Biased Humans, (Un)Biased Algorithms?*

Florian Pethig
University of Mannheim
pethig@uni-mannheim.de

Julia Kroenung
European Business School
julia.kroenung@ebs.edu

February 2022

Abstract

Previous research has shown that algorithmic decisions can reflect gender bias. The increasingly widespread utilization of algorithms in critical decision-making domains (e.g., healthcare or hiring) can thus lead to broad and structural disadvantages for women. However, women often experience bias and discrimination through human decisions and may turn to algorithms in the hope of receiving neutral and objective evaluations. Across three studies (n=1,107), we examine whether women's receptivity to algorithms is affected by situations in which they believe that their gender identity might disadvantage them in an evaluation process. In Study 1, we establish, in an incentive-compatible online setting, that unemployed women are more likely to choose to have their employment chances evaluated by an algorithm if the alternative is an evaluation by a man rather than a woman. Study 2 generalizes this effect by placing it in a hypothetical hiring context, and Study 3 proposes that *relative algorithmic objectivity*, i.e. the perceived objectivity of an algorithmic evaluator over and against a human evaluator, is a driver of women's preferences for evaluations by algorithms as opposed to men. Our work sheds light on how women make sense of algorithms in stereotype-relevant domains and exemplifies the need to provide education for those at risk of being adversely affected by algorithmic decisions. Our results have implications for the ethical management of algorithms in evaluation settings. We advocate for improving algorithmic literacy so that evaluators and evaluatees (e.g., hiring managers and job applicants) can acquire the abilities required to reflect critically on algorithmic decisions.

*The authors thank the section editor, Kirsten Martin, and two anonymous reviewers for their constructive feedback and guidance throughout the review process. The first author acknowledges financial support from a doctoral scholarship of the Landesgraduiertenförderung (LGF) funded by the state of Baden-Württemberg.

1 Introduction

In October 2018, *Reuters* reported that Amazon had abandoned its artificial intelligence (AI) tool for automatically screening résumés because it "showed bias against women" (Dastin 2018). According to the article, the tool "penalized résumés that included the word *women's*, as in *women's chess club captain*." Articles on the "sexist AI" followed in several major news outlets, including *The Wall Street Journal*, *The Guardian*, and the BBC.[1] In response, an Amazon spokeswoman announced that "the program was never used to evaluate applicants" (Shellenbarger 2019). Several days later, the Public Employment Service Austria published the specifications of an algorithm for classifying unemployed citizens according to their chances on the labor market. Of particular public interest was the fact that the algorithm predicted lower chances for unemployed women in comparison with men displaying the same characteristics. Again, public response to the algorithm was swift and negative. Headlines such as "sexist algorithm discriminates against women and mothers" (Gučanin 2018) and "computer says no: algorithm gives women less chances" (Wimmer 2018) appeared in the Austrian media for a period of several weeks (see Reiter 2019).

Although these headlines are recent, research on business ethics has been addressing the problematic societal and ethical implications of algorithmic decision-making since the early 1990s, asserting that "any biases held by the knowledge engineer could also influence the way the decisions are made" (Khalil 1993, p. 318). Nowadays, many algorithms learn automatically from historical data, thus introducing additional sources of bias incorporated in the underlying training data (Martin 2019). Because the transparency of these new types of algorithm is limited, even to developers themselves (see Rai 2020), business ethics researchers have pointed to the importance of making companies accountable for the algorithms they develop and/or employ to avoid negative ramifications for groups underrepresented in the data, including women or minorities (Buhmann et al. 2020; Martin 2019). These scholars conclude that algorithms are inherently value-laden and undermine the chances of

stigmatized groups regularly devalued on the grounds of their social identity (Dovidio et al. 2000). The suspicion has been mooted that biased algorithmic decisions may even lead to structural stigma (Hatzenbuehler 2016; Lepri et al. 2017) resulting in disadvantages "outside of a model in which one person does something bad to another" (Link and Phelan 2001, p. 372). Policy implications from these studies yield important insights with respect to the ethical development, implementation, and management of new algorithms in critical decision-making domains such as healthcare, hiring, or recidivism prediction (Leicht-Deobald et al. 2019).

Despite the abundance of studies addressing technological instances of bias and discrimination triggered by algorithms, little empirical attention has been given to the issue of how women perceive algorithmic evaluators as opposed to human evaluators. This context is highly relevant because gender-based discrimination is common in our society (Bohnet et al. 2016; Moss-Racusin et al. 2012), and women are ability-stigmatized in various domains (mathematics, computer science, etc.) (Carlana 2019). According to Martin (2019, p. 847), this is a crucial and, as yet, under-researched sphere of ethical decision-making that warrants greater attention because it provides insights into "how individuals make sense of the algorithm as contributing to the decision." Importantly, understanding women's receptivity to algorithms lays the foundation for improving the algorithmic literacy of those at risk of being adversely affected by algorithmic bias (Cotter and Reisdorf 2020). Moreover, it provides guidance to policy-makers on how to intervene when companies conceal their algorithmic decisions behind a "veneer of objectivity" (Martin 2019, p. 846).

In this paper, we evaluate women's perceptions of algorithmic evaluators over and against human evaluators across different hiring and career-development settings. Although women tend to be disadvantaged by algorithms *and* humans (e.g., Dastin 2018; Shellenbarger 2019), we draw upon literature in the fields of philosophy and the history of science to argue that the former are commonly portrayed as *less* biased and *more* objective than the latter (Carlson 2019; Gunton et

al. 2021). Accordingly, women are likely to turn to algorithms in the hope of overcoming the significant limitations impairing human decision-making. For example, a recent study has found that consumers exhibit strong preferences for human providers over automated healthcare providers (Longoni et al. 2019). But the authors appeal to future research to investigate whether automated providers are "preferred to human providers when treatment requires the disclosure of stigmatized information" (Longoni et al. 2019, p. 48).

To understand the mechanisms prompting women to prefer evaluations by algorithms to evaluations by humans, we report here on three studies conducted via Amazon's Mechanical Turk (MTurk). Study 1 presents evidence indicating that, in an incentive-compatible setting, unemployed women are significantly more likely to choose evaluations of their employment chances by an algorithm when the alternative choice is an evaluation by a man (*outgroup* condition) as opposed to a woman (*ingroup* condition). This result indicates that women expect more favorable evaluations by algorithms in situations where they believe that their gender identity may be under heightened scrutiny by a human counterpart. Study 2 generalizes this effect to a hypothetical hiring context. Finally, in Study 3, we show that the effect generalizes to a career development context and document *relative algorithmic objectivity*, i.e., the difference in perceived objectivity between the algorithmic evaluator and the human evaluator, as an important driver of women's preferences for algorithmic decisions.

Our paper makes two contributions to research. First, we contribute to business ethics research on algorithmic bias and discrimination (e.g., Buhmann et al. 2020; Khalil 1993; Leicht-Deobald et al. 2019; Martin 2019; Munoko et al. 2020) by examining women's receptivity to algorithmic evaluations in domains where gender-based discrimination is prevalent. Our findings shed light on the subtle contextual cues that may prompt women to expect more favorable evaluations from algorithms than from humans. Second, we advance theoretical understanding of the mechanisms that

determine women's receptivity to algorithms. In particular, we propose that in situations where women perceive an algorithmic evaluator as relatively more objective than a human counterpart, they are more likely to choose the algorithm. We label this mechanism relative algorithmic objectivity and provide initial indications that it may be conditional on the gender of the alternative human evaluator. We thus contribute to the growing body of literature on people's perceptions of algorithms, and psychological mechanisms that influence their reliance on those algorithms (e.g., Castelo et al. 2019; Gunaratne et al. 2018; Longoni et al. 2019).

2 Theoretical Background

2.1 People's Receptivity to Algorithms

As technological advancements have facilitated the storage, retrieval, analysis, and sharing of information in a variety of forms, human involvement in many decision-making processes has become obsolete. Algorithms, commonly defined as computational procedures that use certain inputs in order to generate answers (see Logg et al. 2019), have the ability to carry out automated decisions by adapting to, and learning from, historical data.

