiffBeliefAlgoManager, a male dummy, and each of the treatment dummies yields non-significant coefficients. Thus, we do not find stronger treatment effects for workers who believe that the algorithm discriminates against their gender. We also introduced interactions of each of the treatment dummies and a male dummy, but the coefficients are not significant either. Thus, it is not the case that the treatments effects are different for male and female workers. It is surprising that the transparency treatment, which makes it obvious that women are favored over men, does not have a differential effect on men and women.

difference is driven by BaselineW, where the proportion of optimal choices is significantly smaller for male than for female workers (a two-sided Mann-Whitney test yields p=0.02). The difference is not surprising, as 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 (a two-sided Mann-Whitney test yields p=0.90 and p=0.42, respectively). In NoGenderW, the female workers cannot be favored, thus removing the advantage of the algorithm for female workers. In TranspW, while the algorithm still favors female workers, male workers are aware of it and make better choices between the managers and the algorithm.

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

Result 2 (Workers' algorithm aversion): *Workers are significantly less algorithm-averse toward a gender-blind algorithm than an algorithm using gender. Disclosure of the details of the algorithm does not decrease algorithm aversion. There are no significant treatment differences regarding the optimality of the workers' choice of the algorithm.*

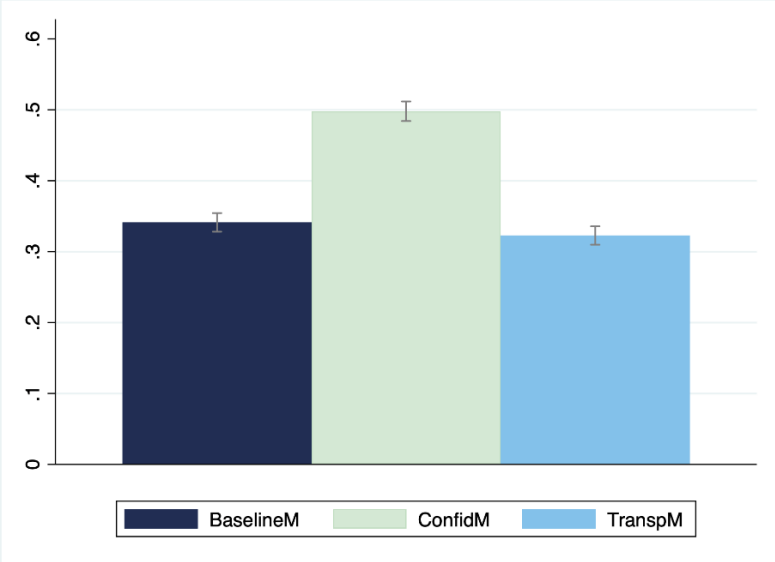
Managers: Delegation to the algorithm

Our main interest is again in the treatment differences, now regarding the managers' decisions of whether to delegate the hiring decision to the algorithm or not. We first present relevant descriptive statistics of the managers' hiring decisions. On average, managers made 11.2 correct hiring decisions out of 20 (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, 11.3 for women (a two-sided Mann-Whitney test yields p=0.02). Thus, we observe a significant overconfidence of men (p<0.01), but not of women (p=0.72), where overconfidence is defined as the difference between the believed number of correct hiring decisions minus the actual number of correct hiring decisions.

We next investigate the managers' choice to delegate to the algorithm. Figure 2 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 shows the causal effect of correcting the managers'

overconfident beliefs regarding their delegation decision. As for TranspM, the proportion of managers delegating to the algorithm is 32.2%, which is not significantly different from BaselineM (a two-sided Mann-Whitney test yields p=0.66) and significantly lower than in ConfidM (p<0.01). Thus, we find support for H3 and no support for H4.



Notes: Black lines correspond to 95% confidence intervals.

Figure 2: Proportion of managers who delegate hiring decision to the algorithm by treatment.

	Delegation (1)	Delegation (2)	Delegation (3)
ConfidM	0.157***	0.155***	0.130***
	(0.043)	(0.043)	(0.044)
TranspM	-0.019	-0.020	-0.045
	(0.042)	(0.042)	(0.043)
Age		-0.000	-0.000
		(0.000)	(0.000)
Male		0.043	0.044
		(0.035)	(0.035)
Overconfidence			-0.018***
			(0.006)
ConfidM*Overconfidence			0.034***
			(0.009)
Observations	754	754	752
Sample	All	All	All

Notes: Marginal effects of probit regression of delegation to algorithm. "Overconfidence" is the difference between the belief of how many hires were correct and the actual number of correct hires. "ConfidM*Overconfid" is the interaction of Overconfid and dummy for treatment ConfidM. Standard errors in parentheses, and * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 4. Determinants of delegation to the algorithm by managers.

