In press at *Journal of Experimental Psychology: General*

Algorithmic Discrimination Causes Less Moral Outrage than Human Discrimination

Yochanan E. Bigman[1,2], Desman Wilson[3], Mads N. Arnestad[4], Adam Waytz[3] and Kurt Gray[2]

[1] Yale University

[2] University of North Carolina at Chapel Hill

[3] Kellogg School of Management, Northwestern University

[4] BI Norwegian Business School

Author's note:

This work has been supported by the National Science Foundation (award #SMA 1714298) to Y.E.B., and a grant from the Charles Koch Foundation via the Center for the Science of Moral Understanding to K.G. The idea for this paper was partially inspired by the book "Feet of Clay" by Terry Pratchett.

A previous version of this work was presented at SPSP 2019 and appeared as a pre-print on PsyArXiv.

Context Paragraph

Considering the ever-increasing cases of algorithm-driven discrimination, it is crucial to understand its psychological consequences. We synthesize work on the role of perceived motivation in moral judgment (e.g., Bigman & Tamir, 2016; Carlson & Zaki, 2018; Reeder et al., 2002) with work on machine morality (e.g., Bigman & Gray, 2018) to predict an *algorithmic outrage deficit*—that discrimination at the "hands" of algorithms will incite less outrage because machines are seen to lack the motivation to be prejudiced. This work sheds light on the causes of moral outrage and reveals how people think about wrongdoing by artificial intelligence (AI).

Abstract

Companies and governments are using algorithms to improve decision-making for hiring, medical treatments, and parole. The use of algorithms holds promise for overcoming human biases in decision-making, but they frequently make decisions that discriminate. Media coverage suggests that people are morally outraged by algorithmic discrimination, but here we examine whether people are *less* outraged by algorithmic discrimination than by human discrimination. Eight studies test this *algorithmic outrage deficit* hypothesis in the context of gender discrimination in hiring practices across diverse participant groups (online samples, a quasi-representative sample, and a sample of tech workers). We find that people are less morally outraged by algorithmic (vs. human) discrimination and are less likely to hold the organization responsible. The algorithmic outrage deficit is driven by the reduced attribution of prejudicial motivation to algorithms. Just as algorithms dampen outrage, they also dampen praise—companies enjoy less of a reputational boost when their algorithms (vs. employees) reduce gender inequality. Our studies also reveal a downstream consequence of algorithmic outrage deficit—people are less likely to find the company legally liable when the discrimination was caused by an algorithm (vs. a human). We discuss the theoretical and practical implications of these results, including the potential weakening of collective action to address systemic discrimination.

Abstract word count: 207

Keywords: moral outrage; motivation attribution; human-robot interaction; discrimination

Discrimination often incites moral outrage (Batson et al., 2007; Russell & Giner-Sorolla, 2011; Spring et al., 2018; Sunstein et al., 1998; Tetlock, 2002), especially when perpetrated by companies (Halzack, 2019)—such as when Walmart was accused of systematically underpaying women and overlooking them for promotion (Covert, 2019). Ultimately, organizational discrimination is perpetrated by other people—those who control hiring, firing, and promotions. However, the rise of artificial intelligence raises a new possibility: organizational discrimination can be perpetrated by *algorithms*. For example, Amazon developed a machine-learning-based algorithm to screen applicant resumes, but the algorithm turned out to discriminate against women, penalizing applicants whose resumes contained terms such as "women's chess club captain" or the names of women's colleges. The algorithm was soon scrapped amid outrage about its systematic gender discrimination, and, according to Amazon, was never actually used (Dastin, 2018).

Although the Walmart and the Amazon cases both involved systematic discrimination that elicited outrage, Amazon's algorithm-driven discrimination seemingly elicited less moral outrage than Walmart's human-driven discrimination. Here we explore whether people truly are more blasé when witnessing discrimination at the "hands" of an algorithm (vs. a human)—what we call an *algorithmic outrage deficit*. We also explore one potential reason for this asymmetry in outrage – because algorithms are perceived as lacking mind (compared to humans; Bigman & Gray, 2018; Srinivasan & Sarial-Abi, 2021) people may perceive the discriminatory decisions of algorithms as less motivated by prejudice—and therefore be less outraged at that discrimination. We also examine a downstream consequence of the algorithmic outrage deficit—that people are less likely to find the company legally liable when it discriminated via algorithm.

The Rise of Artificial Intelligence

The past decades have seen a rapid increase in the integration of autonomous machines and algorithms into human society. Increasingly, tasks that used to be performed by humans are performed by autonomous machines and algorithms, including assembly line work (Levitin et al., 2006), customer service (Bares et al., 2007), house cleaning (Takeshita et al., 2006) and supermarket checkout (Aquilina & Saliba, 2019). Machine-learning has facilitated much of this development, providing data-based predictions that often outperform humans across many areas, from predicting the spread of disease (Chen et al., 2017) to the dynamics of crime (Shapiro, 2017). Algorithms have particularly revolutionized many practices in the business world, helping to manage inventory (Cárdenas-Barrón et al., 2012), distribution chains (Validi et al., 2015), and staff scheduling (Cai & Li, 2000).

The predictive abilities of AI systems are undeniable, but many are uncomfortable with humanity's growing reliance on algorithms. One concern is whether the increased use of machines will take jobs away from humans—similar to other technological revolutions—causing widespread economic unrest (Ford, 2015). Although increased automation appears to increase inequality, which is a driver of social unrest, the rise of AI may also bring people together by emphasizing their shared humanity (Jackson et al., 2020). Another concern about the rise of algorithms is distrust in machines' decision-making capacities. Many legal, medical, and military decisions involve life or death outcomes, and people seem averse to machines making these weighty decisions, in part because machines lack the ability to feel emotions (Bigman et al., 2019; Bigman & Gray, 2018; Gogoll & Uhl, 2018; Kramer et al., 2018; Young & Monroe, 2019). Unlike people who care deeply about their family and friends, machines appear devoid of compassion, and people are hesitant to allow algorithms to make decisions about human lives because of their dispassionate impartiality.

Some may see the impartiality of machines as a disadvantage—at least in some situations (Longoni et al., 2019). Others argue that this impartiality holds the promise to free decision-making from human biases (Mullainathan, 2019). Basic propensities for prejudice can lead humans to discrimination across many domains, including hiring processes. Substantial research reveals that people are biased in their decisions about who to interview, hire, and promote. For example, in one large-scale study, researchers sent out equivalent resumes to employers that differed only by the name at the top— Greg Baker versus Jamal Jones– and employers responded to the white name 50% more than the black name (Bertrand & Mullainathan, 2004). In the face of this type of discrimination, companies are increasingly relying on algorithms to make hiring practices unbiased (Heilweil, 2019).

Unfortunately, the promise of unbiased decision-making is currently unfulfilled, as AIs often discriminate. We previously reviewed one high profile example of algorithm discrimination—the Amazon hiring algorithm—but there are many others, such as racial discrimination by algorithms used to make parole decisions (Angwin et al., 2016) and healthcare decisions (Obermeyer et al., 2019), and gender discrimination by algorithms used to assign credit scores (Stankiewicz, 2019) and to display career ads (Lambrecht & Tucker, 2019). As these algorithms are usually intellectual properties of the companies and are not shared with the public, it is hard to know exactly what caused the discrimination, but there are several possibilities. One possibility is intentional programming, such as when a programmer designs the algorithm to give women a lower score than men. A second possibility is that the algorithm is trained to mimic existing human decisions, thereby showing the same bias as human decision-makers (Cheong et al., 2020). A third possibility is that bias creeps in when the algorithm is trained with existing data—and existing data typically reflects the outcomes of a biased decision process. For

example, a hiring algorithm may penalize women applicants because they show higher turnover—but this higher turnover may be driven by a sexist work environment or lower rates of promotions for women.

Regardless of the ultimate cause of algorithm discrimination, it is important to understand how people respond to cases in which algorithms perpetrate racial and gender discrimination. To answer this question, we need to first understand how people respond to discrimination in general.

Moral Outrage

Although discrimination is often widespread and institutionalized (Fornili, 2018; Goel, 2018; Manuel et al., 2017), salient cases of discrimination are typically seen as unfair. Human reactions to unfairness have deep roots in our evolutionary history and elicit moral outrage (Batson et al., 2007; Russell & Giner-Sorolla, 2011; Spring et al., 2018; Sunstein et al., 1998; Tetlock, 2002) [1]. Moral outrage can serve several important social functions. For example, moral outrage mobilizes people to punish unfair behavior (Fiske & Tetlock, 1997; Gummerum et al., 2016; Nelissen & Zeelenberg, 2009; Salerno & Peter-Hagene, 2013; Spring et al., 2018) which deters uncooperative behaviors (Kurzban et al., 2007; Xiao & Houser, 2005). In addition, moral outrage can promote collective action (Martin et al., 1984; Miller et al., 2011). Indeed, some argue that moral outrage evolved in human society because it served the adaptive function of enforcing cooperation within groups (Spring et al., 2018). When companies discriminate, in addition to eliciting moral outrage, it demotivates employees (Hausknecht et al., 2004), increases

[1] We note that there is an ongoing debate on whether moral outrage is construct which is independent of other types of anger (e.g., Batson et al., 2009, 2007; Hechler & Kessler, 2018). Our focus in this paper is the emotional response to discrimination as an outcome measure, rather than the construct validity of moral outrage.

turnover (Uggerslev et al., 2012), and even causes the public to boycott the discriminating company (Lindenmeier et al., 2012).

As with other aspects of human morality (Machery & Mallon, 2010), responses of moral outrage evolved in social groups with other humans, which raises the question of whether the actions of nonhuman agents—like algorithms—would also generate outrage. One possibility is that discrimination by an algorithm could generate more outrage than discrimination perpetrated by a human. Algorithms are usually implemented to manage or transform entire large-scale operations, which means they can impact a very large number of people. Whereas a single hiring administrator could discriminate against potentially hundreds of applicants, an algorithm—with its limitless throughput—has the capacity to discriminate against thousands of applicants. As people weigh the impact (or potential impact) in their moral judgments (Batson et al., 2007, 2009; Haidt, 2003; Sunstein et al., 1998; Tetlock, 2002), the scalability of an algorithm could elicit substantial outrage.

The novelty of algorithms making hiring decisions may also elicit considerable moral outrage. Moral judgments are typically weaker for descriptively normative acts—acts that are frequent or typical (Gawronski et al., 2017; Malle et al., 2014; Monroe & Malle, 2017). For example, if everyone evades taxes, tax evasion seems less morally wrong. The reason for this link between descriptive normativity and moral judgment is because people conflate descriptive norms (what is) with injunctive norms (what should be) (Eriksson et al., 2015). Acts that are less typical and less frequent, might therefore be more likely to evoke moral outrage. Given that algorithms are infrequently used to make hiring decisions (at least currently), people might be especially outraged when companies use them to perpetrate discrimination.

Although these factors suggest algorithms might generate more moral outrage than the actions of humans, here we suggest that algorithms might actually generate less moral outrage for similar discrimination. We label this prediction the *algorithmic outrage deficit*, and suggest it stems from people's perceptions of the mental states of those perpetrating discrimination as guiding moral judgments. Synthesizing the work emphasizing the role of perceived intentions (Alicke, 2000; Cushman, 2008; Malle et al., 2014; Malle & Knobe, 1997; Pizarro & Tannenbaum, 2011) and perceived motivation (Bigman & Tamir, 2016; Carlson et al., 2022; Carlson & Zaki, 2018; Levine & Schweitzer, 2014; Reeder et al., 2002) with the work showing that people ascribe different mental states to algorithms and robots (Gray & Wegner, 2012; Li et al., 2016; Waytz et al., 2014; Young & Monroe, 2019), we propose that as people are less likely to attribute negative motivations (i.e., prejudice) to an algorithm, they would be less outraged when algorithms discriminate.

Motivation

When someone acts immorally, observers often ask "Why?"—why did the agent behave the way they did? These questions boil down to perceived motivation – what drove the person to act the way they did. Understanding the motivations underlying a person's behavior allows us to make sense of the social world, predict people's future behavior and guides the way we interact with them (Cosmides, 1989; Waytz, Morewedge, et al., 2010). Moral psychology now emphasizes the importance of person-centered moral judgment (i.e., character; Goodwin et al., 2014; Pizarro & Tannenbaum, 2011; Uhlmann et al., 2013, 2014, 2015), and nothing is more central to judging a person's morality than their motivations (Bigman & Tamir, 2016; Carlson et al., 2022; Levine & Schweitzer, 2014; Reeder et al., 2002). Consider the trolley problem, in which a person decides whether or not to actively kill one person to save five others (Foot,

1967). A person who decides to kill one person because they are motivated to save the five people will be judged more positively than a person who did the same action, but did the action because they wanted to push someone to their death (Kahane et al., 2018). We propose that the motivation attributed to a wrongdoer will affect moral outrage.

The motivation that people attribute to others might be especially important in reactions to hiring decisions because there are many possible reasons why one candidate might be preferred over another. Hiring involves weighing many different candidate features, such as education, experience, skills, and general "fit" (Bowen et al., 2011). When a white person or a male candidate is hired over a person of color or a female candidate, there is ambiguity surrounding whether that decision reflects prejudice (e.g., sexism, racism) or a perceived difference in qualifications. Indeed, the complexity of hiring decisions is one of the reasons why it is hard to prove hiring discrimination in the court of law (Kotkin, 2009). The attributional ambiguity of hiring decisions means that general cues about the existence (or lack) of prejudiced motivation could impact people's moral reactions. If a decision-maker seems unlikely—or unable—to harbor ill-will towards a social group, their biased decisions may elicit less outrage. For example, when a woman fails to hire a qualified woman, people may be less likely to perceive discrimination because they assume that other women are not prejudiced against women.

People may assume that algorithms are incapable of harboring prejudice. Research shows that people perceive the mind of artificial agents such as algorithms differently than the mind of humans (Bigman & Gray, 2018; Gray & Wegner, 2012; Malle et al., 2016; Weisman et al., 2017). They are seen as being less able to think rationally and plan their actions, and especially less able to experience emotions (Bigman & Gray, 2018; Gray et al., 2007; Gray & Wegner,

2012). The way people perceive algorithms affects how much people trust them (Gogoll & Uhl, 2018), blame them (Malle et al., 2016; Shank & DeSanti, 2018; Srinivasan & Sarial-Abi, 2021), and want them to make decisions (Bigman & Gray, 2018; Young & Monroe, 2019). Based on this research, we suggest that seeing less mind in machines may contribute to an algorithmic outrage deficit. Humans are perceived as having a full mind—capable of both intention and antipathy—and therefore when they discriminate, they are likely to be seen as motivated by prejudice. Because algorithms are perceived as lacking a full mind, people should be less likely to attribute their behavior to a prejudiced motivation, which should then make observers less outraged when algorithms perpetrate discrimination. We test the algorithmic outrage deficit in the studies described below.

Importantly, although we predict that people will not perceive algorithms *per se* as prejudiced, they might attribute such a prejudiced motivation to the people who created and trained the algorithm. We also test this possibility, examining whether participants perceive more prejudiced motivation in a discriminatory algorithm when it is programmed by a software company known for sexist work conditions. We also test the positive counterpart of algorithm discrimination—how would people respond when algorithms actually help reduce discrimination.

Current Research

We systematically investigate how much people are morally outraged by algorithmic (vs. human) discrimination through 8 studies with diverse methods and samples, including American online samples, a nationally representative sample of the UK, and a sample of employees at Norwegian technology firms. In Study 1 we tested whether people perceived algorithms as less motivated by prejudice than humans, and Study 2 tested whether this difference in motivation

perception leads people to be less outraged by algorithmic discrimination. Study 3 examined how algorithm discrimination affects moral outrage towards the company that used the algorithm. Study 4 investigated the positive counterpart of algorithm discrimination by examining people's judgments of a company that used an algorithm that reduced gender inequality. Study 5 sought to replicate our findings in a sample of professionals in the technology industry, which allowed us to test whether knowledge about AI moderated outrage toward algorithmic discrimination. Study 6 examined whether the algorithmic outrage deficit was moderated by the identity of the programmers (e.g., if they seemed prejudiced), and Study 7 examined the impact of algorithm anthropomorphism on moral outrage at algorithm discrimination. Finally, in Study 8 we tested another downstream consequence of the reduced attribution of prejudiced motivation to algorithms—judgments of legal liability, such that companies may be held less liable for algorithmic vs human discrimination.

