Algorithmic Discrimination Causes Less Moral Outrage than Human Discrimination

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

[1]University of North Carolina at Chapel Hill

[2] BI Norwegian Business School

[3] Kellogg School of Management, Northwestern University

*Corresponding author. Email: ybigman@gmail.com

Abstract

The use of algorithms hold promise for overcoming human biases in decision making. Companies and governments are using algorithms to improve decision-making for hiring, medical treatments, and parole. Unfortunately, as with humans, some of these algorithms make persistently biased decisions, functionally discriminating people based on their race and gender. 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. Six studies test this *algorithmic outrage asymmetry* hypothesis across diverse discrimination in hiring practices (sexism, ageism, racism) and across diverse participant groups (online samples, a quasi-representative sample, and a sample of tech workers). As predicted, people are less morally outraged by algorithmic discrimination. The studies further reveal that this algorithmic outrage asymmetry is driven by the reduced attribution of prejudicial motivation to machines. We also reveal a downstream consequence of algorithmic outrage asymmetry—people are more likely to endorse racial stereotypes after algorithmic discrimination versus human discrimination. We discuss the theoretical and practical implications of these results, including the potential weakening of collective action to address systemic discrimination.

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

Gender discrimination often causes moral outrage, in a similar way to other forms of unfairness (Batson et al., 2007; Russell & Giner-Sorolla, 2011; Spring, Cameron, & Cikara, 2018; Sunstein, Kahneman, & Schkade, 1998; Tetlock, 2002). This outrage often seems especially high when discrimination is perpetrated by companies (Halzack, 2019), such as Walmart, who has been accused of systematically paying women less than men and overlooking women for promotions (Covert, 2019). Companies are collectives of people and so corporate discrimination is ultimately done by the *people* who control hiring, firing, and salaries. However, the rise of artificial intelligence raises a new possibility: systematic discrimination can be perpetrated by *algorithms*. Amazon had developed a machine-learning-based algorithm to screen resumes for software developer applications, and in 2018 it came to light that the algorithm was systematically biased against women, downgrading applicants whose resumes contained terms such as "women's chess club captain" or the names women colleges. The algorithm was soon scrapped amid outrage about its systematic gender (Dastin, 2018).

Although both the Walmart and the Amazon cases involved systematic gender discrimination and elicited much outrage, the case of the Amazon algorithmic discrimination seemingly elicited less moral outrage than that of Walmart. Here we explore whether people truly are more blasé at the "hands" of an algorithm versus a human being—what we call an *algorithmic outrage asymmetry*—and one potential reason for this asymmetry. People may perceive algorithms as less motivated by prejudice, and therefore are less outraged when they discriminate. We also explore one downstream consequences of the algorithmic outrage asymmetry, namely whether people are more likely to endorse stereotypes after discrimination by an 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 can and are performed now by autonomous machines and algorithms, such as working in an assembly line (Levitin, Rubinovitz, & Shnits, 2006), customer service (Bares, Mott, Zettlemoyer, & Lester, 2007), house cleaning (Takeshita, Tomizawa, & Ohya, 2006) and supermarket checkout (Aquilina & Saliba, 2019). Machine learning has facilitated this development, providing data based predictions which often outperform humans in areas such as predicting the spread of disease (Chen, Hao, Hwang, Wang, & Wang, 2017) and crime (Shapiro, 2017). Algorithms have particularly revolutionized many practices in the business world, helping to manage inventory (Cárdenas-Barrón, Treviño-Garza, & Wee, 2012), distribution chains (Validi, Bhattacharya, & Byrne, 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 unrest, the rise of AI may also bring people together by emphasizing their shared humanity (Jackson, Castelo, & Gray, 2020). Another concern about the rise of algorithms is distrust in machines' decision-making capacities. Many choices in the law, medicine, and the military 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 & Gray, 2018; Bigman, Waytz, Alterovitz, & Gray, 2019; Gogoll & Uhl, 2018; Kramer, Borg, Conitzer, & Sinnott-Armstrong, 2018; Young & Monroe, 2019). Unlike people who care deeply about their family and friends, machines are devoid of compassion, and people are hesitant to allow algorithms to make decisions about human lives because of their dispassionate impartiality.

Although some may see the impartiality of machines as a disadvantage—at least in some situations (Longoni, Bonezzi, & Morewedge, 2019)—others argue that this impartiality holds the promise to free decision-making from human biases (Mullainathan, 2019). Basic propensities for ingroup bias and prejudice can lead humans to discrimination across many domains, especially those involving hiring. Substantial research reveals that people are biased in their decisions about who to interview, who to hire, and to 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 obvious discrimination, companies are increasingly relying on algorithms to free hiring practices from bias (Heilweil, 2019).