We proceed from a literature review that we conducted to investigate people's receptivity to algorithms.[2] We analyzed 17 studies and found that individuals regularly prefer to rely on humans rather than algorithms in decision-making contexts. This effect is more pronounced (a) after witnessing algorithm failure (Dietvorst et al. 2015; Prahl and Van Swol 2017), (b) for subjective than objective tasks (Castelo et al. 2019), (c) in moral domains (Bigman and Gray 2018), and (d) among experts than non-experts (Logg et al. 2019). Mediating and moderating mechanisms at the individual level reveal some of the criteria for algorithm rejection. They include uniqueness neglect, i.e., the fear that automated healthcare providers may be unable to take account of individuals'

unique characteristics (Longoni et al. 2019), awareness of the impossibility of offloading responsibility for potentially negative outcomes onto an algorithm (Promberger and Baron 2006), lack of social presence (Langer et al. 2019), and creepiness and discomfort (Castelo et al. 2019; Langer et al. 2019). The inclination to reject a potentially superior algorithm in favor of human judgement has been referred to as "algorithm aversion" (Dietvorst et al. 2015, p. 114).

However, two studies indicate notable exceptions to algorithm aversion. Logg et al. (2019) find that lay people consistently tend to rely on algorithms for numerical tasks with an objective outcome. Similarly, Gunaratne et al. (2018) show that algorithmic investment advice has a greater following than its crowd-sourced alternative, the argument here being that algorithmic authority is more persuasive than the behavior of peers. These results suggest that—at least for certain tasks—individuals may be willing to incorporate the output of opaque algorithms into their own decision-making.

2.2 The Assumption of Mechanical Objectivity[3]

A prominent theme in the previous section was that algorithms tend to be favored for tasks that involve some kind of numerical or quantifiable assessment, i.e., tasks that are viewed as *objective*. Although the meaning of the term objectivity is often taken for granted, no common definition exists (Christin 2016; Gunton et al. 2021). Philosophers of science tend to define the concept by referring to "objective" knowledge as knowledge that is reliable because it is fully detached from the perspective of the person(s) helping to produce it (Gunton et al. 2021). According to this view, objective knowledge is only feasible proposition if we can rule out human agency that might otherwise subjectively influence the interpretation of the collected data. Historically, the idea that objectivity represents an unbiased depiction of reality has put machines at an advantage over humans in producing objective knowledge because technological advances such as photographs

or X-rays have been assessed as displaying "an unmediated representation of natural phenomena" (Christin 2016, p. 30). Daston and Galison (1992) call this *mechanical objectivity*, a notion stemming from the belief that technology is superior to human subjectivity because it is not influenced by values, perspectives, biases, or personal interests (Reiss and Sprenger 2020). In other words, mechanical objectivity reflects a somewhat overenthusiastic and uncritical approach to technology, depicting it as a way of "let[ting] nature speak for itself" (Daston and Galison 1992, p. 81).

Of course, the idea that photographs, X-rays, or even algorithms depict reality "as it is" has long been dismissed (Carlson 2019; Daston and Galison 1992; Gunton et al. 2021). Yet given all that we know about editing photos (Carlson 2019) or the very human (and therefore biased) choices that underlie the development of algorithms (Friedman and Nissenbaum 1996), research so far has shown a remarkable tendency—shared especially by lay people (e.g., Logg et al. 2019)—to trust the output of opaque algorithms in domains where human intervention is viewed as a potential weakness. This indicates that the belief that machines, and especially algorithms, objectively represent a situation is intuitively appealing and often resonates with people's perceptions about machines (Castelo et al. 2019). Accordingly, the concept of mechanical objectivity offers an important lens for understanding how algorithms are perceived over and against humans in situations where people hope that "faithfulness to reality" (Gunton et al. 2021, p. 8) is in their own interests. Building upon these findings, we now illustrate how stigma can influence women's receptivity to algorithms in evaluation contexts.

2.3 Stigma and Algorithms

There has been abundant research in social psychology, sociology, and economics on the question of whether and to what extent individuals interact differently with ingroup and outgroup members and seeking to explain important phenomena such as racial and ethnic conflict, discrimination,

and social exclusion (for an overview, see Chen and Li 2009). One central aspect governing these interactions is whether group members have had a stigma imposed on them—i.e., a social identity that is devalued in a particular context (Crocker et al. 1998; Major and O'Brien 2005; Pescosolido and Martin 2015). As noted in seminal research on the topic, stigma is not located within the person in question but in the fact that that person has an attribute that may lead to devaluation in a certain social context (Crocker et al. 1998; see also Johnson et al. 2011). Contextual cues, such as numerical underrepresentation or the presence of outgroup members (e.g., Inzlicht and Ben-Zeev 2000; Johnson et al. 2011; Parsons et al. 2011), often suffice to indicate to an individual that one of his or her social identities may be the object of devaluation in a certain situation.

Devaluation potential, Steele et al. (2002) posit, is a working hypothesis present in an individual's mind—a *theory of context*—that is activated by contextual cues and may in certain situations affect an individual's vigilance and trust. For example, the image of baseball as a "white man's sport" (Kang 2016; Nightengale 2016), will implant in the minds of ethnic minority pitchers that they may be devalued due to their ethnic identity. An outgroup umpire may then act as a contextual cue for the possibility of devaluation and discrimination, thus heightening their expectations of receiving biased evaluations. As a result, ethnic-minority pitchers will adjust their behavior to encourage less subjective evaluations (Parsons et al. 2011). Similarly, women taking a math test will have in their minds the working hypothesis of female underperformance in math, and the presence of males may then activate the threatening effects of gender stereotypes and undermine their actual performance (Inzlicht and Ben-Zeev 2000).

Attempting to understand how stigma shapes women's receptivity to algorithms takes us back to the concept of mechanical objectivity. In line with this concept, algorithmic decisions as opposed to human decisions will be viewed as "*less* biased without the perceived irrationality, discrimination, or frailties" (Martin 2019, p. 837, emphasis added). Algorithms are perceived as minimizing human

intervention (Christin 2016) and are credited with greater cognitive than emotional abilities (Castelo et al. 2019), which in its turn fuels the perception that they are more successful in objective tasks than subjective tasks. Thus, although humans are generally perceived as better qualified than algorithms for evaluation procedures such as hiring employees, predicting recidivism, or diagnosing a disease (for an overview, see Castelo et al. 2019), the perception that their gender identity is likely to play a role in the assessment process may reverse women's preferences. If it is felt to be likely that subjective and irrational judgements may disadvantage the person being assessed, human ability to solve subjective tasks may not be seen as an advantage over the "mechanical objectivity" of algorithms but rather as a disadvantage. From the perspective of women, reports of gender bias ingrained in algorithms are therefore unlikely to outweigh the daily experiences of bias and discrimination displayed by co-workers or hiring managers (Bohnet et al. 2016; Moss-Racusin et al. 2012). In the following, we outline three hypotheses on how women perceive algorithms in areas where gender bias is prevalent and adversely affects opportunities for women.

3 Hypotheses

Gender bias in hiring has been widely documented independently of the type of evaluator (human vs. algorithm) (e.g., Bohnet et al. 2016; Dastin 2018). Stereotypes are a good explanation for gender bias, as society often views women as less competent than men in business or in science-related domains (Moss-Racusin et al. 2012; Walton et al. 2015). Such oversimplified generalizations of their own social group may sensitize women to cues suggesting that they should anticipate devaluation in these domains. One such cue is the presence of, or evaluation by, men (Inzlicht and Ben-Zeev 2000; Pinel 2004). By contrast, existing reports of algorithmic bias to the detriment of women (e.g., Dastin 2018) have done little to affect the tendency to view algorithms as being less biased than humans (Martin 2019). In conjunction with Steele et al.'s (2002) theory of context, this implies

that women will be more likely to suspect unfair treatment from a man than an algorithm. Men are expected to be more biased and to employ their decision-making powers to the disadvantage of women. By contrast, if the human evaluator is female, women may hope that the ingroup evaluator's decision will be more favorable for them, which makes it less likely that they will choose the algorithm. For this reason, we suggest that women choosing between an algorithmic evaluator and a male evaluator are more likely to go for the algorithm than women choosing between an algorithmic evaluator and a female evaluator.