Table 4 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 indicates that managers who overestimate their hiring success are less likely to delegate the decision to the algorithm. The interaction term of model (3) shows that higher overconfidence is associated with a significantly stronger treatment effect of ConfidM. Finally, analogous to the transparency treatment for the workers, the regressions confirm that providing details about how the algorithm works does not increase the managers' delegation to the algorithm.

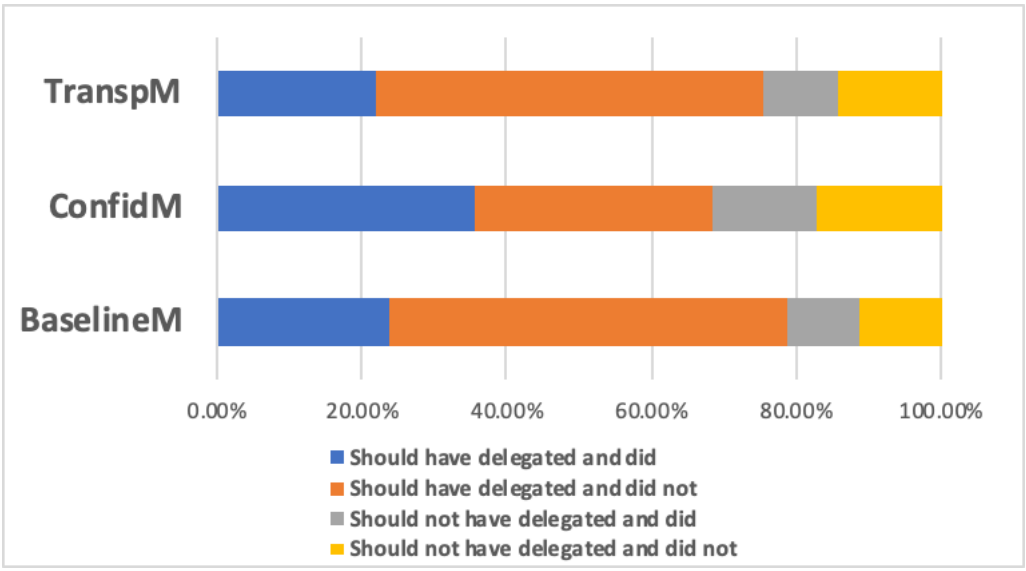


Figure 3. Managers' delegation decisions depending on whether it was optimal or not

Since managers differ in their ability to hire the better worker and in their confidence level, we ask whether managers optimally sort into delegation, and whether our treatments impact the optimality of delegation decisions. Figure 3 shows the breakdown of the managers' decisions to delegate depending on whether it was optimal or not to do so. The proportion of optimal delegation decisions by treatment 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). There are more managers in ConfidM than in BaselineM who should have delegated and did so ($p<0.01$) and who should not have delegated and did not do so ($p=0.06$), as indicated by the longer blue and yellow bars for ConfidM in Figure 3. Furthermore, the proportion of managers who should have delegated and did not (orange bar) is lower in ConfidM than in BaselineM

($p<0.01$). Thus, providing feedback on managers' over- or underconfidence significantly increases the optimality of delegation decisions. 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): *Managers delegate to the algorithm significantly more often and closer to the optimum when they receive feedback on their performance, an effect that is stronger the more overconfident the manager. 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 is a promising result, pointing to a preference for no discrimination (advantageous or disadvantageous) based on gender. We interpret it as direct support for the proposed regulation in the EU that would make illegal any profiling by ethnicity, gender, and other group attributes.

For managers, we replicate the finding of the previous literature that managers delegate decisions to the algorithm too rarely. This costly mistake by managers is caused by overconfidence in their ability to hire the better worker. 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. While mandating the adoption of the algorithm is one way to increase efficiency (as suggested by Glaeser et al., 2021), we show that feedback on past performance might 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 suggests that regulating the transparency of algorithms alone is unlikely to affect the preference for algorithmic decision-making. However, we do not claim that our findings speak against transparency per se, as 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 to be inefficient. Finally, the presentation of the algorithm in our transparency treatments may not be straightforward to understand for some participants. We display the formula used by the algorithm, and we also put it into words. Experimenting, e.g., with graphical representations of algorithms 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 where performance in the job task is perfectly observable ex post. This may overstate the advantages of hiring algorithms relative to environments where the demands on workers are more complex, for example. However, we note that while the observed relative performance of algorithms compared to managers may not 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.