Across these eight studies, we predicted that in cases of discrimination, people will attribute less prejudiced motivation to an algorithm (vs. a human) and that this reduced attribution will lead to less moral outrage. All studies were approved by the IRB of The University of North Carolina at Chapel Hill. We pre-registered all studies except for Study 5. Full study materials and data can be found at https://osf.io/87yu5/?view_only=36fc6f1a3e004665a1e916be5fd180db. For all studies, we report all measures, conditions, data exclusions, and how we determined our sample sizes, acknowledging that variation in sample size occurred for procedural and theoretical reasons.

Study 1: Perceived Prejudiced Motivation

In Study 1 we tested whether people are less likely to attribute prejudiced motivation to an algorithm vs. a human. We also tested whether people see algorithms as more objective than

humans, and whether people see the same action as less discriminatory when it was perpetrated by an algorithm. Participants read about a human HR specialist or an algorithm that discriminated against women in hiring decisions, based on the real story of the algorithm that Amazon used for candidate selection (Dastin, 2018). Following the Amazon case, we described the discriminator as involved in the initial stage of the hiring process—scanning resumes and giving each a rating of between one (lowest) and 5 (highest) stars. We then asked participants to rate the prejudiced motivation of the agent, and how objective and discriminatory the decision-making was. We predicted that participants would attribute less prejudiced motivation to the algorithm, and perceive its actions as both less discriminatory and more objective than identical actions performed by a human.

Method

Participants.

Two hundred and forty participants (116 male, 120 female, 4 other or declined to respond; age: $M = 32.92$, $SD = 10.83$) from the United States completed the study on Prolific in exchange for 50 cents. Accounting for participants who might fail attention checks and therefore be excluded from the analysis, this sample size gives us a power of 0.95 to detect a two-tailed medium effect size (Cohen's $d = 0.5$, calculated with G*Power 3.1.9.2). As specified in the pre-registration (https://aspredicted.org/e2dq8.pdf), we did not include in the analysis participants who failed to answer either the attention check or comprehension question correctly, leading to the exclusion of 10 participants.

Procedure.

Participants were randomly assigned to an Algorithm or a human condition. All participants read the following:

> AeonTech is a high-tech company that creates software to provide interconnectivity between cloud-based enterprise systems.
>
> An external audit found that despite receiving large numbers of applications from women for software engineering positions, AeonTech hires almost no women.
>
> An external audit found the reason for AeonTech's gender discrimination. AeonTech has a two-phased process for hiring software developers. The second stage involves a standard review by a committee of executives, but this committee only receives applications that have been passed forward at the first stage.

Participants in the Algorithm condition read the following (Human condition in parentheses):

> At the first stage, an AeonTech algorithm—SigmaEvalu8, an unsupervised self-learning AI system (an AeonTech HR specialist) scans each application and gives it a rating between one star (lowest fit) and five stars (highest fit). Applicants with the highest ratings are then forwarded to the hiring committee. This self-learning algorithm (HR specialist) systematically gave women a lower star rating than men.

Participants then rated the following dependent variables on a 0 (strongly disagree) to 100 (strongly agree) slider scale.

Assessing perceived discrimination. We assessed perceived discrimination with three items: “SigmaEvalu8 (The HR specialist) discriminated against women”, “SigmaEvalu8 (The HR specialist) treated people differently according to their gender” and “SigmaEvalu8 (The HR specialist) treated men and women differently”. We created a perceived discrimination index by averaging all three items, Cronbach’s $\alpha = .96$.

Assessing perceptions of objectivity. We assessed perceptions of objectivity with three items: "SigmaEvalu8 (The HR specialist) is data-driven", "SigmaEvalu8 (The HR specialist) relies on facts", and "SigmaEvalu8 (The HR specialist) is unaffected by personal opinions". We created a perceived data-driven behavior index by averaging all three items, Cronbach's $\alpha = .87$.

Assessing perceived prejudiced motivation. We then assessed perceived prejudiced motivation with four items: SigmaEvalu8 (The HR specialist) is sexist", "SigmaEvalu8 (The HR specialist) does not want to hire women for high-skilled jobs", "SigmaEvalu8 (The HR specialist) dislikes women", and "SigmaEvalu8 (The HR specialist) is prejudiced". We created a composite perceived prejudiced motivation index by averaging all four items, Cronbach's $\alpha = .91$.

As a comprehension question, we asked participants which of the following best describes the hiring practices at AeonTech: "A hiring committee decided who to hire", "An HR specialist gives each application an initial rating and the hiring committee decides who to hire from the top-rated applicants" or "An algorithm gives each application an initial rating and the hiring committee decides who to hire from the top-rated applicants". Finally, participants provided demographic information.

Results

An independent samples *t*-test revealed that participants perceived the algorithm ($M = 80.17$, $SD = 25.47$) as less discriminatory than the human ($M = 89.09$, $SD = 17.26$), $t(228) = 3.10$, $p = .002$, Cohen's $d = 0.41$. A second *t*-test revealed that participants perceived the algorithm ($M = 51.04$, $SD = 25.79$) as more objective than the human $M = 25.45$, $SD = 23.20$), $t(228) = 7.91$, $p < .001$, Cohen's $d = 1.04$. A third *t*-test revealed that participants perceived the algorithm ($M =$

54.04, *SD* = 29.83) as less motivated by prejudice than the human (*M* = 72.43, *SD* = 22.98), *t* (228)=5.23, *p* < .001, Cohen's *d* = 0.69. Results are shown in Figure 1.

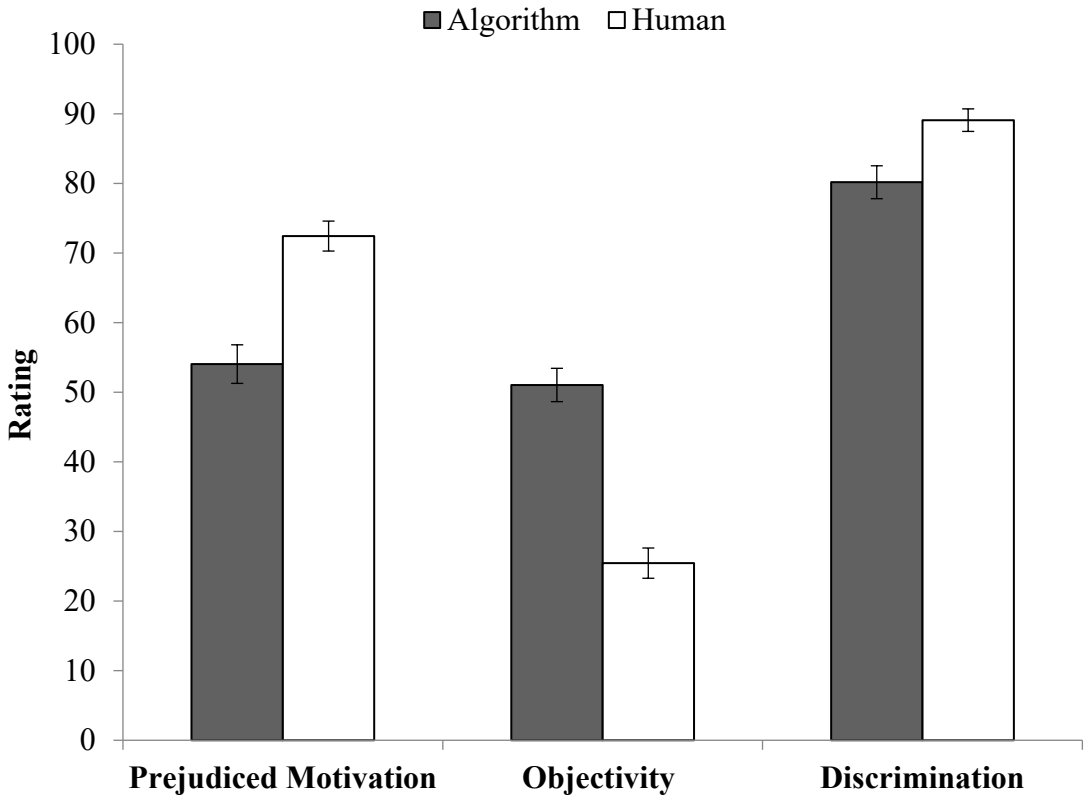

Figure 1. Ratings of discriminatory actions by an algorithm vs. a human (Study 1). Error bars reflect standard errors. All differences are statistically significant ($p < .05$).

Discussion and Replication

These results suggest that for the same action, people see algorithms as less motivated by prejudice than humans. In another study (see "Study 1: Replication" in the supplemental materials; final *N*=216) we replicated this finding with a different paradigm in which participants read about discrimination by an algorithm or a human and wrote what they thought the reason for the discrimination was. Two independent coders rated whether participants attributed the

discriminatory behavior to data or a prejudiced motivation. The results of this replication study were consistent with our findings in Study 1, participants attributed less prejudice to the human discriminator than the algorithm discriminator $t(214) = 5.38$, $p < .001$, Cohen's $d = 0.74$. Not only did people perceive algorithms as less motivated by prejudice than humans, Study 1 revealed that people see algorithms as more objective than humans, and see their actions as less discriminatory than the same actions performed by humans. Given that people perceive discrimination to the extent they perceive prejudice (Major et al., 2003) Study 1 supports the first part of our theory: people see algorithms as less motivated by prejudice than humans. In Study 2 we tested our full model by assessing perceived prejudice motivation directly.

Study 2: Algorithmic Outrage Deficit

In Study 2 we tested our algorithmic outrage deficit hypothesis, examining whether people are less outraged when an algorithm discriminates than when a human discriminates. We also tested whether this effect is mediated by people perceiving algorithms as less motivated by prejudice than humans. Participants read the same story as in Study 1 and reported their moral outrage and the prejudiced motivation they attributed to the agent. We predicted that we would find an algorithmic outrage deficit, such that people will be less outraged when the discrimination was done by an algorithm than when it was done by a human. We further predicted that perceived prejudiced motivation would mediate the algorithmic outrage deficit.

Method

Participants

Eight hundred and four participants[2] (300 male, 486 female, 18 other or preferred not to disclose; age: $M = 32.52$, $SD = 11.54$) from the United States completed the study on Prolific in exchange for 30 cents. Accounting for participants who might fail attention checks and therefore be excluded from the analysis, this sample size gives us a power of 0.80 to detect a two-tailed small effect size (Cohen's $d = 0.2$, calculated with G*Power 3.1.9.2). As specified in the pre-registration (https://aspredicted.org/7e7e3.pdf), we did not include in the analysis participants who failed the attention check, leading to the exclusion of thirty-four participants.

Procedure

All participants first read the following, a slightly modified version of the vignette we used in Study 1:

> AeonTech is a high-tech company that creates software to provide interconnectivity between cloud-based enterprise systems.
> An external audit found some discrimination in AeonTech's hiring practices. Despite receiving large numbers of applications from women for software engineering positions, AeonTech hires almost no women.
>
> An external audit found the reason for AeonTech's gender discrimination. AeonTech has a two-phased process for hiring software developers. The second stage involves a standard review by a committee of executives, but this committee only receives applications that have been passed forward at the first stage.

[2] We initially pre-registered, ran, a sample of 240 participants. The result for moral outrage was not significant in a two-tailed t-test (p =.098). We decided to increase our sample size to see whether this initial non-significant result likely reflects a type II error, or the possibility that the algorithm outrage deficit does not exist in this context. Therefore, to test our hypotheses with more statistical power, we ran additional participants, aiming for a total of 800. We pre-registered adding the participants (https://aspredicted.org/3m8xa.pdf). All of the results hold when Bonferroni correcting for multiple comparisons.

Participants were then randomly assigned to either a Human or an Algorithm condition. In the Algorithm condition participants read the following (Human condition in parentheses):

> At the first stage, an AeonTech algorithm – SigmanEvalu8, an unsupervised self-learning AI system (an AeonTech HR specialist) scans each application and gives it a rating between one star (lowest fit) and five stars (highest fit). Applicants with the highest ratings are then forwarded to the hiring committee. This self-learning algorithm (HR specialist) systematically gave women a lower star rating than men.

We first measured the prejudiced motivation participants perceived the human/algorithm had, using the same items as in Study 1, Cronbach's $\alpha = .92$. We then measured participants' moral outrage.

Measuring Moral Outrage. We used items that focused on the actions, rather than the agent, because participants might think that the algorithm itself is not a viable target of moral outrage. Due to the discussion about the role of disgust in moral outrage (Russell & Giner-Sorolla, 2011; Salerno & Peter-Hagene, 2013) we included one item that measures disgust, as well as items measuring anger and outrage. Specifically, we asked participants to rate their agreement on a 0 (strongly disagree) to 100 (strongly agree) slider, with the following three items (adapted from Russell & Giner-Sorolla, 2011): "I am angry at the HR specialist's/the algorithm's discriminatory actions", " I am outraged by the HR specialist's/the algorithm's discriminatory actions" and "I am disgusted by the HR specialist's/the algorithm's discriminatory actions". We created a moral outrage index by averaging these three items, Cronbach's $\alpha = .96$. To control for intention, we then measured perceived intentionality by

asking participants to rate their agreement with the following item: "The HR specialist/the algorithm intended not to hire women".

Finally, participants answered the same attention check as in Study 1 and provided demographic information.

Results

An independent samples *t*-test revealed that, as predicted, participants perceived the algorithm as being less motivated by prejudice ($M = 59.36$, $SD = 29.75$) than the human HR specialist ($M = 74.67$, $SD = 22.15$), $t(768) = 8.07$, $p < .001$, Cohen's $d = 0.58$, replicating our results from Study 1. A second independent samples *t*-test revealed that, as predicted, participants were less morally outraged by the algorithm's discriminatory actions ($M = 61.10$, $SD = 31.71$) vs. the human HR specialist's discriminatory actions ($M = 73.40$, $SD = 26.53$), $t(768) = 5.82$, $p < .001$, Cohen's $d = 0.42$, supporting our theory about algorithmic outrage deficit. A third independent samples *t*-test revealed that participants perceived the algorithm's discrimination as less intentional ($M = 54.13$, $SD = 33.94$) than the human HR specialist's discrimination ($M = 74.77$, $SD = 25.20$), $t(768) = 9.55$, $p < .001$, Cohen's $d = 0.69$.

Mediation

To test whether perceived prejudiced motivation mediated the effect of agent on moral outrage, we performed a bootstrapping mediation analysis (Preacher & Hayes, 2008; 5000 iterations, model 4) with agent as the IV, coding the algorithm condition as 1 and the human condition as -1, moral outrage as a DV, and perceived prejudiced motivation as a mediator.

As predicted, the effect of agent on moral outrage, $b = -6.15$, $SE = 1.06$, $CI_{.95}$[-8.22, -4.07] was mediated by an indirect effect of perceived prejudiced motivation, $b = -6.55$, $SE =$

0.81, $CI_{.95}$[-8.13, -4.95], see Figure 2. When accounting for the mediation by perceived prejudiced motivation, the direct effect of agent on moral outrage was not significant, $b = 0.40$, $SE = 0.70$, $CI_{.95}$[-0.98, 1.78]. The results of the mediation analysis remain significant when controlling for intention[3]. The reduced moral outrage for algorithm discrimination appears to be driven by people attributing less of a prejudiced motivation to the algorithm (vs. the human).