Hypothesis 1 (H1). *Women prefer an algorithmic evaluator when the alternative is a male (vs. female) evaluator.*

In the past, photographs were believed to depict the "unfiltered" reality. Although this view has not stood the test of time (Daston and Galison 1992; Gunton et al. 2021), people's tendency to believe that modern technologies, such as algorithms, will remove human subjectivity from decision-making processes is still widespread (e.g., Castelo et al. 2019; Logg et al. 2019). Accordingly, people are more likely to trust and use algorithms in situations where algorithms are perceived to be objective and qualified (Castelo et al. 2019).

This means that we will expect women to be most likely to believe that algorithms will arrive at objective decisions when in their view there is a good chance that an alternative human evaluator will be biased against them. Consider the case of Susan. She is applying for a new job because her male supervisor promoted her male co-workers over her head although she was better qualified. Compared to a male hiring manager, Susan may well view a hiring algorithm as the better choice given that the technology involved should enable the algorithm to accurately assess the value of her qualifications. Susan is likely to select the algorithm because she perceives the algorithm as being relatively more objective than a male hiring manager who may be just as prejudiced as her supervisor. By contrast, if the hiring manager is a woman, Susan may be less likely to perceive

the algorithm as relatively more objective than the human because she may expect a female hiring manager to consider her qualifications and not the fact that she is a man or a woman.

This behavior is well-explained by social identity theory (Tajfel 1981). In general terms, outgroup members are often seen as homogeneous, prejudiced, and susceptible to the normative obligation to favor the ingroup (Vivian and Berkowitz 1992). Ingroup members, by contrast, are expected to be more resistant to pressures favoring the ingroup and to produce more objective decisions. This suggests that an algorithm may appear to be relatively more objective than a prejudiced outgroup member, but not more objective than an apparently objective ingroup member. In line with this train of thought, we argue that when women choose between an algorithm and a male evaluator, as opposed to an algorithm and a female evaluator, women will perceive higher relative algorithmic objectivity, defined as the perceived objectivity of the algorithmic evaluator relative to that of the human evaluator. In sum, we suggest that the presence of a male (vs. female) evaluator increases the perception of an algorithm's higher relative objectivity and, in turn, increases reliance on it. Accordingly, we propose

> **Hypothesis 2 (H2).** *Women perceive higher relative algorithmic objectivity when the alternative is a male (vs. female) evaluator.*

> **Hypothesis 3 (H3).** *The preference for an algorithmic evaluator is mediated by higher relative algorithmic objectivity.*

4 Research Overview

Figure 1 shows the research framework. In Studies 1-3, we systematically examined whether a contextual cue in the form of a male evaluator, as opposed to a female evaluator, is likely to increase the probability of women choosing an algorithmic evaluator over a human evaluator (H1).

In Studies 1-2, we used a binary choice variable to measure women's receptivity to algorithms, and in Study 3, we used relative preference between the algorithm and the human as the dependent variable. Furthermore, in Study 3, we focused on the role of relative algorithmic objectivity as a mediator between the gender of the human evaluator and the decision to be evaluated by an algorithm as opposed to a human (H2 and H3).

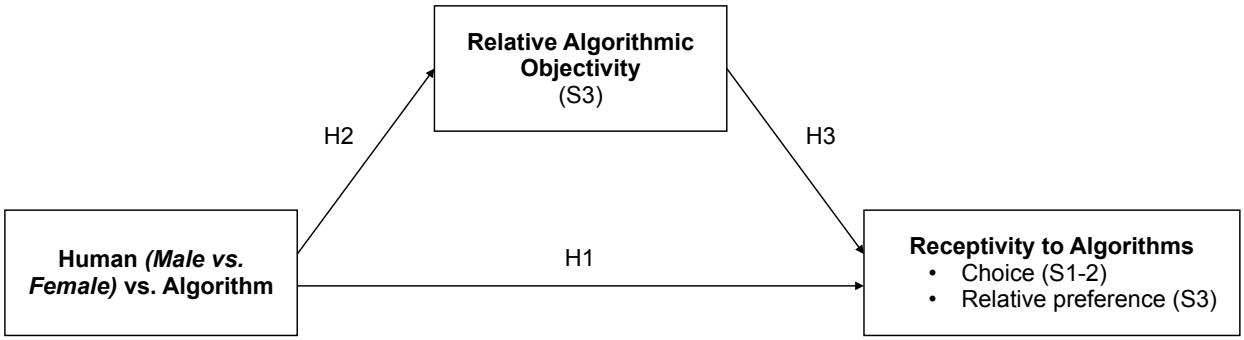

Figure 1: Research Framework

Note: S1 = Study 1; S2 = Study 2; S3 = Study 3.

Studies 1 to 3 were conducted on MTurk and were administered via Unipark. From MTurk, we recruited participants who resided in the United States and had completed at least 500 previous tasks with an approval rate of at least 95%. These criteria are in line with what earlier literature has suggested to ensure high data quality from MTurk (Peer et al. 2014). Furthermore, we included attention checks at the beginning and comprehension checks at the end, and MTurk participants were not allowed to participate in more than one study (except for the screening survey in Study 1).

Responding to a recent call for more transparent research practices in business ethics research (Roloff and Zyphur 2019), we report all data exclusions, manipulations, and measures. Survey materials, data, code for data analysis, and preregistration documents for Studies 2 and 3 are available on the Open Science Framework (OSF).[4] We discuss Studies 1-3 in the order in which we conducted them. We report the data analysis in Jupyter Notebook showing the actual results

that we generated with our code. We utilize the interactive computational environment of Jupyter Notebook, including markup language, as a guide to our analysis. We also provide a link below each figure to the corresponding code that generates that figure. We evaluate the impact of our randomized treatments using the appropriate statistical tests for the variables of interest. Specifically, we use a chi-square test for the binary outcome variables in Studies 1 and 2 and t-tests for the Likert-scale-dependent variables in Study 3. In Tables A1 and A2 in Appendix A, we show that the results are consistent when using ordinary least squares (OLS) regressions and including controls for demographics. For the preparation of our final manuscript, we used RStudio (RStudio Team 2021) together with the *rmarkdown* package (Allaire et al. 2021) and the *bookdown* package (Xie 2020).

4.1 Study 1

4.1.1 Overview

Study 1 focuses on H1. The setting is the evaluation of unemployed women's chances on the labor market. We chose this scenario because women are often stereotype-disadvantaged on the labor market (e.g., Bohnet et al. 2016; Moss-Racusin et al. 2012; Riach and Rich 2002), so unemployed women may legitimately reckon with potential discrimination from a male evaluator. We asked unemployed women whether they believed an algorithm or a human would give them better chances of finding a job within the following six months. We manipulated whether participants had the choice between an algorithm and (a) a male evaluator (outgroup condition) or (b) a female evaluator (ingroup condition). We expected women choosing between an algorithm and a male evaluator to be more likely to select the algorithm than women choosing between an algorithm and a female evaluator. We were also curious to see how unemployed men would respond when asked to choose between an algorithm and a male evaluator over and against an algorithm and a female evaluator.

4.1.2 Participants

We recruited 4,857 participants via MTurk of which 3,352 correctly answered the attention check (69%). Participants passing the attention check earned $0.10 for responding to a short screening survey that included questions on gender, ethnicity, and employment status. 269 participants self-identifying as white, non-Hispanic, and unemployed were carried forward to the second part of the survey. To ensure that employment was a desirable outcome,[5] we only retained respondents who were currently looking for a job. This left us with a sample of 145 participants who were offered the chance to participate in a voluntary bonus task for an additional $0.50. During the bonus task, participants had to decide whether their data should be evaluated by an algorithm or a human and we manipulated the gender of the human evaluator. 136 participants agreed to participate in the bonus task, and 120 correctly selected the gender of the human evaluator at the end of the survey. *Post hoc*, we removed two participants who identified themselves as neither male nor female.[6]

Our final sample encompassed 76 women and 42 men, whose average age was 39.2 years. Average duration of unemployment was 40.5 months, and, on average, participants had been looking for work for 8.5 months. Most participants had worked previously in health care or retail (each 14.4%), followed by finance and insurance (8.5%).