Auster, S. & Pavoni, N. (2021). Optimal delegation and information transmission under limited awareness, ECONtribute Discussion Paper, No. 059, University of Bonn and University of Cologne, Reinhard Selten Institute (RSI), Bonn and Cologne.

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 Review106.5, 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.

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.

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.

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.

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.

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, July). Applicant perceptions of hiring algorithms-uniqueness and discrimination experiences as moderators. In

Academy of Management Proceedings (Vol. 2019, No. 1, p. 18172). Briarcliff Manor, NY 10510: Academy of Management.

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.

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), 2053951718756684.

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*.

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

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.

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 A (online)

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*** (0.007)	-0.094*** (0.010)	-0.157*** (0.042)	-0.443*** (0.011)
Task1 of 1st worker minus 2nd worker	0.161*** (0.004)	0.161*** (0.004)	0.097*** (0.016)	0.153*** (0.015)
Task2 of 1st worker minus 2nd worker	0.113*** (0.005)	0.113*** (0.005)	0.071*** (0.022)	0.190*** (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 2A. 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)	Choice of algorithm (6)
NoGenderW	0.125*** (0.044)	0.126*** (0.044)	0.121*** (0.044)	0.132*** (0.044)	0.142*** (0.044)	0.140*** (0.044)
TranspW	0.040 (0.045)	0.044 (0.045)	0.040 (0.045)	0.040 (0.045)	0.054 (0.045)	0.050 (0.045)
Age		-0.001 (0.001)	-0.000 (0.001)	-0.000 (0.001)	-0.000 (0.001)	-0.000 (0.001)
Male		0.057 (0.037)	0.053 (0.037)	0.030 (0.037)	0.076** (0.038)	-0.152 (0.143)
Task 1			0.010 (0.012)	0.002 (0.013)	-0.000 (0.012)	0.001 (0.012)
Task 2			0.006 (0.009)	0.001 (0.009)	0.003 (0.009)	0.002 (0.009)

Confidence				0.003*** (0.001)	0.003*** (0.001)	0.003*** (0.001)
DiffBeliefAlgo Manager					-0.008*** (0.002)	-0.008*** (0.002)
Male* DiffBeliefAlgo Manager					0.012*** (0.003)	0.011*** (0.003)
BeliefMenTop50						-0.006* (0.003)
Male* BeliefMenTop50						0.008* (0.005)
Observations	744	744	744	744	743	743
R^2	0.011	0.015	0.017	0.028	0.059	0.064
Sample	All	All	All	All	All	All

Notes: OLS regression of choosing the algorithm by workers. "Confidence" is the belief of how many workers out of 100 have lower task-1 and 2 performances than oneself. "DiffBeliefAlgoManager" equals the difference between belief of 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 manager being male. "BeliefMenTop50" is the participants' belief of out of 100 how many of the top 50 workers in terms of performance are men. "Male* BeliefMenTop50" is the interaction of BeliefMenTop50 and dummy for the manager being male. Standard errors in parentheses, and * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 3A. Determinants of choice of the algorithm by workers

	Delegation (1)	Delegation (2)	Delegation (3)
ConfidM	0.157*** (0.043)	0.156*** (0.043)	0.131*** (0.043)
TranspM	-0.019 (0.043)	-0.020 (0.043)	-0.045 (0.044)
Age		-0.000 (0.000)	-0.000 (0.000)
Male		0.043 (0.035)	0.044 (0.035)
Overconfidence			-0.017*** (0.006)
ConfidM*Overconfidence			0.034*** (0.010)
Observations	754	754	752
R^2	0.026	0.029	0.047
Sample	All	All	All

Notes: OLS regression of delegation to algorithm. "Overconfidence" is the difference between belief in how many hires were correct and the actual number of correct hires. "ConfidM*Overconfid" is the interaction of Overconfid and dummy for treatment ConfidM. Standard errors in parentheses, and * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 4A. 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 was not visible for 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 to 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 100 – you have the best score

0 10 20 30 40 50 60 70 80 90 100

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 Only men are hired

0 10 20 30 40 50 60 70 80 90 100

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 Only men are hired

0 10 20 30 40 50 60 70 80 90 100

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

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 to 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 to 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.

	All wrong / All correct 0 2 4 6 8 10 12 14 16 18 20
Number of correctly chosen workers out of 20 pairs ()	

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.**