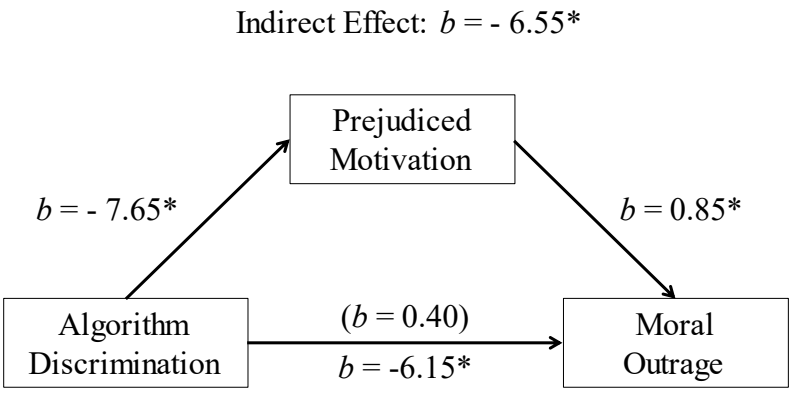

Figure 2. Mediation analysis reveals that perceived prejudiced motivation mediates the effect of agent on moral outrage (Study 2).

* Denotes $p < .05$.

Discussion

[3] An exploratory mediation analysis testing for motivation and intention as mediators revealed that both motivation, b=-5.29, $SE = 0.72$, $CI_{.95}$[-6.74, -3.91], and intention, b =-2.07, $SE = 0.47$, $CI_{.95}$[-3.04, -1.22], were significant mediators. Although motivation is a significant mediator even when accounting for intention, the results suggest that perceived intentions also might contribute to the algorithm outrage deficit.

Study 2 supported our model. The results show that people are less morally outraged by a discriminatory algorithm than a discriminatory human, establishing an algorithmic outrage deficit. We further found that the algorithmic outrage deficit is statistically mediated by perceiving less prejudicial motivation in algorithms, an effect that remains robust even when controlling for perceived intentionality.

Replications

We replicated these findings in five additional studies, reported in full in the supplemental materials. Four (three pre-registered) studies tested the effect of algorithm (vs. human) discrimination on moral outrage, and the fifth (pre-registered) study tested mediation by perceived prejudiced motivation (see "Study 2: Replications A-E" in the supplemental materials). We report the studies in the supplementary materials rather than the main text, because their methodology leaves open more ambiguities than the current Study 2. First, in these replication studies, the agent (the human HR specialist or the algorithm) actually makes the hiring decision rather than just screening the applicant (as in Study 2). Second, we measured moral outrage with a more general set of items, asking participants how morally outraged they were, how unjust they thought the actions were, how immoral they thought the actions were and how wrong they thought they were, all on a 1 to 7 scale. The results of these replication studies are presented in Figure 3.

We found support for the algorithmic outrage deficit in all studies. Replication 2A (N = 122) examined race discrimination, Cohen's d = 0.80. Replication 2B (N = 241) examined age discrimination, Cohen's d = 0.34. Replication 2C (N = 241) examined gender discrimination, Cohen's d = 0.46. Replication 2D (N = 1503) examined gender discrimination in a quasi-representative sample from the UK, Cohen's d = 0.26. Replication 2E (N = 240) examined race

discrimination, Cohen's $d = 0.39$, and also tested for mediation by perceived prejudiced motivation. The mediation analysis replicated the results of the mediation analysis in Study 2, and found that the reduced outrage at algorithm discrimination, $b = -0.23$, $SE = 0.08$, $p = .004$, was mediated by people perceiving the algorithm as less motivated by prejudice than the human, $b = -0.30$, $SE = 0.06$, $CI_{.95}[-0.42, -19]$.

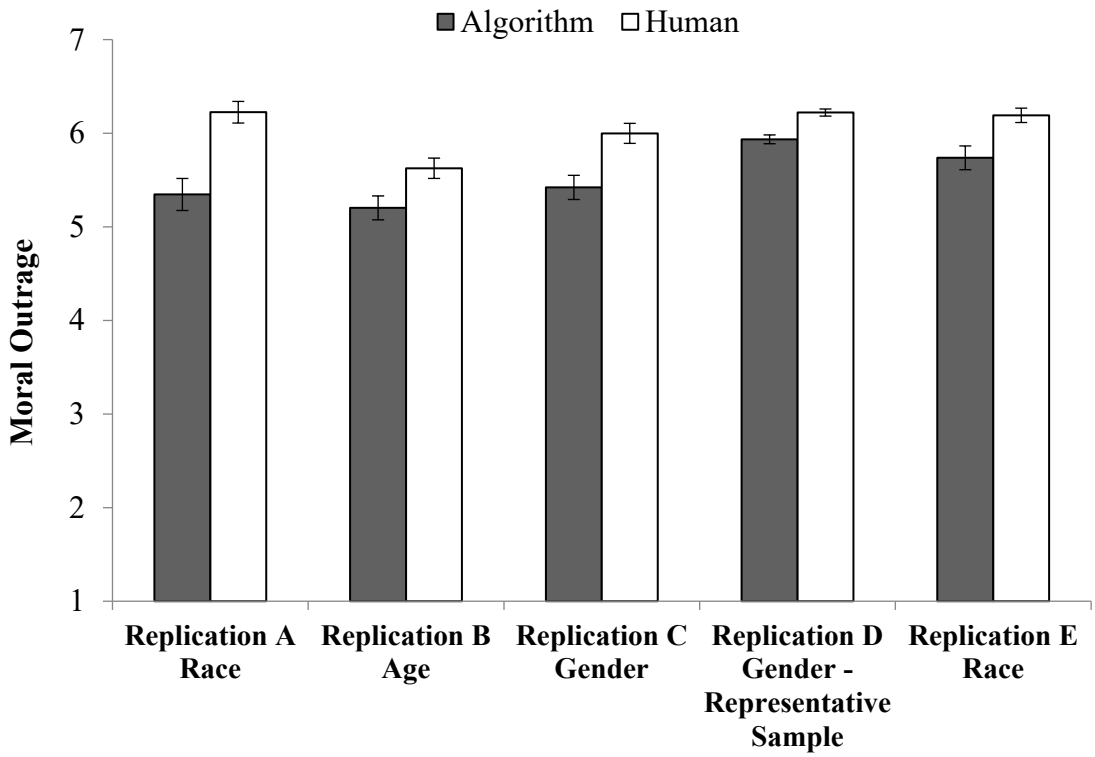

Figure 3. Mean moral outrage by agent (Study 2: Replications A-E). Error bars reflect standard errors. All differences are statistically significant ($p < .05$).

Study 3: Outrage at the Company

In Study 2 we examined how algorithmic discrimination affects outrage at the discrimination itself. However, beyond general outrage at the event, people might be outraged at the company that used the discriminatory algorithm (or human). In Study 3 we expanded our investigation and examined how algorithm discrimination affects judgment of the company that used the algorithm for decision-making. Focusing on the company required us to take into account that people might not want algorithms to make such decisions (Bigman & Gray, 2018) or view algorithm decision-making in HR decisions as generally unfair (Newman et al., 2020). Therefore, people might be more outraged initially at companies for using algorithms for hiring decisions, even if the increase in outrage at discrimination by algorithms is smaller than the increase in outrage at discrimination by humans. To account for this possibility, we measured moral outrage twice. The first time after telling participants that an algorithm or a human was used to screen resumes in the company, and the second time after telling them that the algorithm or the human discriminated against women. We predicted that people will be initially more outraged at the company that used an algorithm for screening resumes than the company that used a human HR specialist to screen resumes, but that the increase in outrage due to discrimination would be smaller for the company that used an algorithm. We measured how permissible people thought it was for the algorithm to make these decisions (Bigman & Gray, 2018) and perceptions of the algorithm's fairness (Newman et al., 2020) as well to test if these documented effects might affect people's moral outrage at the company for the use of the algorithm.

Procedure

Participants

Four hundred and eighty-two participants (222 male, 253 female, 7 other or declined to respond; age: $M = 31.26$, $SD = 10.44$) from the United States completed the study on Prolific in exchange for 60 cents. Accounting for participants who might fail attention checks and therefore be excluded from the analysis, this sample size gives us a power of 0.95 to detect a two-tailed small to medium effect size (Cohen's $d = 0.35$, calculated with G*Power 3.1.9.2). As specified in the pre-registration (https://aspredicted.org/fc2ni.pdf), we did not include in the analysis participants who failed to answer either the attention check or comprehension question correctly, leading to the exclusion of 24 participants.

Procedure

The procedure had four stages. First, participants read about the agent (algorithm/human) screening resumes at a company. Second, participants reported their moral outrage. Third, participants read about discrimination via the agent. Fourth, participants reported their moral outrage at the company again.

Agent introduction

Participants were randomly assigned to an Algorithm or a Human condition. In the Algorithm condition participants first read the following (Human condition in parentheses):

> AeonTech is a high-tech company that creates software to provide interconnectivity between cloud-based enterprise systems.
>
> AeonTech has a two-phased process for hiring software developers.
>
> In the first phase, an algorithm—SigmaEvalu8, an unsupervised Bayesian AI system (an HR specialist) scans each application. Each application gets between one star (lowest fit) and five stars (highest fit), according to the initial evaluation of the algorithm (HR specialist). Applicants with the highest scores are then forwarded to the hiring committee.

> At the second phase, a hiring committee of software engineers and senior executives goes over the applications that got the highest initial scores, conducts interviews and decides who to hire.

Participants then completed the following two measures (all items were answered on a 0 (strongly disagree) to 100 (strongly agree) slider scale).

Measurement of judgments pre-discrimination.

Outrage. We measured participants initial moral outrage at the company by asking them to rate their agreement with three items: "I am angry at AeonTech's hiring practices", "I am outraged by AeonTech's hiring practices" and "I am disgusted by AeonTech's hiring practices". We created an index of moral outrage at the company by averaging all three items (Cronbach's $\alpha = .93$).

Permissibility. We then measured how permissible they thought it was for the agent to screen resumes by asking them to rate their agreement with the following three items (adapted from Bigman & Gray, 2018): "It is appropriate for SigmaEvalu8 (the HR specialist) to screen resumes", "SigmaEvalu8 (The HR specialist) should be the one to screen resumes" and "SigmaEvalu8 (The HR specialist) should be forbidden from screening resumes" (reversed scored). We created an index of permissibility by averaging all three items (Cronbach's $\alpha = .82$).

Discrimination manipulation

After measuring pre-outcome moral outrage and permissibility, we told participants about the discriminatory outcome. Participants in the Algorithm condition read the following (Human condition in parentheses):

> Despite receiving large numbers of applications from women for software engineering positions, AeonTech has hired almost no women.
> The reason for this has to do with the initial screening. AeonTech's algorithm, SigmaEvlau8H, (HR specialist) systematically gave women a lower star rating than men. Therefore, women almost never made it to the second phase, and were not in the final list that the hiring committee considered.

Measurement of judgments post-discrimination

Outrage, perceived motivation. After reading about the discrimination, we measured participants' moral outrage at the company again, using the same items as they did in the first measurement (Cronbach's $\alpha = .97$), and perceived prejudiced motivation, using the same items as in Studies 1-2 (Cronbach's $\alpha = .93$).

Fairness. We then measured how fair participants thought the company's hiring process was, by asking participants to rate their agreement with the following four items (adapted from Newman et al., 2020): "The way AeonTech decides who to hire seems fair", "AeonTech's process for deciding who to hire was fair", "The decision who to hire was fair" and "The outcome of the hiring process at AeonTech was fair" (Cronbach's $\alpha = .94$).

As an exploratory measure, we then measured participants' beliefs about men being better than women in math with the following item: "Statistically, men are better at math than women". Finally, participants completed the same attention check as in Studies 1-2 and provided demographic information.

Results

Motivation, permissibility, fairness and math ability

An independent samples *t*-test revealed that participants found it less permissible for the algorithm (M = 57.38, SD = 22.81) to screen resumes than the HR specialist (M = 73.37, SD = 20.32), t (456)=7.91, $p < .001$, Cohen's d = 0.74, replicating previous work showing that people are averse to algorithms making certain high-stake decisions (Bigman & Gray, 2018). A second independent samples *t*-test revealed that participants thought the algorithm was less motivated by prejudice (M = 61.74, SD = 29.44) than the human HR specialist (M = 68.95, SD = 28.89), t (456)= 2.65, p = .008, Cohen's d = 0.25, replicating our results from Studies 1-2. A third independent samples *t*-test did not find a difference between how fair participants thought the decisions were when an algorithm made them (M = 17.89, SD = 20.89) than when a human did (M = 20.32, SD = 22.98), t (452)[4] = 1.52, p =.129, failing to replicate previous showing that people find hiring decision by algorithms less fair than hiring decisions by humans (Newman et al., 2020). This result is likely because we measured fairness *after* participants learned about the discriminatory outcome while previous work examined perceptions of fairness independently of discrimination. A fourth independent samples *t*-test did not find a difference in participants agreement that statistically men might be better than women at math when an algorithm discriminated against women (M = 19.81, SD = 25.93) than when a HR specialist discriminated against women (M = 19.67, SD = 24.36), t (443) = 0.057, p =.955.

Moral Outrage

If discrimination by an algorithm causes less moral outrage at the company than discrimination by a human, we would expect a smaller increase in moral outrage following algorithm (vs. human) discrimination. We tested this in a 2 (agent: human, algorithm) x 2 (time:

[4] The degrees of freedom vary across these tests because some participants did not answer all of the questions.

pre-discrimination, post-discrimination) mixed-model ANOVA predicting moral outrage at the company. We found a main effect for time, such that participants were more morally outraged after hearing about the discrimination (M = 62.35, SD = 21.78) than before hearing about the discrimination (M = 17.58, SD = 21.78), $F(1, 456) = 827.36$, $p < .001$, partial $\eta^2 = .645$, suggesting that across conditions, discrimination increased moral outrage. We also found a main effect for agent, such that participants were more outraged at the company when it used an algorithm to screen applicants (M = 44.03, SD = 20.96) than when it used an HR specialist to screen applicants (M = 35.79, SD = 21.29), $F(1, 456) = 17.46$, $p < .001$, partial $\eta^2 = .037$. However, this effect was qualified by a significant time x agent interaction, $F(1, 456) = 6.27$, $p = .013$, partial $\eta^2 = .014$. Follow-up pair-wise comparisons revealed that while pre-discrimination participants were more outraged at the company when it used algorithms for screening applicants (M = 23.58, SD = 23.50) than when it used an HR specialist (M = 11.42, SD = 17.93), $F(1, 456) = 38.56$, $p < .001$, partial $\eta^2 = .078$, post-discrimination the difference between algorithm discrimination (M = 64.50, SD = 29.41) and human discrimination (M = 60.15, SD = 34.03) was not significant, $F(1, 456) = 2.14$, $p = .144$. These results suggest that while participants were initially more outraged at the company for using an algorithm, the increase in moral outrage at the company following discrimination by an algorithm was smaller than the increase following discrimination by a human. We note that post-discrimination, the difference between outrage at the company that used an algorithm and the company that used a human is not significant, suggesting a possible boundary for algorithm outrage deficit. Anchoring people on their pre-existing outrage toward the use of algorithms in hiring might have produced outrage equivalent to human-based discrimination after discrimination has occurred.

Mediation by perceived prejudiced motivation

We conducted a mediation analysis to test whether perceived prejudice mediated the reduced increase in outrage following discrimination by an algorithm (vs. human). We used the increase in moral outrage (calculated as outrage post-discrimination minus outrage pre-discrimination) as our DV, agent (Human = -1, Algorithm = 1) as our IV, and perceived prejudiced motivation as the mediator. The analysis (Preacher & Hayes, 2008; 5000 iterations, model 4), revealed that, as predicted, the effect of agent on moral outrage, b = -3.90, SE = 1.56, p = .013, was mediated by an indirect effect of attribution of prejudiced motivation, b = -2.46, SE = 0.95, $CI_{.95}$[-4.30, -0.55]. When accounting for the mediation by attribution of prejudiced motivation, the direct effect of agent on moral outrage was not significant, b = - 1.44, SE = 1.26, $CI_{.95}$[-3.92, 1.03]. The smaller increase in moral outrage for discrimination by an algorithm appears, therefore, to be driven by people attributing less of a prejudiced motivation to the algorithm (vs. the human), consistent with our findings from Study 2.