4.1.3 Procedure

The study began with a demographic survey on the participants' gender, age, ethnicity, employment status, education, obligations to care for family members, and disability.[7] Participants who matched our selection criteria (i.e., white, non-Hispanic, and unemployed) were asked which sector their last job (if any) was in, whether they were currently looking for a job, duration of unemployment, and duration of job search. Next, participants were given the opportunity to participate in a bonus task. At the beginning of the bonus task, we presented them with their answers to the previous

questions and asked whether they would authorize us to evaluate their employment chances on the basis of that data. Table B1 in Appendix B shows the profile of a fictional participant who would have matched our selection criteria.

The bonus task informed participants that they would be randomly matched with another unemployed participant from our sample and that they would receive $0.50 if they were granted a higher chance of finding employment in the next six months than the randomly matched participant. To have their employment prospects assessed, participants could choose between an algorithmic evaluator and a human evaluator ("use the human [algorithm] to determine my bonus"). Furthermore, we informed participants that if the randomly matched participant chose a different evaluator, each of them would be paid according to the evaluation outcome of their chosen evaluator. For example, if the initial participant chose the human evaluator and the other person the algorithmic evaluator, the initial participant would be paid in accordance with the human evaluation and the other person in accordance with the algorithmic evaluation.

If participants (voluntarily) chose to participate in the bonus task, they were assigned to one of two conditions: in the outgroup condition, the gender of the human evaluator was male, and in the ingroup condition, the gender of the human evaluator was female (vice versa for male participants). To make the gender attribute of the human evaluator less salient, we followed Bohnet et al. (2016) and displayed three different attributes of the human evaluator: profession, gender, and ethnicity. Profession (research assistant) and ethnicity (Caucasian) were held constant across both conditions. After making their choice, participants answered one open-ended question asking participants why they chose to have their bonus determined by the algorithm or the human evaluator, depending on which they had chosen. Lastly, they were asked what the gender of the human evaluator was. After the study was over, we awarded a $0.50 bonus to all participants in the bonus task and informed them that we would not be evaluating their data.

4.1.4 Results and Discussion

In line with our hypothesis, women were found to be significantly more likely to choose an algorithmic evaluation if the alternative choice was an evaluation by a man as opposed to a woman (see Figure 2). Whereas 66% chose the algorithm to evaluate their bonus in the outgroup condition, only 39% chose to use the algorithm in the ingroup condition ($\chi^2(1,\ N = 76) = 5.40,\ p = 0.020$). Moreover, 71% of the men assigned to an outgroup human evaluator chose the algorithm, while 60% of the men assigned to an ingroup evaluator chose the algorithm. These two conditions did not differ significantly ($\chi^2(1,\ N = 42) = 0.49,\ p = 0.482$).

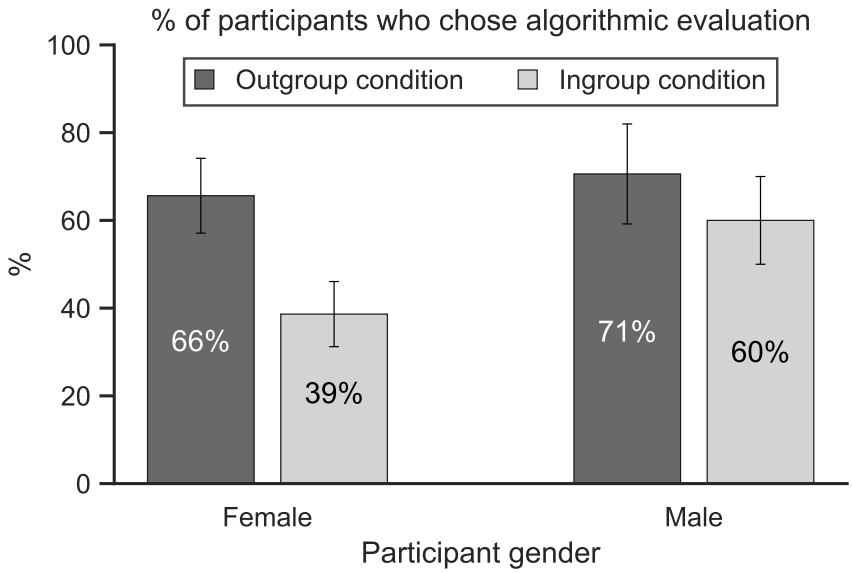

Figure 2: Women Were Significantly More Likely to Prefer an Algorithmic Evaluator When the Alternative was a Male Evaluator as Opposed to a Female Evaluator (Study 1)

Note: Error bars represent ± standard error. The Python code is available from https://osf.io/k89cg/.

Unemployed women who had the choice between an algorithm and a male evaluator were more likely to choose the algorithm than women who had the choice between an algorithm and a female evaluator. These results lend support to H1 and suggest that unemployed women expect more favorable treatment from algorithms when the alternative choice is an outgroup human evaluator

as opposed to an ingroup human evaluator. For unemployed men, we found a directional but non-significant difference between both conditions. We also found that in both conditions unemployed men tend to opt for the algorithmic evaluation.

4.2 Study 2

4.2.1 Overview

So far, we have documented the outcome that when the alternative was a male rather than a female evaluator, unemployed women were more likely to choose an algorithmic evaluator. In Study 2, our aim was to generalize the finding to a hypothetical hiring scenario, a context where the (mis-)use of algorithms is increasingly widespread. In addition, we wanted to ensure that our findings were not idiosyncratic to the specific sampling population of unemployed women and also that they were robust to the use of a much larger sample size. In the study, the participants were faced with a hypothetical scenario and were asked to imagine that they were applying for their dream job in this scenario. They were required to answer three demographic questions (age, gender, education), and they were told that the answers were used by the company to evaluate their application. As in Study 1, we manipulated the gender of the human evaluator (male vs. female), and once again, we expected women to be more likely to select the algorithm if the alternative evaluator was a man rather than a woman.

4.2.2 Participants

Via MTurk we recruited 1,100 participants via MTurk of whom 784 participants passed the attention check (71%). Participants were paid $0.50 for completing the study. If we found that a single MTurk ID participated in our survey more than once, we preregistered to include only the response with the earliest timestamp, leading to 16 exclusions. We also removed 306 participants who failed

at least one of the two comprehension checks at the end of the survey, and 3 participants, who identified as neither male nor female. The final sample consisted of 164 women and 226 men (mean age = 38.4).

4.2.3 Procedure

The survey began by telling the participants that the study would be asking them to decide whether their hypothetical application should be evaluated by an algorithm or a human. On the next page, they were presented with an attention check question indicating whether they had understood this instruction. If they failed to answer correctly, the survey excluded them from participating, and we collected no further data from them. Participants who answered correctly were then given a scenario to read in which they were asked to imagine applying for a new job. The scenario read as follows:

> Imagine that you have recently lost your job due to the global pandemic and are looking for new opportunities. During your search, you find an online advertisement for the job you have long dreamed of, and you decide to apply for it. On the website you read that the company has introduced a new system for hiring employees. This system allows applicants to decide whether their data should be evaluated by a human or by an algorithm. According to the company, the new system allows applicants to increase their chances of being hired by choosing the option (human or algorithm) that they think will grant them the highest chances of getting the job.

After reading the scenario, participants were asked to complete the application process by providing information on their age, gender, and education. On the next page, they were told that they had received an email from the company, asking them to choose whether their application should be evaluated by a human or an algorithm ("Who do you want to be evaluated by?"). This question served as our dependent variable. Participants were assigned to an ingroup or outgroup condition. We manipulated the gender of the human evaluator by giving them the option to choose between a pictogram of an algorithm along with the caption "Algorithm" and a pictogram of a man

[woman] along with the caption "Human (Male [Female])" (see Figure B1 in Appendix B). On the same page, participants were asked to briefly substantiate their decision.

At the end of the survey, participants answered two comprehension-check questions to ensure that they had properly understood the scenario. The first question asked them about the gender of the human evaluator depicted (male vs. female) and the second question asked them about the benefit the company claimed to have achieved in connection with the new system ("increase the chances of applicants to be hired").