Discussion

The results of Study 3 paint a nuanced picture of the effect of algorithm discrimination on moral outrage towards the company that used the algorithm. Participants were initially more morally outraged at the company for using algorithms for such decisions. However, consistent with our theory and our findings from Study 2, the increase in moral outrage following discrimination was smaller for discrimination by an algorithm than for discrimination by a human, albeit the difference in outrage at the company post-discrimination was not significant. Consistent with our previous findings, the reduced increase in moral outrage following algorithm (vs. human) discrimination was mediated by people attributing less prejudiced motivation to a discriminatory algorithm than a discriminatory human. One limitation of Study 3 is that although we measured outrage at the company using the same items (e.g., “I am outraged by AeonTech’s

hiring practices"), these items might have a different meaning in different contexts. Before reading about the discrimination, participants may have focused on the fact that AeonTech was using an algorithm for hiring decisions (and not general outrage at the company). After reading about the discrimination, participants maybe focused on the fact that AeonTech was discriminatory. Accordingly, there is some ambiguity about the exact meaning of the differences in outrage scores between pre- and post-discrimination.

In Studies 1-3 we focused on people's responses to algorithms that discriminate against women, a group that has faced discrimination within the technology sector. In Study 4 we examine people's reactions to algorithms that favor women.

Study 4: Positive and Negative Outcomes

In Study 4 we expanded our investigation to cases in which an algorithm (vs. a human) behaves in a way that people evaluate positively. According to our theory, if perceived motivation affects reactions to algorithm behavior, then reactions to an algorithm that acts positively (i.e., increases rather than decreasing gender equality) will not be judged as positively as a person—who presumably acts upon a motivation to do good. In other words, not only is there an algorithmic outrage deficit but also an algorithm praise deficit.

To test this idea, participants were randomly assigned to read about companies that either discriminated against women or had gender equality. Each participant read about two companies. In one company the source for this was a human HR specialist who screened resumes, and in the other company the source was an algorithm that screened resumes. Because moral outrage does not have a positive equivalent, we measured participants' broad positive and negative evaluations of the companies. We predicted that companies that used algorithms would be

judged less extremely in both cases: less negatively when promoting gender inequality and less positively promoting gender equality.

Participants

Two hundred and forty-one participants (128 male, 113 female; age: $M = 35.12$, $SD = 10.51$) from the United States completed the study on Prolific in exchange for 60 cents. Accounting for participants who might fail attention checks and therefore be excluded from the analysis, this sample size gives us a power of 0.95 to detect a medium effect size (Cohen's $d = 0.5$, calculated with G*Power 3.1.9.2). As specified in the pre-registration (https://aspredicted.org/zr7m8.pdf), we did not include in the analysis participants who failed to answer either of the attention checks correctly, leading to the exclusion of 30 participants.

Procedure

All participants first read a description of two high-tech companies, one using a human HR specialist for screening applicants and the other, an algorithm. Participants were randomly assigned to either a discrimination or an equality condition. In the discrimination condition participants read the following (equality condition in parentheses):

> AeonTech and CompSolutions are high-tech companies. Both companies hire a significantly lower (higher) percentage of women than the average at high-tech companies. They are on the bottom (top) 5% of high-tech companies in regards to gender equality. 95% of high-tech companies hire a larger (smaller) proportion of women than AeonTech and CompSolutions do.
> A series of external audits revealed why. At AeonTech, Sigma-Evalu8, a machine-learning-based algorithm was used to rate applicants. The algorithm systematically evaluated women more negatively (positively) than it evaluated men. At CompSolutions, a human HR specialist was used to rate applicants. The HR specialist systematically evaluated women more negatively (positively) than it evaluated men.

Assessing positive and negative evaluations of the companies. We then measured participants' positive and negative evaluations of each company by asking participants to rate their agreement with the following six statements on a 0 (strongly disagree) to 100 (strongly agree) slider for each company (positive evaluation in parentheses, company order was randomized): "AeonTech/CompSolutions deserves blame (praise) for its hiring decisions", "AeonTech/CompSolutions should be punished (rewarded) for its hiring decisions" and "I am angry at (happy with) AeonTech's/CompSolutions' hiring decisions". We created indices of negative moral evaluations by averaging the three negative items (Cronbach's $\alpha s > .89$), and for positive moral evaluations by averaging the three positive items (Cronbach's $\alpha s > .92$).

Manipulation and attention checks. To test whether participants saw both companies as more moral in the high gender equality and as less moral in the low gender equality condition, we asked participants to rate their agreement with the following two items on a 0 (strongly disagree) to 100 (strongly agree) slider: "AeonTech and CompSolutions hiring decisions are moral" and "AeonTech and CompSolutions hiring decisions are immoral". We then asked participants two attention checks. First, whether AeonTech used an algorithm and CompSolutions a human HR specialist or vis versa. Second, whether AeonTech and CompSolutions hired more or less women than other high-tech companies. Finally, participants provided demographic information.

Results

Manipulation check

A mixed-model 2 (evaluation type: positive, negative; within-subject) x 2 (condition: discrimination, equality; between-subject) ANOVA revealed that the companies were evaluated

more negatively in the discrimination condition (M = 62.58, SD = 29.17) than in the equality condition (M = 31.43, SD = 31.08), $F(1, 209) = 56.17$, $p < .001$, partial $\eta^2 = .212$, and more positively in the equality condition (M = 49.86, SD = 29.13) than in the discrimination condition (M = 24.44, SD = 25.47), $F(1, 209) = 45.29$, $p < .001$, partial $\eta^2 = .178$, interaction: $F(1, 209) = 63.46$, $p < .001$, partial $\eta^2 = 0.213$, supporting the success of our manipulation.

Evaluation of company

We conducted a 2 (outcome: discrimination, equality; between-participant) x 2 (agent: algorithm, human; within-participant) x 2 (evaluation: positive, negative; within-participant) mixed model ANOVA to test our hypothesis that people would evaluate discrimination by an algorithm (vs. a human) less negatively, and also see equality-supporting decisions by an algorithm (vs. a human) less positively.

As expected, we found a significant three-way interaction, $F(1, 209) = 16.63$, $p < .001$, partial $\eta^2 = .074$. A series of follow-up pairwise comparisons revealed that, as predicted, participants evaluated the company less negatively in the discrimination condition when it used an algorithm (M = 54.44, SD = 27.87) than a human (M = 59.74, SD = 27.81), $F(1, 209) = 6.13$, $p = .014$, partial $\eta^2 = .028$. We further found that, as predicted, participants had a less positive evaluation of the company for equality when it used an algorithm (M = 41.88, SD = 27.87) than a human (M = 50.83, SD = 27.62), $F(1, 209) = 19.53$, $p < .001$, partial $\eta^2 = .085$. We also found an unexpected effect such that participants evaluated equality by an algorithm (M = 28.20, SD = 26.65) more negatively than equality by a human (M = 20.23, SD = 23.94), $F(1, 209) = 14.78$, $p < .001$, partial $\eta^2 = .066$. The difference in positive evaluation of the discriminating company that used an algorithm (M = 14.36, SD = 19.04) and the company that used a human (M = 14.45, SD = 22.73), was not significant, $p = .965$. See Figure 4. Controlling for the order in which

participants rated the companies (AeonTech first or CompSoluation first) did not change the results.

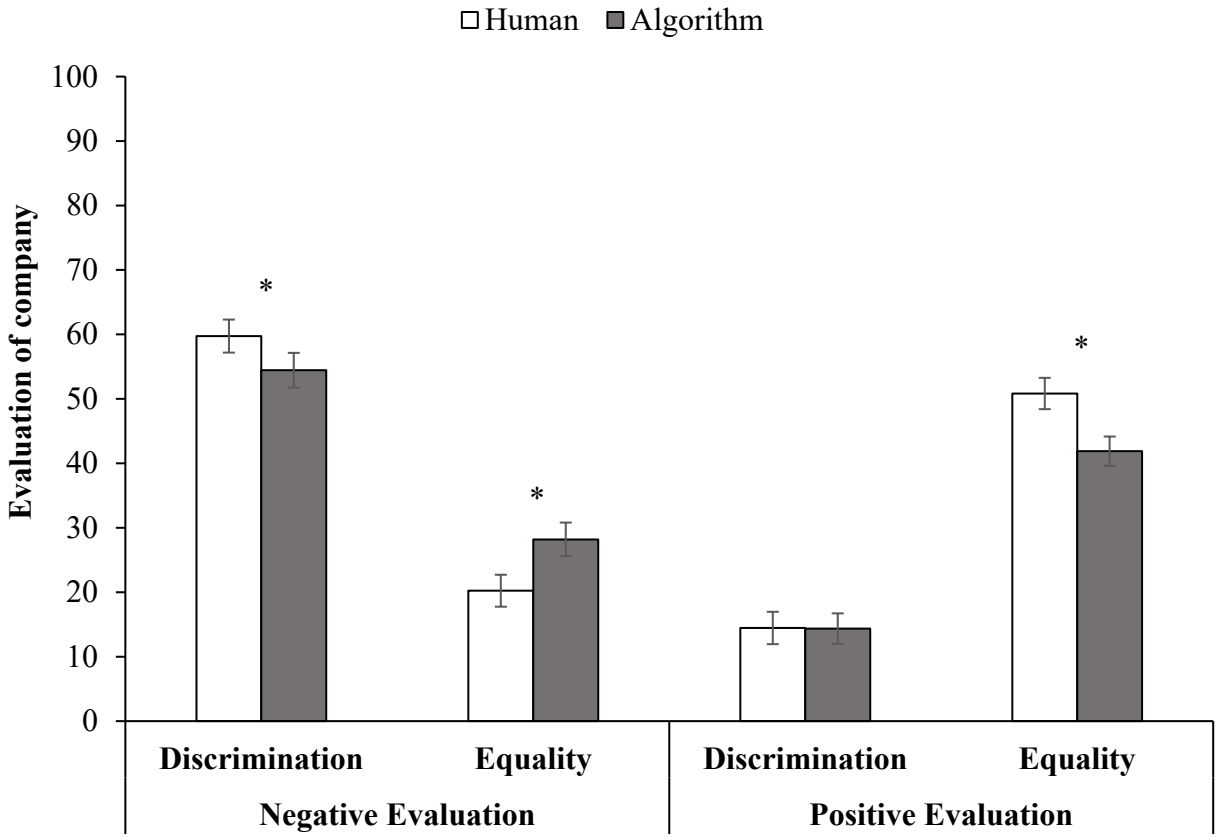

Figure 4. Negative and positive evaluations of companies that discriminated or had gender equality that used either a human HR specialist or an algorithm for screening applicants, Study 4.

* Denotes $p < .05$

The ANOVA further revealed significant effects for type of evaluation (positive vs. negative), partial $\eta^2 = .061$, agent, partial $\eta^2 = .036$, type of evaluation x outcome interaction, partial $\eta^2 =$

.394, and type of evaluation x agent interaction, partial η^2 = .022. As these analyses are not important for the current investigation, we do not elaborate on them (full data is available at OSF).

Discussion

The results of Study 4 show that the use of algorithms for screening applicants has a mitigating effect on how people evaluate the company. People evaluated the company less negatively when the algorithm screening resulted in low gender equality but also evaluate the company less positively when the algorithm screening resulted in high gender equality. This pattern is consistent with our theoretical framework. Since perceived motivation affects judgment, the acts of ostensibly motivationless algorithms affect judgments of the company less than the acts of motivated humans.

In Study 3 we did not find a difference in outrage post-discrimination between algorithm and human discrimination, while in Study 4 we did find a significant difference in negative evaluation. One possible reason for this apparent discrepancy is that in Study 3 we measured moral outrage, and in Study 4 general evaluations, such as blame and punishment, and these different measures might differ in how sensitive they are to the motivation of the agent. Another possibility is that the joint evaluation in Study 4 might make the difference between the algorithm and the human more salient (Hsee et al., 1999).

Studies 5-7: Potential Moderators of the Algorithmic outrage deficit

After finding that people are less morally outraged by algorithm discrimination and that algorithm discrimination causes less moral outrage at the company that uses the algorithm, we turn to explore possible moderators. In Study 5 we examine whether knowledge of AI moderates

the effect of discrimination by algorithm on moral outrage (with professional technology workers). In Study 6 we examine whether the identity of the programmers influences this effect. In Study 7 we examine whether people will be more outraged at discrimination by an algorithm when the algorithm is anthropomorphized. Understanding factors that mitigate algorithmic outrage deficit can outline the scope of the effect, and also provide additional support for the role of perceived prejudice in algorithmic outrage deficit.

Study 5: Tech Workers

In Study 5 we sampled workers in the technology industry in Norway. We had two goals in this study. First, we wanted to test the generalizability of the algorithmic outrage deficit using a sample of people that actually experienced a hiring process for the type of job we describe in our manipulation. Second, we wanted to explore whether the algorithmic outrage deficit might be a result of a lack of knowledge about algorithms. Participants in this study read about gender discrimination in hiring decisions done by a human or an algorithm and reported how outraged they were as well as other questions. We predicted that, as in our previous studies, people will be less outraged when algorithms discriminate.[5]

Method

Participants. We recruited participants working in the tech industry by approaching the HR managers of five Norwegian tech companies. The five organizations all provide financial

---

[5] We conducted this study before studies 1-4, and prior to getting valuable feedback from reviewers. This is why the vignette and the measures are slightly different than the ones in Studies 1-4 (but are similar to the replications reported in the discussion of Study 2, and to Study 6).

technology and enterprise technology services for banking and finance. The HR managers in the companies forwarded an email invitation to the study to all employees who work with technology. Out of the 292 people who started the survey 206 (159 male, 42 female, 5 other/preferred not to answer; age: $M = 34.51$, $SD = 10.63$) completed it, of which 51 failed the attention check and were excluded from the analysis. As we were not able to estimate in advance how fast data collection would be from this sample, we did not pre-register this study. We report all conditions and measures.

Procedure and measures. Participants were randomly assigned to an Algorithm or a Human condition. The study was conducted in Norwegian; we describe the English translation. In the Algorithm condition participants read the following (Human condition in parentheses):

> Imagine the following:
>
> COMPNET, an Artificial-Intelligence-based computer program (Mr. Davie, an HR specialist), was given ultimate power in the hiring process of programmers and engineers in your company two years ago.
>
> It was recently found that COMPNET (Mr. Davie) was biased against women when rating applicants' resumes'. COMPNET (Mr. Davie) put penalties on any resume using the word "women's", as in "women's chess club captain".
>
> This prevented many talented and qualified women engineers from getting high-paid jobs at your company.

Assessing outrage. After reading the scenario, participants rated their moral outrage. To assess moral outrage we used five items. The first item asked participants "Which of the following best expresses your opinion of the discriminatory actions of Mr. Davie/Compnet" (1 = completely acceptable; 3 = objectionable; 5 = absolutely shocking; 7 = outrageous). The other items asked participants to rate their agreement with the following four statements: "I am morally outraged by the discriminatory actions of Mr. Davie/Compnet", "The discriminatory actions of Mr. Davie/Compnet are unjust", "The discriminatory actions of Mr. Davie/Compnet

were immoral" and "The discriminatory actions of Mr. Davie/Compnet were wrong" (1 = Strongly disagree; 7 = Strongly agree). We then created a composite moral outrage index by averaging all five items, Cronbach's $\alpha = .86$.

Additional measures. We asked participants how worried and how concerned they would be about such discrimination (inter-item correlation: $r = .61$, $p < .001$), to what extent they thought that the human or the algorithm should be fired and replaced/discarded, and the extent to which they thought the company should make a public apology, do an internal audit and make an effort to hire more women (Cronbach's $\alpha = .69$). We found no significant difference between conditions for these variables (ps > .11).

Knowledge about AI. We then asked participants "compared to the average Norwegian, how knowledgeable are you about AI" (1 = much less knowledgeable; 7 = much more knowledgeable). Finally, as an attention check we asked participants whether a human or a software made hiring decisions in the story they read about and to provide demographic information.