4.2.4 Results and Discussion

In line with Study 1, 46% of women stated that they would choose to have the algorithm evaluate their application in the outgroup condition and only 25% stated that they would choose the algorithm in the ingroup condition ($\chi^2(1,\ N = 164) = 8.09,\ p = 0.004$).[8] 36% of the men stated that they would choose to have the algorithm evaluate their application in the outgroup condition, while 26% stated that they would choose the algorithm in the ingroup condition ($\chi^2(1,\ N = 226) = 2.73,\ p = 0.098$). In contrast to Study 1, this difference is marginally significant. It is, however, much smaller than the difference observed in the women.

Taken together, Studies 1 and 2 provide support for H1, demonstrating that when the alternative choice was a man as opposed to a woman, women were more likely to favor an algorithm. In the next study, we look at relative algorithmic objectivity as a driver of women's receptivity to algorithms.

4.3 Study 3

4.3.1 Overview

Study 3 tested the process by which the gender of the human evaluator affects women's receptivity to algorithms. Accordingly, it focuses jointly on H1, H2, and H3. Our goal for H1 was to further

probe the findings of Studies 1 and 2 by assessing women's relative preference for an algorithmic evaluator over and against a human evaluator in a career development setting. In contrast to the two previous studies, we used a seven-point Likert scale instead of binary choice as the dependent variable. Our goal for H2 and H3 was to focus on the role of relative algorithmic objectivity as a mechanism that drives women's receptivity to algorithms. Individuals have a clear tendency to believe that algorithms are able to overcome the restrictions imposed by human subjectivity (Carlson 2019). This being the case, we expected women to perceive algorithms as more objective than humans in situations where they believe that a human evaluator might disadvantage them. We adapted the four-item scale of self-perceived objectivity from Uhlmann and Cohen (2007) to measure the perceived objectivity of the human evaluator and the algorithmic evaluator (Table 1). To construct our measure of relative algorithmic objectivity, we subtracted the average score of perceived objectivity for the human evaluator from that of the algorithmic evaluator. We expected that women would perceive higher relative algorithmic objectivity when given the choice between an algorithm and a man as opposed to an algorithm and a woman. Furthermore, we expected relative algorithmic objectivity to function as a mediator between the outgroup and ingroup conditions and the relative preference between a human evaluator and an algorithmic evaluator.

4.3.2 Participants

Via MTurk we recruited 1,141 participants, 774 of whom correctly answered the attention check question (68%). Participants earned $0.50 for completing a short scenario with follow-up questions. If a single MTurk ID participated in our survey more than once, we preregistered to include only the first response, resulting in 7 exclusions. In addition, we removed 94 participants who failed the comprehension check at the end of the survey, and 6 participants, who identified as neither male nor female. The final sample consisted of 288 women and 311 men. The average age was 40.3.

4.3.3 Procedure

Participants completed an online survey. They read the same introductory text as in Study 2 and completed the same attention check question.[9] Afterwards, they were asked for their demographic information. On the next page, they read the following scenario, in which age and gender of the respondent were displayed in accordance with the answers to the demographic questions on the previous page.

> Imagine that you are a 35-year-old female employee. Your company has created a new business unit for which it is looking for internal employees. You consider a role in this new business unit as an opportunity to develop your career and decide to apply for it. The process requires you to send an internal application to human resources (HR). They review your résumé and, depending on their evaluation, it is forwarded to the head of the new business unit. Before you submit your internal application, you read the following message:
>
> *We have recently implemented new software that uses an algorithm to evaluate the skills of our employees. You can now choose whether your application should be evaluated by the algorithm or by a human. Please choose the option that you believe will give you the highest chance of successfully advancing your career within our company.*

On the same page, they found the profile of the human resources (HR) manager (adapted from Espino-Pérez et al. 2018), which consisted of four attributes: position, number of years with the company, gender, and age. We randomly switched the positions of age and gender to ensure that our findings were robust to the presentation order of the attributes. As in Study 1, position (HR manager), number of years with the company (5 years), and age (39 years) were held constant across both conditions. We chose 39 years as the age of the HR manager because it was closest to the mean age of participants in Studies 1 and 2. Our dependent variable was a single question measuring the relative preference between the human and the algorithm, a question adapted from Longoni et al. (2019) ("Would you like the human or the algorithm to evaluate your application?"). Participants answered on a seven-point Likert scale with labeled neutral midpoints (1 = definitely the human; 4 = indifferent between the human and the algorithm; 7 = definitely the algorithm). At the bottom of the page, we asked participants to explain their preference.

Table 1: Measures for Relative Algorithmic Objectivity in Study 3

In the scenario presented, I expected the human's [algorithm's] judgement	
Item	Scale
1. ...would be reasonable and logical.	1 = very strongly disagree; 11 = very strongly agree
2. ...would objectively consider all of the facts.	1 = very strongly disagree; 11 = very strongly agree
3. ...would be based on logical analysis.	1 = very strongly disagree; 11 = very strongly agree
4. ...would be rational and objective.	1 = very strongly disagree; 11 = very strongly agree

On the following two pages, participants responded to four items each on the perceived objectivity of the algorithmic evaluator and the human evaluator respectively (Table 1). All items were adapted from the self-perceived objectivity scale (Uhlmann and Cohen 2007). We randomized the order in which participants filled out the two scales to mitigate any priming effects that might occur if participants always completed one scale before the other. We also randomized the order of the items in each scale. Finally, participants answered one comprehension check question in which they were asked to indicate the correct gender of the human evaluator.[10]

4.3.4 Results and Discussion

Our results showed that, although smaller than in Studies 1 and 2, there was a significant difference in relative preference between the outgroup and ingroup condition (3.52 vs. 3.02, $t(288) = 1.97$, $p = 0.0498$). This result indicates that H1 also holds in a career development setting, as the women displayed a higher relative preference for the algorithm when the alternative was a male HR manager as opposed to a female HR manager. For the men, we found no significant difference between the outgroup and ingroup condition (3.01 vs. 2.66, $t(311) = 1.59$, $p = 0.114$).

To construct single measures, we averaged the items pertaining to perceived objectivity of the human evaluator ($\alpha = 0.95$) and perceived objectivity of the algorithmic evaluator ($\alpha = 0.91$). We subtracted the averaged measure of perceived human objectivity from that of perceived algorithmic objectivity so that a higher score indicates higher relative algorithmic objectivity. Women showed marginally higher relative algorithmic objectivity in the outgroup condition than in the ingroup

condition (0.60 vs. -0.27, $t(288) = 1.88$, $p = 0.061$). Notably, the signs of both values were in line with our expectations. Relative algorithmic objectivity was positive in the outgroup condition but negative in the ingroup condition.

To test our mediation hypothesis, we used Model 4 of the PROCESS macro (release 3.5.3) for SPSS (version 27) with 20,000 bootstraps (Hayes 2017). The result was marginally significant at a 94% confidence interval (CI) that excluded 0 (indirect effect = 0.284, 94% $CI = [0.002, 0.573]$). Specifically, relative algorithmic objectivity increased with the presence of an outgroup vs. ingroup evaluator ($a = 0.868$, $SE = 0.461$, $p = 0.061$), and the greater the relative algorithmic objectivity became, the more likely women were to prefer the algorithm to the human ($b = 0.328$, $SE = 0.026$, $p < 0.001$). Because the direct effect of outgroup vs. ingroup evaluator on relative preference was not significant, we have evidence of full mediation ($c' = 0.211$, $SE = 0.203$, $p = 0.298$). For the men, we found a marginally significant difference in relative algorithmic objectivity (0.35 vs. -0.34, $t(311) = 1.79$, $p = 0.074$), and the mediation test was marginally significant at a 92% CI with 20,000 bootstraps (indirect effect = 0.220, 92% $CI = [0.006, 0.448]$).

All told, our results tend to support the outcome that women find algorithms versus men more objective than algorithms versus women and that this perception increases women's preferences for an algorithmic evaluator. In the following section, we discuss the implications of the findings in Studies 1-3.

5 General Discussion

The overall aim of our paper is to explore how women make sense of algorithmic evaluations in situations where they may be disadvantaged on the grounds of their gender identity. First, Studies 1-3 confirmed H1, indicating that women anticipated less biased evaluations from algorithms when they had the choice between an algorithmic evaluator and a male evaluator over and against a

choice between an algorithmic evaluator and a female evaluator. Second, we identified relative algorithmic objectivity as a driver of women's receptivity to algorithms. Study 3 showed that relative algorithmic objectivity mediated women's relative preferences between an algorithmic evaluator and a human evaluator.

5.1 Theoretical Implications

First, we contribute to the research on algorithmic bias and discrimination by shedding light on women's perspective on algorithmic decision-making. Earlier work has mostly focused on the ethical management of algorithms (Lepri et al. 2017; Munoko et al. 2020), uncovering problematic biases in the underlying training data and organizational structures (predominantly male developers, etc.) (Demetis and Lee 2018; Leicht-Deobald et al. 2019). This important stream of research has primarily been drawn upon in developing policy implications for companies and developers (Buhmann et al. 2020; Khalil 1993; Martin 2019). Our study of women's perceptions of algorithmic evaluations was inspired by the reports of adverse affects of algorithms on women's career opportunities, for example, in the delivery of STEM (science, technology, engineering and math) advertisements (Lambrecht and Tucker 2019) or in hiring processes (Dastin 2018). In particular, our results show that subtle contextual cues, such as the gender of an alternative human evaluator, can have severe implications for women's receptivity to algorithmic evaluations in domains where their gender identity may be under greater scrutiny. When companies implement algorithmic assessments as an alternative to human assessments, women may thus tend to prefer the use of the algorithm because they perceive the alternative as less favorable. Accordingly, our study reflects the increasing importance of business ethics research on algorithmic literacy (Cotter and Reisdorf 2020) and the hazards of people being tricked into accepting algorithmic evaluations. Because gender discrimination is widespread in our society, future research should consider how women can be alerted to the fact

that algorithms often reflect the same biases as those present in human decision-making processes (Martin 2019).

Second, more broadly, we contribute to the interdisciplinary literature on the receptivity to algorithms in decision-making situations. Although earlier research has shown that individuals are reluctant to choose algorithms over their own judgement (e.g., Dietvorst et al. 2015), the judgement of friends (e.g., Yeomans et al. 2019), or judgement from healthcare professionals (e.g., Longoni et al. 2019), our results indicate that it may be different if individuals feel threatened that their group membership may disadvantage them in an evaluation process. Drawing upon Steele et al.'s (2002) theory of context, our focus on women's receptivity to algorithms in stereotyped domains fills a current gap in the broader literature on consumers' decision to rely on algorithmic judgements. Our results indicate that the complex and dynamic social contexts in which algorithms are embedded may be a fruitful avenue for future studies to explore. More specifically, our study underlines the necessity, when making assumptions about the acceptability of algorithms for evaluation tasks, of bearing in mind the fact that women are frequently disadvantaged in non-algorithmic evaluation settings. More generally, our empirical findings can serve as a suitable springboard for future research on the acceptance of algorithms not only for quantitative forecasting (e.g., Gunaratne et al. 2018; Logg et al. 2019), but also for more subjective evaluation tasks.

Third, we identify and test a novel psychological driver of women's receptivity to algorithms, namely relative algorithmic objectivity. As pointed out in earlier conceptual work (Carlson 2019), people tend to believe that algorithms are able to overcome the limitations of human subjectivity. This in its turn increases their receptivity to algorithmic evaluations. We build on this argument in our paper and define relative algorithmic objectivity as the difference between the perceived objectivity of algorithmic and human evaluators. Moreover, we provide initial empirical evidence that the gender of an alternative human evaluator may affect how objectively women perceive

algorithms compared to humans in situations where their gender identity is stigmatized. Thus, we enlarge on earlier research and its cautions about a "veneer of objectivity" (Martin 2019) by showing that relative algorithmic objectivity may be particularly pronounced when women believe that their gender identity is subject to heightened scrutiny on the part of the alternative (human) evaluator. In addition, we establish relative algorithmic objectivity as a mediator, thus extending work on mechanisms influencing receptivity to algorithms in critical evaluation settings (e.g., Langer et al. 2019; Longoni et al. 2019). While in the past algorithmic objectivity has been proposed speculatively as an explanation of why people tend to rely on algorithms rather than humans (Christin 2016), our work now provides initial empirical support for this perspective.

5.2 Practical Implications

Our work has practical implications for women, policy-makers, and organizations. First of all, structural stigma is widespread in society, and it is important for women to remain aware of potential discrimination from algorithms. Many of these algorithms are utilized in key decision-making domains but operate inconspicuously in the background. This is highly alarming because it means that the victims of discrimination have no readily identifiable culprit in the traditional sense. An algorithm cannot be reprimanded, fired, or taken to court for exhibiting bias. For women, the discrimination ingrained in algorithms used across different organizations could thus turn out to be more problematic than biased humans—especially if algorithms are thought to be impartial and objective. Stigmatized groups in general and women in particular, should therefore strive to find out all they can about the pitfalls of algorithmic evaluation so that they can put their finger on potential bias in the results of those evaluations. Additionally, their perspectives are a valuable component in public debate on algorithmic bias, contributing accounts of stigmatization and discrimination experienced in real life.

Second, for policy-makers it is important to foster algorithmic literacy in society through educational programs so that the pros and cons of utilizing algorithms for evaluation purposes stand out more clearly. This will help applicants to understand how their data is evaluated and is equally important for HR managers to understand potential bias in their training data. Algorithmic literacy, the ability to understand and reflect on algorithmic decisions, is considered a key skill in today's information society (Cotter and Reisdorf 2020) and alleviates the danger of widespread manipulation through algorithms. On a par with education on data privacy, algorithmic literacy should be appreciated as a skill of crucial significance for interaction with digital technologies.

Third, our findings are both good and bad news for organizations intending to employ algorithms for evaluation purposes and fear backlashes. We find that some people are generally open to having their résumés screened by algorithms, especially if the alternative is an outgroup member. However, particularly in our hypothetical scenarios in Studies 2 and 3, we find that human evaluators are generally preferred to algorithmic evaluators across both genders and conditions. In Study 2, only 33% preferred to be evaluated by the algorithm ($n = 390$), and in Study 3 the average relative preference for the algorithm was below the midpoint (3.03 out of 7, $n = 599$). A practical indication for organizations may therefore be to use algorithmic and human evaluations in concert so as to facilitate the anticipation and identification of biased evaluations.

5.3 Limitations and Future Research

One limitation of our study has to do with the experimental nature of our work. Our settings do not reflect the high-stakes settings typical of organizational contexts. To increase the ecological validity of our studies, we used actual behavior as the dependent variable in Study 1. Generally, future work might consider employing field experiments (Nelson et al. 2020) to shed further light on women's perceptions of algorithms in real-world evaluation scenarios.

The experimental setup also limits the sample to a specific geographic location, the United States, which may not be representative of other populations. Additionally, collecting data via MTurk may be problematic for reasons that are idiosyncratic to the use of crowdworking platforms, including (1) pay, (2) inattentiveness, or (3) representativeness. We tried to offset these limitations in several ways: First, most studies paid participants more than the federal minimum wage. For example, in Studies 2 and 3, we paid participants $0.50 for a short scenario that took them on average 3.74 minutes (Study 2) and 3.80 minutes (Study 3) to complete, the equivalent of $8.02/hour and $7.88/hour, respectively. We also checked our MTurk requester profile on the TurkerView[11] platform, which allows MTurkers to rate the quality of their requesters. The feedback we received confirmed that our pay was viewed as generous by many MTurkers. Second, in all studies, we used questions to exclude participants who did not pay attention. Notably, in Studies 2 and 3, we used highly effective attention checks corresponding directly with the content of the studies. This avoids the common problem that MTurk respondents are experienced in detecting these checks, which makes them less effective. Additionally, we used comprehension checks at the end of all studies to ensure that our manipulations were successful. Also, we used single scenarios in all studies to mitigate the effects of respondent fatigue. Third, in this paper, we only report MTurk studies with experimental methods (as opposed to simple, cross-sectional surveys). Experiments have been shown to work particularly well on MTurk and earlier studies have confirmed the replicability of experimental findings generated on MTurk with nationally representative samples (e.g., Berinsky et al. 2012; Coppock 2019).