Results

An independent samples *t*-test revealed that, as predicted, participants were less outraged when the discrimination was done by an algorithm ($M = 6.16$, $SD = 1.55$) than when the discrimination was done by a human ($M = 6.60$, $SD = 0.98$), $t(153) = 2.15$, $p = .033$, Cohen's $d = 0.34$.

To examine whether algorithm outrage is dependent on people's knowledge about AIs, we tested whether self-reported knowledge about AI moderated the algorithmic outrage deficit using a bootstrapping moderation analysis (we used Preacher & Hayes, 2008; 5000 iterations,

model 1). We found a significant Knowledge about AI x Condition interaction, $b = 0.24$, $SE = 0.10$, $t(151) = 2.46$, $p = .015$. A follow-up analysis revealed that while there was no difference between conditions for people 1 SD below the average of knowledge about AI ($p = .793$), there was a significant difference for people with an average knowledge about AI (conditional effect: $b = 0.21$, $SE = 0.10$, $t(151) = 2.09$, $p = .038$) and people 1 SD above the average of knowledge about AI (conditional effect: $b = 0.46$, $SE = 0.14$, $t(151) = 3.24$, $p = .002$), see Figure 5. To further examine this interaction, we ran two regression analyses, one for the Algorithm condition and one for the Human condition, each testing the relation between AI knowledge and moral outrage. AI Knowledge was negatively related (marginally significant) to moral outrage in the Algorithm condition ($\beta = -.23$, t (69)= -1.958, $p = .054$), suggesting that the more self-reported knowledge people had about AIs, the less outraged they were at discrimination by an algorithm. In contrast, in the Human condition, the relation between AI knowledge and moral outrage was not significant ($\beta = 0.15$, t (82) = 1.36, $p = .177$). Knowledge about AI, therefore, didn't mitigate the algorithmic outrage deficit, but actually aggravated it. The more people said they knew about AIs, the less outraged they were at algorithm (vs. human) discrimination.

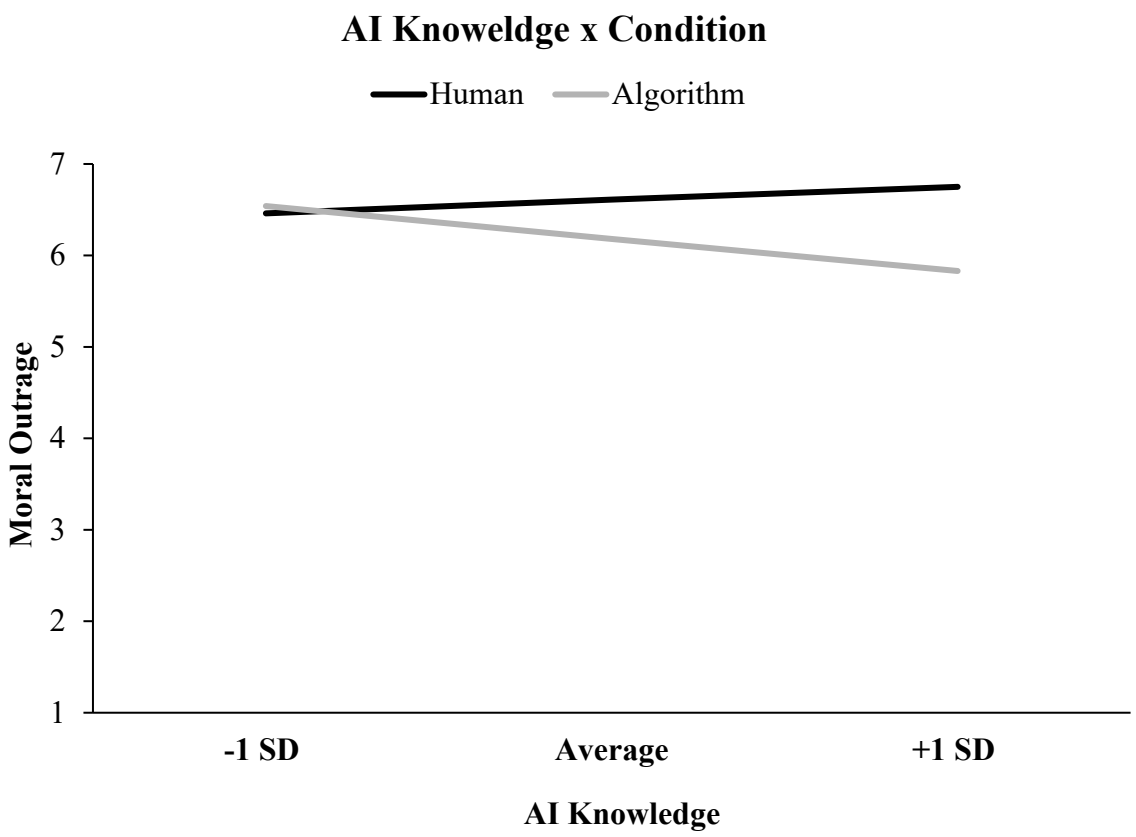

Figure 5. The interaction effect of AI knowledge and condition on moral outrage (Study 5), p = .015.

Discussion

The results of Study 5 generalized our findings in a sample of people who work in the high-tech sector, are familiar with algorithms, are in another country (Norway), and shows the robustness of the phenomenon. The results also point to an interesting influence of knowledge of AIs on algorithmic outrage deficit. The more people know about AIs, the less outraged they are at algorithm (vs. human) discrimination. This suggests that algorithmic outrage deficit is not due to people's lack of knowledge about AIs. It is possible that people who know more about AIs are less likely think that AIs are not motivated by prejudice, but the data we collected in this study

did not allow us to test this possibility. We further examined the role of attributed prejudiced motivation in Study 6.

Study 6: Manipulating Programmers

In our previous studies we found that perceived prejudiced motivation statistically explains (mediates) the algorithmic outrage deficit. However, mediation analysis provides only limited support to arguments about causality. A more powerful way to examine causality is by manipulating the mediator (Pirlott & MacKinnon, 2016). In Study 6, to test the causal role of perceived prejudiced motivation we manipulated perceived prejudiced motivation by manipulating the identity of the algorithm's programmers. There are two reasons why manipulating the identity of the programmers might affect the motivation attributed to the algorithm. The first is that people attribute to objects properties of disliked individuals who have been brief contact with them. For example, people are unwilling to wear a shirt that was previously worn by Hitler (Rozin et al., 1986). This attribution can be even stronger for objects that were created by disliked people, such as sexist programmers. The second is that people might think that sexist programmers will program their sexism into the algorithms they create. Participants read about gender discrimination in hiring decisions and were randomly assigned to one of four conditions. The first two conditions were similar to those in Study 5, describing gender discrimination in hiring decision either by an algorithm or by a human. In the two new conditions we provided participants with information about the identity of the algorithm's programmers. In one condition they were described as working for a known sexist company, and in the other as working for a more egalitarian company. We predict that in addition to replicating our previous findings, people will be more outraged at discrimination by an algorithm when the algorithm was programmed by a more sexist company than when it was programmed by a more

egalitarian company, as people might perceive the algorithm as sharing to some extent the motivation of its creators. We note that there is some research suggesting that people might be actually more outraged at socially responsible companies that behave unethically (King & McDonnell, 2012). However, according to our theory, discrimination is not just an action, it is an action in a context, and the context is prejudiced motivation. We therefore predict that since people will attribute less of a prejudiced motivation to an algorithm programmed by an egalitarian company, people will be less outraged when such an algorithm discriminates.

Method

Participants. Nine hundred and sixty-four participants[6] from the US and Canada (47.9% male, 51.6% female, 0.4% other or preferred not to disclose; age: $M = 36.93$, $SD = 11.96$) completed the study on Amazon's Mechanical Turk in exchange for 40 cents. As specified in the pre-registration (https://aspredicted.org/xq68n.pdf), we did not include in the analysis participants who failed to answer any of the attention check/comprehension questions correctly, leading to the exclusion of 181 participants.

Procedure. We randomly assigned participants to one of four conditions. Two of these conditions were similar to those used in Study 5, but instead of telling participants to imagine this happened in their company, we told them to imagine this happened in Amazon. In the

[6]As specified in the pre-registration, we started with 480 participants. However, one of the effects that was significant in our previous studies, the comparison between the "Algorithm-control" and the "Human HR specialist" conditions, was not significant in this sample ($p = .115$). In order to understand if this is a type I error in our previous studies or a type II error in this study, we ran an additional 480 participants, for increased statistical power. We pre-registered these additional participants, see https://aspredicted.org/4r5iq.pdf. We note that all tests are significant even after applying the Bonferroni correction for multiple comparisons.

Human condition participants read that a human, Mr. Davie, discriminated against women. In the "Algorithm Control" condition participants read that an algorithm, CompNet, discriminated against women. We included two additional conditions, in which participants read that CompNet discriminated against women and were giving information about the identity of CompNet's programmers. In the "Sexist Programmers" condition participants read the following:

> "COMPNET was developed by a company named Beyond Computers. Beyond Computers, founded and managed by men, is known in the industry as being a hostile work environment for women. 95% of its programmers are men, and men are systematically paid more than women."

In the "Egalitarian Programmers" condition participants read that:

> "COMPNET was developed by a company named Beyond Computers. Beyond Computers, founded and managed by women, is known in the industry as being very women friendly. It has an equal number of men and women programmers, and pays men and women exactly the same."

After reading the scenario participants reported their moral outrage using the same items we used Study 5, with one difference: the items were framed as asking about "CompNet's/Mr. Davie's behavior", rather than "discriminatory actions" (Cronbach's $\alpha = .91$). As a manipulation check, participants rated the prejudiced motivation they attributed to the agent, using the items we used in Studies 1-2 ($\alpha = .70$). As an attention check we then asked participants who made the hiring decision in the story they read: a human, a software without mention of its programmers, a software programmed by a company founded and managed by women, or a software

programmed by a company founded and managed by men. Finally, participants provided demographic information.

Results

Manipulation check – attribution of prejudiced motivation. A one-way ANOVA revealed that, as predicted, condition affected the attribution of prejudiced motivation, $F(3, 779) = 82.76$, $p < .001$, $\eta^2 = .21$, see Figure 6. We then ran a follow-up planned and pre-registered contrasts. The first contrast revealed that participants attributed less of a prejudiced motivation to CompNet in the Algorithm Control condition ($M = 4.20$, $SD = 1.35$) than in the Human condition ($M = 5.10$, $SD = 0.92$), $t(779) = 7.97$, $p < .001$, Cohen's $d = 0.77$, replicating our findings from Studies 1-2. The second contrast revealed that, as predicted, people attributed a more prejudiced motivation to CompNet in the Sexist Programmers condition ($M = 5.05$, $SD = 0.98$) than in the Egalitarian Programmers condition ($M = 3.79$, $SD = 1.21$), $t(779) = 12.08$, $p < .001$, Cohen's $d = 1.14$. Another contrast, which we did not pre-register, did not find a difference between the Human condition ($M = 5.10$, $SD = 0.92$) and the Sexist Programmers condition ($M = 5.05$, $SD = 0.98$), $p = .656$.

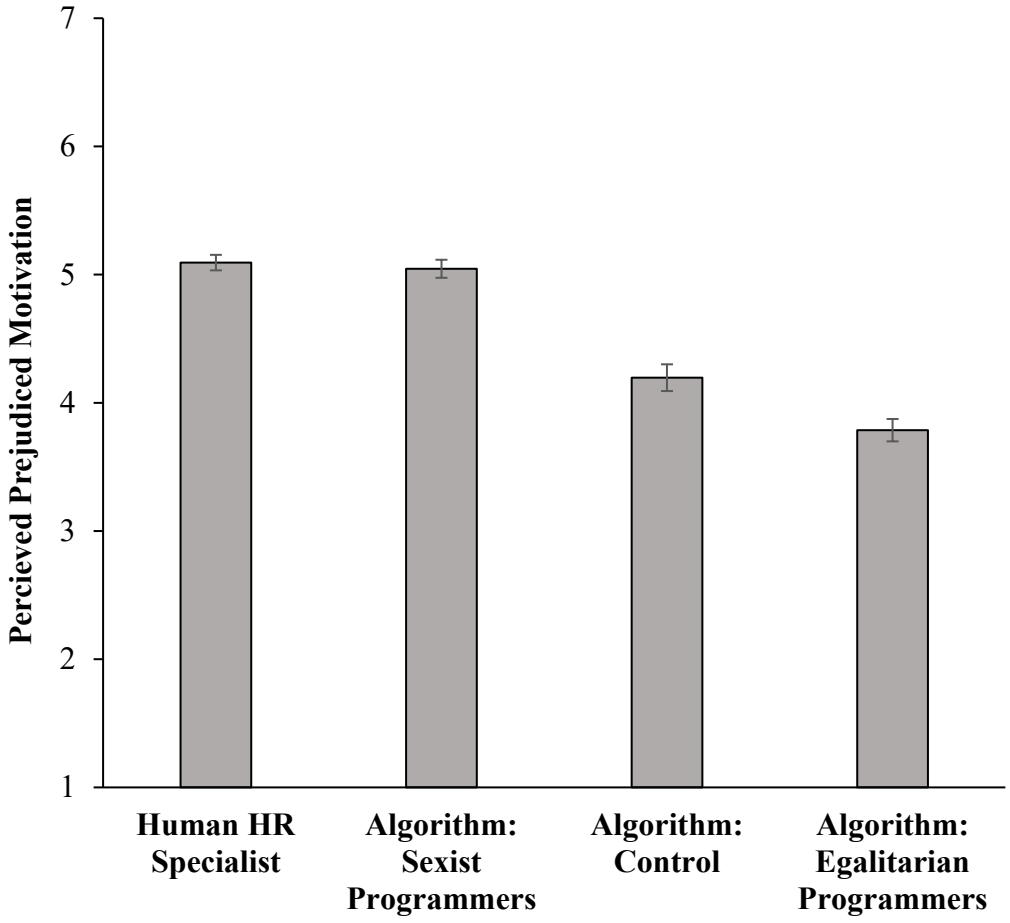

Figure 6. Perceived prejudiced motivation by condition (Study 6). All differences are significant ($p < .05$) expect for between the Human condition and the Sexist Programmers condition. Error bars reflect standard errors.

Moral outrage. A one-way ANOVA revealed that, as predicted, condition affected moral outrage, $F(3, 779) = 26.28$, $p < .001$, $\eta^2 = .10$, see Figure 7. We ran two follow-up planned contrasts. The first contrast revealed that, as predicted, participants were less morally outraged by the discrimination in the Algorithm Control condition ($M = 5.52$, $SD = 1.32$) than in the Human condition ($M = 5.85$, $SD = 1.08$), $t(779) = 2.74$, $p = .006$, Cohen's $d = 0.28$, replicating

our findings from Studies 2-5. The second contrast revealed that, as predicted, people were more outraged by CompNet in the Sexist Programmers condition (M = 5.83, SD = 1.09) than in the Egalitarian Programmers condition (M = 4.96, SD = 1.34), $t(779) = 7.56$, $p < .001$, Cohen's d = 0.79. Another contrast, which we did not pre-register, did not find a significant difference between the Human condition (M = 5.85, SD = 1.08) and the Sexist Programmers condition (M = 5.83, SD = 1.09), $p = .498$.

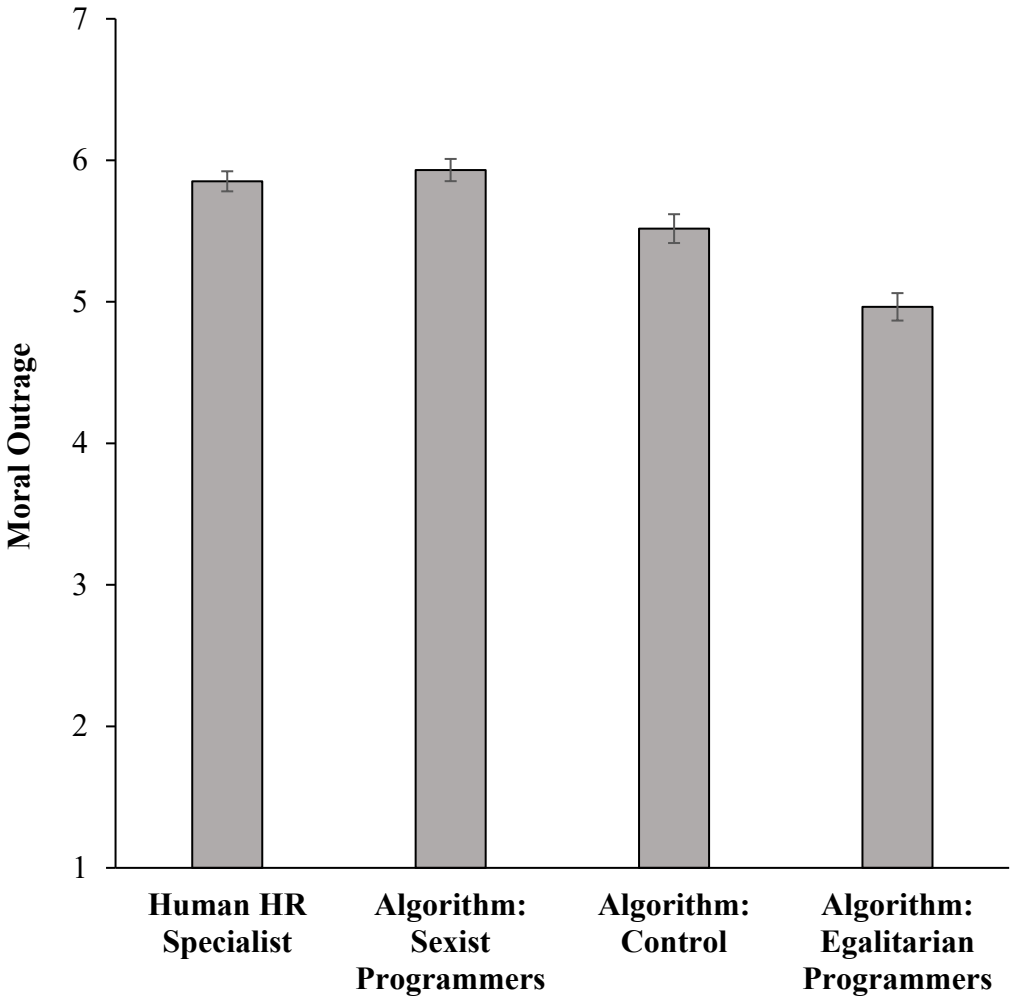

Figure 7. Moral outrage by condition (Study 6). All differences are significant ($p < .05$) except for between the Human condition and the Sexist Programmers condition. Error bars reflect standard errors.

Discussion

Study 6 provides further causal support for our hypothesis about the role of perceived prejudiced motivation in algorithmic outrage deficit, by examining the identity of the programmers as a moderator. For the same discriminatory actions, when the programmers are known to be sexist, people perceived the algorithm as having a stronger prejudiced motivation and were more outraged at the discrimination. On the other hand, when the programmers are known to be egalitarian, people perceived the algorithm as less motivated by prejudice and were less morally outraged when it discriminated. The results of Study 6 also point to a possible positive outcome for companies who have a diverse workforce. Even if the product (the algorithm) they create ends up discriminating, people will be less outraged at the discrimination, because they will see the algorithm as less motivated by prejudice, due to the diversity of its programmers. Study 6 rules out an alternative explanation for our findings. In all conditions, people attributed mid-level and higher prejudiced motivation to algorithms, suggesting that our results are not due to people seeing algorithms as inherently incapable of prejudiced motivation, but rather of people attributing less such motivation to algorithms. We note that our results in this study might be due to the halo effect (Nisbett & Wilson, 1977). Specifically, people's outrage might be affected by learning about the sexist or egalitarian company, and not due to an effect of the programmer's identity on how people perceive the algorithm. To address this possibility and further examine the causal role of perceived motivation on moral outrage, in Study 7 we approached this differently – by manipulating anthropomorphism.

Study 7: Anthropomorphism

In Study 7 we wanted to further examine the role of perceived prejudiced motivation in algorithmic outrage deficit. Rather than doing so by manipulating the identity of the programmers, as we did in Study 6, in Study 7 we did so by manipulating anthropomorphism – the extent that which people attribute human-like abilities, such as motivation, to non-human entities (de Visser et al., 2016; Schroeder & Epley, 2016; Waytz et al., 2014). This allowed us to examine the role of perceived motivation while keeping the identity of the agent constant (always an algorithm) and without mentioning the identity of the programmers (they are never mentioned). Participants read about an algorithm low on anthropomorphism or an algorithm high on anthropomorphism that discriminated against women in hiring decisions. We predicted that people will perceive anthropomorphized algorithms as motivated by prejudice more than non-anthropomorphized algorithms, causing participants to be more morally outraged at discrimination by the anthropomorphized algorithm.

Method

Participants

Eight hundred and one participants (378 male, 407 female, 16 other or preferred not to disclose; age: $M = 32.00$, $SD = 11.93$) from the United States completed the study on Prolific in exchange for 3 dollars[7]. Accounting for participants who might fail attention checks and therefore be excluded from the analysis, this sample size gives us a power of 0.80 to detect a two-tailed small effect size (Cohen's $d = 0.2$, calculated with G*Power 3.1.9.2). As specified in

[7] This study was conducted through a different lab's prolific account, with different payment norms.

the pre-registration (https://aspredicted.org/q397v.pdf), we did not include in the analysis participants who failed the attention check, leading to the exclusion of sixty-four participants.

Procedure

Participants were randomly assigned to a high-anthropomorphism or a low-anthropomorphism condition. All participants first read the following paragraph:

> AeonTech is a technology company that has an automated process for hiring software developers. They rely on an algorithm—SigmaEvalu8, an unsupervised Bayesian AI system that scans each application and assigns it between one (lowest fit) and five stars (highest fit). Applicants with the highest scores are then forwarded to the hiring committee.
>
> Then, in the high-anthropomorphism, participants read the following:
>
> The algorithm is just like a human being in that it develops its own tastes and social preferences from scanning past job applications. You can think of it as a human on a hiring committee. It gets a feeling for which features of an application it likes in predicting future job performance and then chooses new applications based on the opinions it has formed. In other words, the algorithm is just like a human that thinks about what it wants in terms of a job candidate and uses its beliefs to select people it likes based on these features.
>
> Despite receiving large numbers of applications from women for software engineering positions, AeonTech has hired almost no women.
>
> The reason for this has to do with the initial screening. AeonTech's algorithm, SigmaEvalu8, through forming its own humanlike preferences for a particular social category, systematically gave women a lower star rating than men. Therefore, women almost never made it to the second phase, and were not in the final list that the hiring committee considered.
>
> In the low-anthropomorphism condition, participants read the following:
>
> The algorithm is a tool that computes information from scanning past job applications. You can think of it as a supercomputer. It calculates what features of an application are useful in predicting good performance and scores new applications based on this information. In other words, the algorithm is a mathematical process that extracts useful features from existing data sources and uses them to select optimal job candidates. Despite receiving large numbers of applications from women for software engineering

> positions, AeonTech has hired almost no women.
> The reason for this has to do with the initial screening. AeonTech's algorithm, SigmaEvalu8, through producing its positive computational ouput for a particular social category, systematically gave women a lower star rating than men. Therefore, women almost never made it to the second phase, and were not in the final list that the hiring committee considered.

Assessing likeness as a manipulation check. Participants then rated their agreement with the four statements on a 0 (strongly disagree) to 100 (strongly agree) slider. Two items measured perceptions of human-likeness: "SigmanEvalu8 behaves like a human being" and "SigmaEvalu8 has a mind". We created an index of perceived human-likeness by averaging these two items, r =.61. The other two items measured perceived machine-likeness: "SigmaEvalu8 operates like a tool" and "SigmaEvalu8 works like a machine". We created an index of perceived machine-likeness by averaging these two items, r =.71.

Assessing perceived prejudiced motivation and moral outrage. We then measured perceived prejudiced motivation (Cronbach's α = .89) and moral outrage (Cronbach's α = .96), using the same items as in Studies 1-2 and 6.

Participants then answered two attention check questions. The first asked whether an algorithm or a human HR specialist screened the applications. The second asked whether the story they read described gender discrimination in hiring practices. Finally, participants provided demographic information.

Results

Manipulation check.

To test whether our anthropomorphism manipulation was successful in changing participants' perceptions of the algorithm, we conducted a 2 (condition: high-anthropomorphism, low- anthropomorphism; between-participants) x 2 (likeness: human, machine; within-participants) mixed model ANOVA. The ANOVA revealed a main effect for likeness, $F(1, 734) = 817.34$, $p < .001$, partial $\eta^2 = .527$, such that across conditions, participants perceived the algorithm as higher in machine-likeness ($M = 73.54$, $SD = 24.75$) than in human-likeness ($M = 30.22$, $SD = 26.92$). The main effect for condition was also significant, $F(1, 737) = 14.69$, $p < .001$, partial $\eta^2 = .020$, such that across types of likeness, participants perceived the algorithm as being higher in human and machine likeness in the high-anthropomorphism condition ($M = 53.69$, $SE = 0.68$) than in the low-anthropomorphism condition ($M = 50.00$, $SE = 0.69$).

These main effects were qualified by a significant condition x likeness interaction, $F(1, 737) = 117.96$, $p < .001$, partial $\eta^2 = .138$. Follow-up pairwise comparisons revealed that participants perceived the algorithm as being more humanlike in the high-anthropomorphism condition ($M = 40.18$, $SD = 26.63$) than in the low-anthropomorphism condition ($M = 19.94$, $SD = 23.11$), $F(1, 734) = 120.95$, $p < .001$, partial $\eta^2 = .141$. The analysis further revealed that participants attributed to the algorithm more machine-likeness in the low-anthropomorphism condition ($M = 80.06$, $SD = 21.65$) than in the high-anthropomorphism condition ($M = 67.20$, $SD = 25.93$), $F(1, 734) = 53.20$, $p < .001$, partial $\eta^2 = .068$. These results suggest that our anthropomorphism manipulation successfully changed the way the participants perceived the algorithm.

Perceived prejudiced motivation and moral outrage.

An independent samples *t*-test revealed that, as predicted, participants attributed more prejudiced motivation to the algorithm in the high-anthropomorphism condition ($M = 59.79$, $SD = 28.65$) than in the low-anthropomorphism condition ($M = 46.42$, $SD = 33.12$), $t(735) = 5.87$, $p < .001$, Cohen's $d = 0.43$. Contrary to our prediction, a second independent samples *t*-test did not find a significant difference in moral outrage between the high-anthropomorphism condition ($M = 51.83$, $SD = 31.93$) and the low-anthropomorphism condition ($M = 52.12$, $SD = 33.82$), $t(735) = 0.12$, $p = .905$.

Mediation

Following our pre-registration and to further test the relation between anthropomorphism, perceived prejudiced motivation, and moral outrage, we performed a bootstrapping mediation analysis (Preacher & Hayes, 2008; 5000 iterations, model 4) with anthropomorphism as the IV, coding the low-anthropomorphism condition as -1 and the high-anthropomorphism condition as 1, moral outrage as a DV, perceived prejudiced motivation as a mediator.

The analysis revealed, as predicted, a significant indirect effect of anthropomorphism on moral outrage mediated by perceived prejudiced motivation $b = 4.55$, $SE = 0.81$, $CI_{.95}$[2.97, 6.13]. The direct effect of anthropomorphism on moral outrage was significant and negative, $b = -4.40$, $SE = 0.95$, $CI_{.95}$[-6.27, -2.53], see Figure 8. Taken together, these results show anthropomorphism has a dual effect on moral outrage. It causes people to perceive the anthropomorphized algorithm as being more motivated by prejudice and therefore feel more outraged. But there is another factor, evident by the significant direct effect, causing them to be less outraged at the anthropomorphized algorithm.

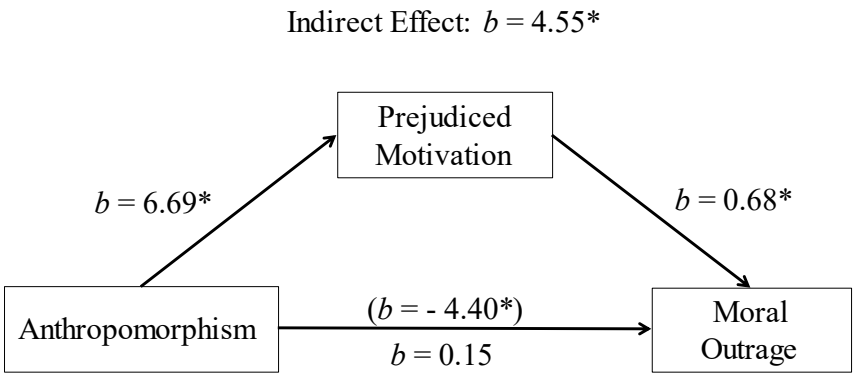

Figure 8. Mediation analysis reveals that anthropomorphism indirectly affects moral outrage via perceived prejudiced motivation (Study 7). While the total effect of anthropomorphism is not significant, the indirect effect vie perceived prejudiced motivation is significant ($b = 4.55$, $SE = 0.81$, $CI._{95}[2.99, 6.16]$).

Discussion

The results of Study 7 provide further support for the role of perceived motivation in algorithmic outrage deficit. Participants perceived highly anthropomorphized algorithms as being more motivated by prejudiced motivation, causing them, indirectly, to be more outraged at algorithm discrimination. While in our previous studies the perceived motivation was confounded with the identity of the agent (algorithm vs. human), or the identity of the programmers, Study 7 shows that even for the same agent – an algorithm – the perceived prejudiced motivation affects moral outrage. However, the overall effect on moral outrage was more nuanced. In addition to the significant indirect effect vie perceived prejudiced motivation, we found a significant direct effect in the opposite direction: taking into account the increased outrage due to an increase in perceived prejudiced motivation, people were less outraged at

discrimination by the highly-anthropomorphized algorithm. This is consistent with research showing that anthropomorphism tends to increase people's trust in algorithms and robots (de Visser et al., 2016; Waytz et al., 2014) and with research showing that anthropomorphism increases people's forgiveness after wrong-doing (Tang & Gray, 2021; Yam et al., 2020). Overall, the non-significant main effect of anthropomorphism could be explained by these opposing forces—perceived motivation (activated by anthropomorphism) increases moral outrage (as our other studies show), but anthropomorphism generally can orient people toward positive evaluations of a machine, engendering trust and forgiveness.

In Studies 1-7 we examined the effect of algorithm discrimination on moral outrage and general evaluations of companies who use algorithms. In Study 8 we examine a possible downstream consequence of algorithm discrimination.

Study 8: Perceived Legal Liability

In Study 8 we examined a possible downstream consequence of people perceiving algorithms as less motivated by prejudice than humans – company liability. Title VII of the United States Civil Rights Act of 1964 prohibits discrimination based on gender, race, color or national origin, or religion (Civil Rights Act, 1964). One type of discrimination lawsuit that can be filed under Title VII is cases of intentional discrimination (Civil Rights Act, 1964). Since people see algorithms as less motivated by prejudice than humans, they might see companies as less liable for intentional discrimination when an algorithm (vs. a human) discriminates. In Study 8, participants read about gender discrimination in a high-tech company by an algorithm or a human and rated the perceived prejudiced motivation of the agent. Participants then read that the female applicants who were not hired are suing the company for intentional discrimination under Title VII. We asked participants how liable they thought the company is. We predicted that the

company will be judged as less liable when an algorithm (vs. a human) discriminated, and that this will be mediated by participants perceiving the algorithm as less motivated by prejudice than the human.

Method

Participants. To assess responses from the group not subject to discrimination in our scenario, we sampled only men for this study. Two hundred and forty participants from the US and Canada (age: $M = 32.19$, $SD = 11.46$) completed the study on Prolific in exchange for 45 cents. As specified in the pre-registration (https://aspredicted.org/de5r2.pdf), we did not include in the analysis participants who failed to answer any of the attention checks correctly, leading to the exclusion of fifteen participants.