Lastly, although companies increasingly use algorithms to evaluate applicants and such software is popular on the market (e.g., pymetrics[12]), it is doubtful whether companies will voluntarily allow individuals to choose between an algorithmic and a human evaluator, or inform them that their data had been evaluated by an algorithm. Nevertheless, there is growing public awareness of the

need to hold organizations and governments accountable for their use of algorithms (Martin 2019; Taylor 2019). Accordingly, our study reveals important consumer perceptions of algorithms that can support governments and organizations.

6 Conclusion

In this paper, we examine women's perceptions of algorithmic evaluators as opposed to human evaluators in settings where their gender identity may be stigmatized. Although earlier research has reported that people tend to be against the use of algorithms in evaluation settings, our results support the notion that women's receptivity to algorithmic evaluations in stereotype-relevant domains is sensitive to the gender of the alternative human evaluator (male vs. female). Drawing upon related research on mechanical objectivity and stigma, this study thus provides initial empirical evidence on women's receptivity to algorithms that has implications for the ethical management of algorithms in such evaluation settings as the hiring of employees.

Notes

1. The cited media coverage is available from https://osf.io/24psu/.
2. The literature review is available from https://osf.io/9a8ny/.
3. This title is borrowed from Carlson (2019).
4. Data for Study 1 is available from the first author upon request. All other materials are available from https://osf.io/axgp2/.
5. This question is used by U.S. Bureau of Labor Statistics to identify individuals who are considered to be marginally attached to the labor force and available to take on a job. See https://www.bls.gov/cps/definitions.htm (accessed February 2023).
6. We did not screen out participants who identified as neither male nor female and removed them after they had participated in the study because they could not be assigned to an ingroup condition.
7. The demographic survey is based on a set of items developed by the Public Employment Service Austria, which has developed a model designed to ascertain unemployed citizens' job prospects. See https://www.ams-forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf (accessed February 2023).
8. We originally preregistered to collect data from 500 respondents. After data collection, we found that 37% of participants had failed our second comprehension check ("Which advantage did the company claim regarding the new system?"). Combined with the fact that we had only 39% of women in our sample, the sample size was significantly lower than we had anticipated, and we found no significant difference between the two conditions (44% vs. 31%, $\chi^2(1, N = 119) = 2.23$, $p = 0.136$). However, when we included those participants who had correctly answered the first

comprehension check ("What was the gender of the human evaluator?") but failed to correctly answer the second comprehension check, we found a significant difference between both conditions (47% vs. 27%, $\chi^2(1, N = 194) = 7.76$, $p = 0.005$). After careful consideration, we therefore decided to recruit 200 additional participants (in a single batch) and adhere to our plan to remove all participants who failed either one of the two comprehension checks. This is the result that we report in this paper. Incidentally, if we include all participants who failed the second comprehension check in the final sample, the result remains strong and highly significant (48% vs. 26%, $\chi^2(1, N = 267) = 14.34$, $p < 0.001$).

9. We have no concerns about using the same attention check question because we prevented participants from Study 2 from participating in Study 3 and the success rate of the attention check question in Study 3 (68%) was below that of Study 2 (71%).

10. Participants also answered two items on perspective taking, which were, as preregistered, collected for exploratory purposes only and are not further discussed in this study.

11. See https://turkerview.com.

12. See https://www.pymetrics.ai.

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., et al. (2021). *rmarkdown: Dynamic documents for r*. https://github.com/rstudio/rmarkdown

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis*, *20*(3), 351–368.

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

Bohnet, I., Van Geen, A., & Bazerman, M. (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, *62*(5), 1225–1234.

Buhmann, A., Paßmann, J., & Fieseler, C. (2020). Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse. *Journal of Business Ethics*, *163*(2), 265–280.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, *134*(3), 1163–1224.

Carlson, M. (2019). News Algorithms, Photojournalism and the Assumption of Mechanical Objectivity in Journalism. *Digital Journalism*, *7*(8), 1117–1133.

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

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, *99*(1), 431–457.

Christin, A. (2016). From daguerreotypes to algorithms: machines, expertise, and three forms of objectivity. *ACM Computers & Society*, *46*(1), 27–32.

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, *7*(3), 613–628.

Cotter, K., & Reisdorf, B. C. (2020). Algorithmic Knowledge Gaps: A New Horizon of (Digital) Inequality. *International Journal of Communication*, *14*, 745–765.

Crocker, J., Major, B., & Steele, C. M. (1998). Social Stigma. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 504–553). Boston; New York: McGraw-Hill.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Accessed 21 August 2021

Daston, L., & Galison, P. (1992). The image of objectivity. *Representations*, *40*, 81–128.

Demetis, D. S., & Lee, A. S. (2018). When Humans Using the IT Artifact Becomes IT Using the Human Artifact. *Journal of the Association for Information Systems*, *19*(10), 929–952.

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.

Dovidio, J. F., Major, B., & Crocker, J. (2000). Stigma: Introduction and overview. In T. F. Heatherton, R. E. Kleck, M. R. Hebl, & J. G. Hull (Eds.), *The social psychology of stigma* (pp. 1–28). Guilford Press.

Espino-Pérez, K., Major, B., & Malta, B. (2018). Was it race or merit?: The cognitive costs of observing the attributionally ambiguous hiring of a racial minority. *Cultural Diversity and Ethnic Minority Psychology*, *24*(2), 272.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Gučanin, J. (2018). Sexistischer AMS-Algorithmus benachteiligt Frauen und Mütter. https://wienerin.at/sexistischer-ams-algorithmus-benachteiligt-frauen-und-mutter. Accessed 21 August 2021

Gunaratne, J., Zalmanson, L., & Nov, O. (2018). The Persuasive Power of Algorithmic and Crowdsourced Advice. *Journal of Management Information Systems*, *35*(4), 1092–1120.

Gunton, R. M., Stafleu, M. D., & Reiss, M. J. (2021). A General Theory of Objectivity: Contributions from the Reformational Philosophy Tradition. *Foundations of Science*, 1–15.

Hatzenbuehler, M. L. (2016). Structural stigma: Research evidence and implications for psychological science. *American Psychologist*, *71*(8), 742.

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (2nd ed.). New York: Guilford Press.

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, *11*(5), 365–371.

Johnson, S. E., Richeson, J. A., & Finkel, E. J. (2011). Middle class and marginal? Socioeconomic status, stigma, and self-regulation at an elite university. *Journal of Personality and Social Psychology*, *100*(5), 838–852.

Kang, J. C. (2016). The Unbearable Whiteness of Baseball. https://www.nytimes.com/2016/04/10/magazine/the-unbearable-whiteness-of-baseball.html. Accessed 21 August 2021

Khalil, O. E. M. (1993). Artificial decision-making and artificial ethics: A management concern. *Journal of Business Ethics*, *12*(4), 313–321.

Kizilcec, R. F. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY: ACM.

Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science*, *65*(7), 2966–2981.

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, *27*(3), 217–2348.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics*, *160*(2), 377–392.

Lepri, B., Staiano, J., Sangokoya, D., Letouzé, E., & Oliver, N. (2017). The tyranny of data? the bright and dark sides of data-driven decision-making for social good. In *Transparent data mining for big and small data* (pp. 3–24). Springer.

Link, B. G., & Phelan, J. C. (2001). Conceptualizing Stigma. *Annual Review of Sociology*, *27*, 363–385.

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

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

Major, B., & O'Brien, L. T. (2005). The social psychology of stigma. *Annual Review of Psychology*, *56*, 393–421.

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics*, *160*(4), 835–850.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479.

Munoko, I., Brown-Liburd, H. L., & Vasarhelyi, M. (2020). The Ethical Implications of Using Artificial Intelligence in Auditing. *Journal of Business Ethics*, *167*(2), 209–234.

Nelson, L., Simester, D., & Sudhir, K. (2020). Introduction to the Special Issue on Marketing Science and Field Experiments. *Marketing Science*, *39*(6), 1033–1038.