Procedure. All participants first read the same scenario as in Study 1 about gender discrimination in hiring decisions at a high-tech company. We then measured the prejudiced motivation participants attributed to the agent they read about using the same items as in Studies 1, 2, 6 and 7 (Cronbach's $\alpha = .94$). After that, participants read about female programmers who are suing the company for intentional liability:

> Protective Title VII of the Civil Rights Act of 1964 makes it illegal to discriminate against people based on gender.
> One type of discrimination lawsuit that can be filed under Title VII are cases of intentional discrimination, which require discriminatory intent. In other words, it requires showing intentional prejudice.
> A group of female programmers who applied for jobs at AeonTech are suing AeonTech under Title VII.
> As a reminder, an audit found that the reason AeonTech didn't hire women is that SigmaEvalu8 (the HR specialist) gave women a lower rating than men.

Assessing liability. To assess how much participants thought the company is liable for intentional discrimination, we asked participants to rate how much they disagree or agree with

the following three statements on a 0 (strongly disagree) to 100 (strongly agree) slider: "I would support a verdict that said the company had discriminatory intent", "The company is liable because it acted out of intentional prejudice" and "The company is not liable, because there is no evidence of intentional discrimination" (reverse scored). We then created a composite stereotype endorsement index by averaging all three items, Cronbach's $\alpha = .90$.

Participants then answered two attention check questions. The first asked whether an HR specialist or an algorithm did the initial screening in the story they read. The second asked whether the company they read about hardly hired any women or hired a similar number of men and women. Finally, participants provided demographic information.

Results

Attribution of prejudiced motivation. An independent samples *t*-test revealed that participants attributed less prejudiced motivation to SigmaEvalu8 ($M = 48.25$, $SD = 33.39$) than they did to the HR specialist ($M = 67.21$, $SD = 24.10$), $t(223) = 4.89$, $p < .001$, Cohen's $d = 0.65$.

Liability. An independent samples *t*-test also revealed that participants found the company less liable when the discrimination was done by an algorithm ($M = 49.08$, $SD = 31.59$) than a human ($M = 66.58$, $SD = 35.63$), $t(223) = 4.57$, $p < .001$, Cohen's $d = 0.61$.

Mediation. To test whether perceived prejudiced motivation mediated the effect of agent on liability, we performed a bootstrapping mediation analysis (Preacher & Hayes, 2008; model 4, 5000 iterations), coding the algorithm condition as 1 and the human condition as -1. As predicted, the effect of agent on liability, $b = -8.75$, $SE = 1.91$, $p < .001$, was mediated by an indirect effect of attribution of prejudiced motivation, $b = -6.43$, $SE = 1.32$, $CI_{.95}[-9.13, -3.91]$. When accounting for the mediation by attribution of prejudiced motivation, the direct effect of

agent on stereotype endorsement was not significant, b = -2.32, SE = 1.47, $CI_{.95}$[-5.22, 0.57]. The decrease in liability judgments when the discrimination was done by an algorithm appears, therefore, to be driven by people attributing less of a prejudiced motivation to the algorithm (vs. the human), see Figure 9.

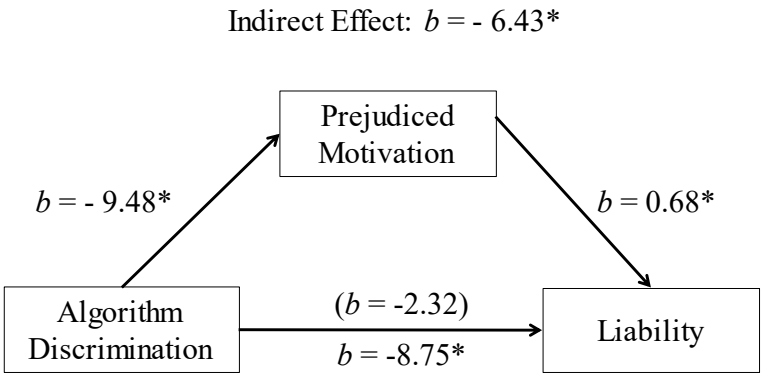

Figure 9. Mediation analysis reveals that perceived prejudiced motivation mediates the effect of agent on liability judgments (Study 8). * Denotes $p < .05$.

Discussion

The results of Study 8 demonstrate one possible downstream effect of algorithm discrimination. People find companies less liable when they discriminate because of human bias than when they discriminate because of algorithm bias, an effect mediated by people perceiving algorithms as less motivated by prejudiced motivation.

General Discussion

Eight studies provide evidence for the idea of *algorithmic outrage deficit*. People are less outraged when gender discrimination is perpetrated by algorithms versus other people because

algorithms are not perceived to possess the motivation to discriminate. We found that people attribute less of a prejudiced motivation to algorithms than humans (Study 1), leading people to be less outraged at discrimination by an algorithm than discrimination by a human (Study 2). Algorithm discrimination had a nuanced effect on people's moral outrage at the company using the algorithms. Learning that a company used an algorithm (vs. a human) for screening applicants initially increased moral outrage (consistent with Bigman & Gray, 2018). However, relative to this baseline of outrage, people were less outraged when an algorithm (vs. a human) caused the discrimination (Study 3). The use of algorithms (vs. humans) for hiring decisions mitigated how people evaluated a company. Companies were judged less negatively for negative outcomes such as low gender equality and less positively for positive outcome such as high gender equality (Study 4). The algorithmic outrage deficit is not due to people's lack of knowledge about AIs. The more people know about AI, the less outraged they were at algorithm discrimination (Study 5). We found converging evidence for the role of perceived motivation (Studies 6-7). When the discriminatory algorithm was created by sexist programmers, people perceived it as more motivated by prejudice and were more outraged (Study 6). In contrast, when the algorithm was created by egalitarian programmers it was perceived as less motivated by prejudice and people were less outraged (Study 6). Furthermore, when the algorithms were anthropomorphized, people attributed to them more prejudiced motivation, causing an increase in moral outrage (Study 7). Finally, we also found that people are less likely to find the company legally liable for intentional discrimination when an algorithm (vs. a human) was the cause of discrimination (Study 8).

Algorithms are playing an increasingly active role in several domains where they sometimes discriminate, including: the legal system (Angwin et al., 2016), healthcare

(Obermeyer et al., 2019), banking (Stankiewicz, 2019), advertising (Lambrecht & Tucker, 2019) and HR decisions (Dastin, 2018). As a first step, in our studies, we focused on hiring decisions, but we suspect that a similar effect will be found in other domains such as sentencing, parole, consumer targeting, and medical assessment as well. Most of our studies examined gender discrimination, inspired by the algorithm Amazon used (Dastin, 2018), but some of our studies examined cases of race (see "Study 2 Replication A: Race Discrimination" in the supplemental materials) and age discrimination (see "Study 2 Replication B: Age Discrimination" in the supplemental materials) and found the same pattern, showing generalizability beyond the type of discrimination. We used diverse samples including online panels from Mturk and Prolific, a quasi-representative sample (see "Study 2 Replication D: Quasi-Representative Sample" in the supplemental materials), and workers in high-tech companies (Study 5). Our samples include people from the US, UK, Canada, and Norway. This diversity of samples demonstrates the robustness of our findings.

We do not argue that the algorithmic outrage deficit is irrational or a bias. It is indeed possible that algorithms are less likely to unfairly discriminate between people according to their race, age, and gender than humans (Mullainathan, 2019). However, it is still crucial to understand how people respond to algorithms when they do show bias.

Our research contributes to a growing literature on how people respond to the presence of algorithms in day-to-day life. Algorithms are seen as more impartial and less capable of bias than human decision-makers (Jago & Laurin, 2021), leading people to prefer them as decision-makers in some situations (Bigman et al., 2021; Logg et al., 2019). However, in some cases people show an aversion to algorithm decision-making (Dietvorst et al., 2015) and prefer for algorithms not to act as decision-makers. At least sometimes, this aversion can be a result of the flip side of

algorithm impartiality – algorithms have a reductionist approach to human nature that lacks empathy and emotion which are seen as necessary for some decisions (Bigman & Gray, 2018; Longoni et al., 2019; Newman et al., 2020). Our research shows another consequence of the perceived impartiality of algorithms – people perceive them as lacking prejudiced motivation and are therefore less morally outraged when they discriminate. We note that although algorithms do discriminate, algorithm biases are easier to correct than human biases (Mullainathan, 2019) and can actually be used to detect human discrimination (Logg, 2019).

Implications

Our research has both theoretical and practical implications. First, our research contributes to research on moral outrage. By demonstrating the role of attribution of prejudiced motivation we contribute to the literature suggesting that moral outrage is not only a response to harm (Hechler & Kessler, 2018). While previous research on moral outrage compared intentional to non-intentional harm (Hechler & Kessler, 2018; Russell & Giner-Sorolla, 2011), our current work proposes that a broader set of mental states – motivations – affect moral outrage. People care not only whether an agent intentionally discriminated – performed the action with the belief that it will lead to a specific outcome – or not, but also about the agent's motivation – why they discriminated (see Carlson et al., 2022, for a discussion of the distinction between intention and motivation). We note that while our results from Study 2 suggest that intention might also play a role in explaining moral outrage (see Footnote 3), perceived motivation mediates algorithm outrage deficit above and beyond perceived intention. In doing so, our work contributes to the growing literature in moral psychology highlighting the role of perceived motivation in moral judgment (e.g., Bigman & Tamir, 2016; Carlson & Zaki, 2018; Levine & Schweitzer, 2014; Reeder et al., 2002).

Second, our work has implications for research in human-robot interaction. Specifically, our research contributes to the literature on how people perceive the mental states of artificial agents such as robots and algorithms (Bigman & Gray, 2018; Graaf & Malle, 2019; Gray & Wegner, 2012; Schein & Gray, 2015; Waytz et al., 2014; Weisman et al., 2017; Young & Monroe, 2019) by showing that people are less likely to attribute prejudiced motivation to algorithms. While previous work focused on the mind (mental capacities) robots and algorithms are perceived to have (e.g., Bigman & Gray, 2018), our work explores motivation, the content of a specific mental state. Looking at this more detailed level of attributions opens new promising venues for investigating how people react, respond, and interact with algorithms and other artificial agents.

Third, our research complements current work in computer science, legal studies, and other disciplines on how to create fair algorithms (Abdul et al., 2018; Ananny, 2016; Kusner & Loftus, 2020; Sandvig et al., 2016; Selbst & Barocas, 2018; Wachter et al., 2017; Zou & Schiebinger, 2018). While work on "algorithm ethics" discusses how to create fair algorithms, our work starts exploring the psychological response of biased algorithms, and how they differ from the psychological response to biased humans. Understanding these differences is a first necessary step to address the unique challenges that biased algorithms pose to society.

Finally, our research has implications for organizations that use algorithms for decisions such as hiring. Fairness perceptions of the recruitment and selection procedure affect the overall reputation of an organization (Chapman et al., 2005; Ryan & Ployhart, 2000). Furthermore, applicants who are hired at the end of an unfair selection process will develop lower levels of organizational commitment, less organizational citizenship behaviors, and higher turnover (Hausknecht et al., 2004; Uggerslev et al., 2012). Our research suggests that, at least for some

decisions, emphasizing the role of algorithms in the hiring process might increase perceptions of fairness, increase worker satisfaction, and reduce moral outrage when the outcomes of the process are discriminatory. In addition, if companies recognize that while "blaming the algorithm" could protect the company reputationally, it might have the adverse effect of making discrimination feel more permissible. Furthermore, the results of Study 6 suggest that when creating algorithms, companies can reduce public outrage when those algorithms discriminate by having a diverse team of programmers develop these algorithms.

Limitations and future directions

Despite the robustness of our findings, our work has several limitations. People might be less outraged at discrimination by algorithms because algorithms are easier to de-bias than humans (Mullainathan, 2019). We note that one reason why it might be easier to de-bias algorithms than humans is that algorithms lack the prejudiced motivation that might cause some people to discriminate. Future research should investigate this interesting possibility. There also might be individual and cultural differences in the way people respond to discrimination by algorithms. For example, it is possible that people who are high on anthropomorphizing non-humans (Waytz, Cacioppo, et al., 2010) might attribute more motivation to algorithms and therefore not show reduced outrage at discrimination by machines.

Our results suggest that people might attribute to algorithms characteristics of their creators and programmers (see Study 6). This finding raises several interesting questions – when would people see algorithms as independent of their creators? Who would people blame for harm done by algorithms: the algorithms, the programmers, or the company that uses the algorithms? Future research is needed to investigate this question.

The results of Study 5 show that the more people know about AIs the less outraged they are by algorithm discrimination. Presumably, this is because people with high knowledge about AIs have different beliefs about why algorithms discriminate. It is also possible that the results we obtained in our studies are due to people having certain beliefs about the reasons for algorithm discrimination. We explored this in a short study where participants read about a screening algorithm that discriminated against women. We asked participants (N=183 after exclusions) how much they thought the algorithm was 1) trained to mimic previous decisions; 2) trained to make decisions based on the views and beliefs of its programmers; and 3) trained to make its own independent decisions. Participants showed similar levels of agreement with all of these statements, F (2, 176) = 0.79, p = .465. These results suggest that there is not a strong distinction in people's beliefs about the reasons for algorithm discrimination. Still, future research is needed to explore how beliefs about the technical reasons for algorithm discrimination (beyond lack of prejudiced motivation) affect moral outrage.

There are also specific sub-findings in certain studies that could benefit from future research. In Replication D, reported in the discussion of Study 2, we found a significant interaction with gender (F(1, 1267) = 5.07, p = .025, partial η^2 = .004), such that while men were less outraged at gender discrimination by an algorithm (F(1, 1267) = 26.02, p < .001, partial η^2 = .020, women's reduced outrage at algorithm discrimination was only marginally significant (F(1, 1267) = 3.78, p = .052, partial η^2 = .003) (see supplemental materials). This analysis was exploratory and further research is needed to systematically explore whether the group suffering from the discrimination reacts differently to algorithm discrimination than an unaffected group.

Across studies, there are also differences in effect sizes that need to be explored. In the replications reported in the discussion of Study 2 we found algorithmic outrage deficit for race

and age discrimination in addition to gender discrimination. The effect sizes ranged between Cohen's *d*'s of 0.26 and 0.80. Future research is needed to fully understand the source of the variance in effect sizes between the different types of discrimination and different populations.

Finally, our research focused on how people who are not affected by the discrimination of an algorithm respond to it. Another interesting question is how people who were targets of discrimination will respond to discrimination by an algorithm. Research shows that people who are discriminated against suffer from negative psychological consequences such as depression (Finch et al., 2000; Noh et al., 1999) and anxiety (Soto et al., 2011). Would people show less or more of these responses when they are discriminated against by an algorithm rather than a human? Further research is needed to explore this question, which will help us understand the full psychological consequences of discrimination by algorithms.

Concluding remarks

The increasing abilities and prevalence of machine-learning-based AI and autonomous machines raise new ethical and societal concerns. Here we highlight one of them, the *algorithmic outrage deficit*. We find that people attribute less prejudiced motivation to algorithms and consequently are less morally outraged by discrimination by algorithms. Beyond the contribution of this research to the study of moral outrage and human-robot interaction, our work has a warning sign for society. Algorithms carry the promise of being fairer than humans. However, when they are not, people's defenses against injustice might be lowered when the agent is an algorithm, making it easier for discrimination to go unnoticed and unopposed.

References

Full study materials and data can be found at:

https://osf.io/87yu5/?view_only=36fc6f1a3e004665a1e916be5fd180db

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*, 1–18. https://doi.org/10.1145/3173574.3174156

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574. https://doi.org/10.1037/0033-2909.126.4.556

Ananny, M. (2016). Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science Technology and Human Values*, *41*(1), 93–117. https://doi.org/10.1177/0162243915606523

Angwin, J., Larson, J., Surya, M., & Lauren, K. (2016). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Aquilina, Y., & Saliba, M. A. (2019). An automated supermarket checkout system utilizing a SCARA robot: preliminary prototype development. *Procedia Manufacturing*, *38*, 1558–1565. https://doi.org/10.1016/j.promfg.2020.01.130

Bares, W., Mott, B., Zettlemoyer, & Lester, J. (2007). *US Patent 7,305,345 B2*.