Nightengale, B. (2016). Adam Jones on MLB's lack of Kaepernick protest: 'Baseball is a white man's sport'. https://eu.usatoday.com/story/sports/mlb/columnist/bob-nightengale/2016/09/12/adam-jones-orioles-colin-kaepernick-white-mans-sport/90260326/. Accessed 21 August 2021

Parsons, C. A., Sulaeman, J., Yates, M. C., & Hamermesh, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *American Economic Review*, *101*(4), 1410–1435.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031.

Pescosolido, B. A., & Martin, J. K. (2015). The Stigma Complex. *Annual Review of Sociology*, *41*(1), 87–116.

Pinel, E. C. (2004). You're Just Saying That Because I'm a Woman: Stigma Consciousness and Attributions to Discrimination. *Self and Identity*, *3*(1), 39–51.

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

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141.

Reiss, J., & Sprenger, J. (2020). Scientific Objectivity. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2020.). https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity/; Metaphysics Research Lab, Stanford University.

Reiter, A. (2019). Das Amt und meine Daten. https://www.zeit.de/2019/20/digitale-verwaltung-behoerden-aemter-effizienzsteigerung-probleme. Accessed 21 August 2021

Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal*, *112*(483), F480–F518.

Roloff, J., & Zyphur, M. J. (2019). Null findings, replications and preregistered studies in business ethics research. *Journal of Business Ethics*, *160*(3), 609–619.

RStudio Team. (2021). *RStudio: Integrated development environment for r*. Boston, MA: RStudio, PBC. http://www.rstudio.com/

Shellenbarger, S. (2019). A Crucial Step for Averting AI Disasters. https://www.wsj.com/articles/a-crucial-step-for-avoiding-ai-disasters-11550069865. Accessed 21 August 2021

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology*, *34*, 379–440.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology* (pp. xiii–369). Cambridge; New York: Cambridge University Press.

Taylor, J. (2019). People should be held accountable for AI and algorithm errors, rights commissioner says. https://www.theguardian.com/law/2019/dec/17/people-should-be-held-accountable-for-ai-and-algorithm-errors-rights-commissioner-says. Accessed 21 August 2021

Uhlmann, E. L., & Cohen, G. L. (2007). "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, *104*(2), 207–223.

Vivian, J. E., & Berkowitz, N. H. (1992). Anticipated bias from an outgroup: An attributional analysis. *European Journal of Social Psychology*, *22*(4), 415–424.

Walton, G. M., Murphy, M. C., & Ryan, A. M. (2015). Stereotype Threat in Organizations: Implications for Equity and Performance. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 523–550.

Wimmer, B. (2018). Computer sagt nein: Algorithmus gibt Frauen weniger Chancen beim AMS. https://futurezone.at/netzpolitik/computer-sagt-nein-algorithmus-gibt-frauen-weniger-chancen-beim-ams/400345297. Accessed 21 August 2021

Xie, Y. (2020). *Bookdown: Authoring books and technical documents with r markdown*. https://github.com/rstudio/bookdown

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

Appendix A. Robustness of Results

In this section, we repeat the main analyses from the paper using ordinary least squares (OLS) regressions. Additionally, we assess the robustness of our findings by including the demographic variables collected during each study. The results reported in Tables A1 and A2 are consistent with those reported in the paper.

Table A1: Regression Results of Studies 1 and 2

	Dependent variable: *Choice (1 = algorithm, 0 = human)*					
	Study 1			Study 2		
	(1) Full sample	(2) Only men	(3) Only women	(4) Full sample	(5) Only men	(6) Only women
Outgroup condition	0.224*	0.107	0.300**	0.149**	0.104†	0.210**
	(0.089)	(0.161)	(0.112)	(0.047)	(0.062)	(0.074)
Education	0.021	0.049	0.026	−0.001	0.006	−0.014
	(0.036)	(0.071)	(0.043)	(0.019)	(0.026)	(0.029)
Age	0.000	0.001	0.000	0.000	0.002	−0.002
	(0.004)	(0.009)	(0.006)	(0.002)	(0.003)	(0.003)
Gender	−0.185†			0.046		
	(0.102)			(0.049)		
Care for family/others	−0.026	0.150	−0.148			
	(0.098)	(0.185)	(0.115)			
Disability	−0.261*	−0.284	−0.180			
	(0.110)	(0.209)	(0.151)			
Place	0.024	−0.004	0.038			
	(0.036)	(0.060)	(0.044)			
Unemployment duration	0.001	0.000	0.002†			
	(0.001)	(0.001)	(0.001)			
Constant	0.666*	0.450	0.183	0.177	0.140	0.381†
	(0.313)	(0.461)	(0.378)	(0.142)	(0.169)	(0.193)
Observations	118	42	76	390	226	164
R-squared	0.125	0.120	0.171	0.028	0.015	0.052

Note:
Although Studies 1 and 2 had binary outcome variables, we used OLS regressions because they are easier to interpret. The results, however, are robust to using logistic regressions. Robust standard errors are in parentheses.
** $p < 0.01$, * $p < 0.05$, † $p < 0.10$.

Table A2: Regression Results of Study 3

	Dependent variable: *Relative algorithmic objectivity*			Dependent variable: *Relative preference for algorithm*					
	(1) Full sample	(2) Only men	(3) Only women	(4) Full sample	(5) Only men	(6) Only women	(7) Full sample	(8) Only men	(9) Only women
Outgroup condition	0.778**	0.695†	0.865†	0.422*	0.347	0.504*	0.171	0.127	0.220
	(0.300)	(0.388)	(0.466)	(0.167)	(0.222)	(0.253)	(0.137)	(0.184)	(0.208)
Age	−0.001	0.001	−0.002	0.005	0.005	0.006	0.005	0.005	0.006
	(0.012)	(0.016)	(0.017)	(0.007)	(0.010)	(0.010)	(0.006)	(0.008)	(0.008)
Gender	0.154			0.405*			0.355*		
	(0.301)			(0.169)			(0.139)		
Relative algorithmic objectivity							0.323***	0.316***	0.328***
							(0.017)	(0.027)	(0.023)
Constant	−0.499	−0.371	−0.184	2.027***	2.476***	2.779***	2.188***	2.593***	2.839***
	(0.637)	(0.644)	(0.777)	(0.353)	(0.409)	(0.432)	(0.296)	(0.325)	(0.385)
Observations	599	311	288	599	311	288	599	311	288
R-squared	0.012	0.010	0.012	0.022	0.009	0.015	0.350	0.313	0.370

Note:
The table shows the regression results of the mediation analysis. Columns (1)-(3) report the effect of the outgroup-vs.-ingroup condition on relative algorithmic objectivity (i.e., the mediator). Columns (4)-(6) report the baseline effect of the outgroup-vs.-ingroup condition on relative preference for the algorithm. In Columns (7)-(9), relative algorithmic objectivity was added to the baseline. Robust standard errors are in parentheses. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$.

Appendix B. Selected Materials Used

In this section, we list some of the materials used to facilitate the comprehension of the paper. All survey materials are available from https://osf.io/7yek2/.

Table B1: Profile of Fictional Participant in Study 1

Question	Answer
Gender	Male
Age	28
Hispanic	No
Racial/Ethnic group	White
Education	4-Year College Degree
Care for family/others	No
Place	A very big city (> 1 million people)
Disability	No
Prior industry	Software
Unemployment duration	6 months
Job search	4 months

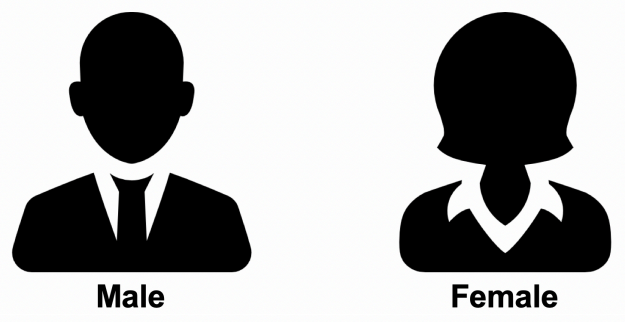

Figure B1: Pictograms Used in Study 2 (Source: Icons Made by Freepik from www.flaticon.com)