Batson, C. D., Chao, M. C., & Givens, J. M. (2009). Pursuing moral outrage: Anger at torture. *Journal of Experimental Social Psychology*, *45*(1), 155–160. https://doi.org/10.1016/j.jesp.2008.07.017

Batson, C. D., Kennedy, C. L., Nord, L. A., Stocks, E. L., Fleming, D. A., Marzette, C. M., Lishner, D. A., Hayes, R. E., Kolchinsky, L. M., & Zerger, T. (2007). Anger at unfairness: Is it moral outrage? *European Journal of Social Psychology*, *37*(6), 1272–1285. https://doi.org/10.1002/ejsp.434

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal?: A Field Experiment on Labor Market Discrimination. In *The American Economic Review* (Vol. 94, Issue 4, pp. 991–1013). Routledge. https://doi.org/10.4324/9780429499821-53

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, *145*(12), 1654–1669. https://doi.org/10.1037/xge0000230

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding Robots Responsible: The Elements of Machine Morality. *Trends in Cognitive Sciences*, *23*(5), 365–368. https://doi.org/10.1016/j.tics.2019.02.008

Bigman, Y. E., Yam, K. C., Marciano, D., Reynolds, S. J., & Gray, K. (2021). Threat of racial and economic inequality increases preference for algorithm decision-making. *Computers in Human Behavior*, *122*(May). https://doi.org/10.1016/j.chb.2021.106859

Bowen, D. E., Ledford, G. E., & Nathan, B. R. (2011). Hiring for the organization, not the job. *Executive*, *5*(4), 35–51. https://doi.org/10.5465/ame.1991.4274747

Cai, X., & Li, K. . (2000). A genetic algorithm for scheduling staff of mixed skills under multi-criteria. *European Journal of Operational Research*, *125*(2), 359–369. https://doi.org/10.1016/S0377-2217(99)00391-4

Cárdenas-Barrón, L. E., Treviño-Garza, G., & Wee, H. M. (2012). A simple and better algorithm to solve the vendor managed inventory control system of multi-product multi-constraint economic order quantity model. *Expert Systems with Applications*, *39*(3), 3888–3895. https://doi.org/10.1016/j.eswa.2011.09.057

Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., & Crockett, M. J. (2022). How inferred motives shape moral judgment. *Manuscript under Review*.

Carlson, R. W., & Zaki, J. (2018). Good deeds gone bad: Lay theories of altruism and selfishness. *Journal of Experimental Social Psychology*, *75*, 36–40. https://doi.org/10.1016/j.jesp.2017.11.005

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant Attraction to Organizations and Job Choice: A Meta-Analytic Review of the Correlates of Recruiting Outcomes. *Journal of Applied Psychology*, *90*(5), 928–944. https://doi.org/10.1037/0021-9010.90.5.928

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease Prediction by Machine Learning Over Big Data From Healthcare Communities. *IEEE Access*, *5*, 8869–8879. https://doi.org/10.1109/ACCESS.2017.2694446

Cheong, M., Lederman, R., McLoughney, A., Njoto, S., Ruppanner, L., & Wirth, A. (2020). *Ethical Implications of AI Bias as a Result of Workforce Gender Imbalance*. https://www.unibank.com.au/-/media/unibank/about-us/member-news/report-ai-bias-as-a-

result-of-workforce-gender-imbalance.ashx

Civil Rights Act of 1964, Pub. L. No. 78 Stat. 241 (1964). https://www.govinfo.gov/content/pkg/STATUTE-78/pdf/STATUTE-78-Pg241.pdf

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276. https://doi.org/10.1016/0010-0277(89)90023-1

Covert, B. (2019). Nearly two decades ago, women across the country sued Walmart for discrimination. They're not done fighting. *Time*. https://time.com/5586423/walmart-gender-discrimination/

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

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

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

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. https://doi.org/10.1037/xge0000033

Eriksson, K., Strimling, P., & Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, *129*, 59–69. https://doi.org/10.1016/j.obhdp.2014.09.011

Finch, B. K., Kolody, B., & Vega, W. A. (2000). Perceived Discrimination and Depression among Mexican-Origin Adults in California. *Journal of Health and Social Behavior*, *41*(3), 295–313.

Fiske, A. P., & Tetlock, P. E. (1997). Taboo trade-offs: Reactions to transactions that transgress the spheres of justice. *Political Psychology*, *18*(2), 255–297. https://doi.org/10.1111/0162-895X.00058

Foot, P. (1967). The Problem of Abortion and the Doctrine of the Double Effect. In *Oxford Review* (Issue 5, pp. 5–15).

Ford, M. (2015). *The Rise of the Robots: Technology and the Threat of Mass Unemployment*. Oneworld publications.

Fornili, K. S. (2018). Racialized Mass Incarceration and the War on Drugs. *Journal of Addictions Nursing*, *29*(1), 65–72. https://doi.org/10.1097/JAN.0000000000000215

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343–376. https://doi.org/10.1037/pspa0000086

Goel, S. (2018). Third generation sexism in workplaces: Evidence from India. *Asian Journal of*

Women's Studies, *24*(3), 368–387. https://doi.org/10.1080/12259276.2018.1496616

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*(March), 97–103. https://doi.org/10.1016/j.socec.2018.04.003

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726

Graaf, M. M. A. De, & Malle, B. F. (2019). People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. *Proceedings of the International Conference on Human-Robot Interaction, HRI'19*.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science (New York, N.Y.)*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & López-Pérez, B. (2016). Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology*, *65*, 94–104. https://doi.org/10.1016/j.jesp.2016.04.004

Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Science*. Oxford University Press.

Halzack, S. (2019). Peloton, Nike, Walmart and other brands get savaged online, but are fine in real life. *Bloomberg*. https://www.bloomberg.com/opinion/articles/2019-12-16/nike-peloton-walmart-etc-savaged-online-fine-in-real-life

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant Reactions to Selection Procedures: An Updated Model and Meta-Analysis. *Personnel Psychology*, *57*(3), 639–683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

Hechler, S., & Kessler, T. (2018). On the difference between moral outrage and empathic anger: Anger about wrongful deeds or harmful consequences. *Journal of Experimental Social Psychology*, *76*(March), 270–282. https://doi.org/10.1016/j.jesp.2018.03.005

Heilweil, R. (2019). *Artificial intelligence will help determine if you get your next job*. Vox. https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576–590. https://doi.org/10.1037/0033-2909.125.5.576

Jackson, J. C., Castelo, N., & Gray, K. (2020). Could a rising robot workforce make humans less prejudiced? *American Psychologist*, *November*. https://doi.org/10.1037/amp0000582

Jago, A. S., & Laurin, K. (2021). Assumptions About Algorithms' Capacity for Discrimination. *Personality and Social Psychology Bulletin*, 014616722110161. https://doi.org/10.1177/01461672211016187

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian

psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

King, B., & McDonnell, M.-H. (2012). Good Firms, Good Targets: The Relationship between Corporate Social Responsibility, Reputation, and Activist Targeting. *SSRN Electronic Journal*, *1990*, 12–30. https://doi.org/10.2139/ssrn.2079227

Kotkin, M. J. (2009). Diversity and discrimination: a look at complex bias. *William and Mary Law Review*, *50*(5), 1439–1500.

Kramer, M. F., Borg, J. S., Conitzer, V., & Sinnott-Armstrong, W. (2018). When Do People Want AI to Make Decisions ? *Proceedings of First Annual AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES-18)*.

Kurzban, R., Descioli, P., & Obrien, E. (2007). Audience effects on moralistic punishment☆. *Evolution and Human Behavior*, *28*(2), 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. *Nature*, *578*(7793), 34–36. https://doi.org/10.1038/d41586-020-00274-3

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. https://doi.org/10.1287/mnsc.2018.3093

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, *53*, 107–117. https://doi.org/10.1016/j.jesp.2014.03.005

Levitin, G., Rubinovitz, J., & Shnits, B. (2006). A genetic algorithm for robotic assembly line

balancing. *European Journal of Operational Research*, *168*(3), 811–825. https://doi.org/10.1016/j.ejor.2004.07.030

Li, J., Zhao, X., Cho, M.-J., Ju, W., & Malle, B. F. (2016, April 5). *From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars*. https://doi.org/10.4271/2016-01-0164

Lindenmeier, J., Schleer, C., & Pricl, D. (2012). Consumer outrage: Emotional reactions to unethical corporate behavior. In *Journal of Business Research* (Vol. 65, Issue 9, pp. 1364–1373). https://doi.org/10.1016/j.jbusres.2011.09.022

Logg, J. M. (2019). Using Algorithms to Understand the Biases in Your Organization. In *Harvard Business Review*. https://hbr.org/2019/08/using-algorithms-to-understand-the-biases-in-your-organization

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*(February), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Machery, E., & Mallon, R. (2010). Evolution of Morality. In *The Moral Psychology Handbook* (pp. 3–46). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199582143.003.0002

Major, B., Kaiser, C. R., & McCoy, S. K. (2003). It's not my fault: When and why attributions to

prejudice protect self-esteem. *Personality and Social Psychology Bulletin*, *29*(6), 772–781. https://doi.org/10.1177/0146167203029006009

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101–121.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, *March*, 125–132. https://doi.org/10.1109/HRI.2016.7451743

Manuel, S. K., Howansky, K., Chaney, K. E., & Sanchez, D. T. (2017). No Rest for the Stigmatized: A Model of Organizational Health and Workplace Sexism (OHWS). *Sex Roles*, *77*(9–10), 697–708. https://doi.org/10.1007/s11199-017-0755-x

Martin, J., Brickman, P., & Murray, A. (1984). Moral outrage and pragmatism: Explanations for collective action. *Journal of Experimental Social Psychology*, *20*(5), 484–496. https://doi.org/10.1016/0022-1031(84)90039-8

Miller, D. T., Effron, D. A., & Zak, S. V. (2011). From moral outrage to social protest: The role of psychological standing. *The Psychology of Justice and Legitimacy*, *November*, 103–124. https://doi.org/10.4324/9780203837658

Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology:*

General, *146*(1), 123–133. https://doi.org/10.1037/xge0000234

Mullainathan, S. (2019, December 6). Biased Algorithms Are Easier to Fix Than Biased People. *The New York Times*. https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

Nelissen, R. M. A., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making*, *4*(7), 543–553.

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*(June), 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250–256. https://doi.org/10.1037//0022-3514.35.4.250

Noh, S., Beiser, M., Kaspar, V., Hou, F., & Rummens, J. (1999). Perceived Racial Discrimination , Depression , and Coping : A Study of Southeast Asian Refugees in Canada Author (s): Samuel Noh , Morton Beiser , Violet Kaspar , Feng Hou and Joanna Rummens Published by : American Sociological Association Stable URL : *Journal of Health and Social Behavior*, *40*(3), 193–207.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, *66*(September 2015), 29–38. https://doi.org/10.1016/j.jesp.2015.09.012

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil.* (pp. 91–108). American Psychological Association. https://doi.org/10.1037/13091-005

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology*, *83*(4), 789–803. https://doi.org/10.1037/0022-3514.83.4.789

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the Laws of Sympathetic Magic in Disgust and Other Domains. *Journal of Personality and Social Psychology*, *50*(4), 703–712. https://doi.org/10.1037/0022-3514.50.4.703

Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger, but not moral disgust, responds to intentionality. *Emotion (Washington, D.C.)*, *11*(2), 233–240. https://doi.org/10.1037/a0022598

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' Perceptions of Selection Procedures and Decisions: A Critical Review and Agenda for the Future. *Journal of Management*, *26*(3), 565–606. https://doi.org/10.1177/014920630002600308

Salerno, J. M., & Peter-Hagene, L. C. (2013). The Interactive Effect of Anger and Disgust on Moral Outrage and Judgments. *Psychological Science*, *24*(10), 2069–2078. https://doi.org/10.1177/0956797613486988

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2016). When the algorithm itself is a racist: Diagnosing ethical harm in the basic Components of Software. *International Journal of Communication*, *10*(June), 4972–4990.

Schein, C., & Gray, K. (2015). The eyes are the window to the uncanny valley: Mind perception, autism and missing souls. *Interaction Studies*, *16*(2), 173–179. https://doi.org/10.1075/is.16.2.02sch

Schroeder, J., & Epley, N. (2016). Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General*, *145*(11), 1427–1437. https://doi.org/10.1037/xge0000214

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, *87*(3), 1085–1139. https://doi.org/10.2139/ssrn.3126971

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, *86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Shapiro, A. (2017). Reform predictive policing. *Nature*, *541*(7638), 458–460. https://doi.org/10.1038/541458a

Soto, J. A., Dawson-Andoh, N. A., & BeLue, R. (2011). The relationship between perceived discrimination and Generalized Anxiety Disorder among African Americans, Afro Caribbeans, and non-Hispanic Whites. *Journal of Anxiety Disorders*, *25*(2), 258–265. https://doi.org/10.1016/j.janxdis.2010.09.011

Spring, V. L., Cameron, C. D., & Cikara, M. (2018). The Upside of Outrage. *Trends in Cognitive*

Sciences, *22*(12), 1067–1069. https://doi.org/10.1016/j.tics.2018.09.006

Srinivasan, R., & Sarial-Abi, G. (2021). When Algorithms Fail: Consumers' Responses to Brand Harm Crises Caused by Algorithm Errors. *Journal of Marketing*, *85*(5), 74–91. https://doi.org/10.1177/0022242921997082

Stankiewicz, K. (2019). *Twitter complainer says Apple is to blame for credit card issues*. CNBC. https://www.cnbc.com/2019/11/11/apple-shouldnt-pass-the-blame-on-gender-bias-says-complainant.html

Sunstein, C. R., Kahneman, D., & Schkade, D. (1998). Assessing punitive damages. *Yale Law Journal*, *107*(50), 2071–2153.

Takeshita, T., Tomizawa, T., & Ohya, A. (2006). A house cleaning robot system - Path indication and position estimation using ceiling camera. *2006 SICE-ICASE International Joint Conference*, 2653–2656. https://doi.org/10.1109/SICE.2006.315049

Tang, S., & Gray, K. (2021). Feeling empathy for organizations: Moral consequences, mechanisms, and the power of framing. *Journal of Experimental Social Psychology*, *96*(September), 104147. https://doi.org/10.1016/j.jesp.2021.104147

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*(3), 451–471. https://doi.org/10.1037//0033-295X.109.3.451

Uggerslev, K. L., Fassina, N. E., & Kraichy, D. (2012). Recruiting Through the Stages: A Meta-Analytic Test of Predictors of Applicant Attraction at Different Stages of the Recruiting Process. *Personnel Psychology*, *65*(3), 597–660. https://doi.org/10.1111/j.1744-

6570.2012.01254.x

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*, 72–81. https://doi.org/10.1177/1745691614556679

Uhlmann, E. L., Zhu, L. L., & Diermeier, D. (2014). When actions speak volumes: The role of inferences about moral character in outrage over racial bigotry. *European Journal of Social Psychology*, *44*(1), 23–29. https://doi.org/10.1002/ejsp.1987

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334. https://doi.org/10.1016/j.cognition.2012.10.005

Validi, S., Bhattacharya, A., & Byrne, P. J. (2015). A solution method for a two-layer sustainable supply chain distribution model. *Computers and Operations Research*, *54*, 204–217. https://doi.org/10.1016/j.cor.2014.06.015

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, Explainable, and Accountable AI for Robotics. *Science Robotics*, *2*(6), eean6080.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010).

Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410–435. https://doi.org/10.1037/a0020240

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *2017*, 201704347. https://doi.org/10.1073/pnas.1704347114

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7398–7401. https://doi.org/10.1073/pnas.0502399102

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2020). Robots at work: People prefer—and forgive—service robots with perceived feelings. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000834

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, *85*(August 2018). https://doi.org/10.1016/j.jesp.2019.103870

Zou, J., & Schiebinger, L. (2018). Design AI so that its fair. *Nature*, *559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8