Unfortunately, the promise of unbiased decision-making is currently unfulfilled, as AIs have been found to discriminate in many cases. We reviewed one high profile example in the first paragraph—the Amazon hiring algorithm—but there are many others, such as race discrimination in algorithms used for in parole decisions (Angwin, Larson, Surya, & Lauren, 2016) and healthcare (Obermeyer, Powers, Vogeli, & Mullainathan, 2019), and gender discrimination in credit scores (Stankiewicz, 2019) and in the display of online STEM career ads (Lambrecht & Tucker, 2019). The reasons that machine-learning based algorithms often discriminate is because they are often trained on biased datasets (Brunet, Alkalay-Houlihan, Anderson, & Zemel, 2019; Torralba & Efros, 2011), which reflect existing social inequities. For example, the bias of Amazon's hiring algorithm was driven in part by the algorithm learning that in years prior, software engineers who were hired at Amazon were predominantly men (Dastin, 2018). Regardless of the causes for algorithm discrimination, it is important to understand how people respond to cases in which algorithms perpetrate racial and gender discrimination. In order to answer this question we need to first understand how people respond to discrimination in general.

Moral Outrage

Although discrimination is often wide-spread and institutionalized (Fornili, 2018; Goel, 2018; Manuel, Howansky, Chaney, & Sanchez, 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]. Theorizing suggests that this outrage serves a number of important social functions. For example, moral outrage mobilizes people to punish unfair behavior (Fiske & Tetlock, 1997; Gummerum, Van Dillen, Van Dijk, & López-Pérez, 2016; Nelissen & Zeelenberg, 2009; Salerno & Peter-Hagene, 2013; Spring et al., 2018) which deters uncooperative behaviors (Kurzban, Descioli, & Obrien, 2007; Xiao & Houser, 2005). In addition, moral outrage leads to and promotes collective action (Martin, Brickman, & Murray, 1984; Miller, Effron, & Zak, 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 it demotivates employees (Hausknecht, Day, &

[1] We note that there is an ongoing debate on whether moral outrage is a real psychological 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 on the emotional response to discrimination as an outcome measure, rather than on the construct validity of moral outrage.

Thomas, 2004), increases turnover (Uggerslev, Fassina, & Kraichy, 2012), and even causes the public to boycott the discriminating company (Lindenmeier, Schleer, & Pricl, 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 sexism and racism perpetrated by an algorithm could generate more outrage than discrimination perpetrated by a single person. Algorithms are usually implemented to manage or transform entire large-scale operations, which means they have the ability to impact a large number of people. Whereas a single hiring administrator could discriminate against potentially dozens of applicants, an algorithm—with its limitless throughput—has a capacity to discriminate against hundreds or thousands of applicants. As people weigh the impact (or potential impact) in their moral judgments (Batson, Chao, & Givens, 2009; Batson et al., 2007; 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 descriptive normative acts—acts that are frequent or typical (Gawronski, Armstrong, Conway, Friesdorf, & Hütter, 2017; Malle, Guglielmo, & Monroe, 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). 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 there are some reasons that algorithms might generate more moral outrage than the actions of human beings, here we suggest that algorithms might actually generate less outrage for similar discrimination. We label this prediction the *algorithm outrage asymmetry*, 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; Levine & Schweitzer, 2014; Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002) with the work showing that people ascribe different mental states to robots (K. Gray & Wegner, 2012; Li, Zhao, Cho, Ju, & Malle, 2016; Waytz, Heafner, & Epley, 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 and immorally, one pressing question in the minds of perceivers is "Why?"—why did the agent behaved the way they did? Understanding why people behave the way they do is important because it allows us to make sense of the social world, predict their future behavior and guides the way we interact with them (Cosmides, 1989; Waytz, Morewedge, et al., 2010). Understanding why an agent behaved the way they did provides valuable information about the moral character of the agent. A resurgence in moral psychology emphasizes that morality is about judging the character of other people (Goodwin, Piazza, & Rozin, 2014; Pizarro & Tannenbaum, 2011; Uhlmann, Pizarro, & Diermeier, 2015; Uhlmann, Zhu, & Diermeier, 2014; Uhlmann, Zhu, & Tannenbaum, 2013), and an important part of moral character is the motivations that shape the persons behavior (Bigman & Tamir, 2016; Levine &

Schweitzer, 2014; Reeder et al., 2002). Consider for example the notorious trolley problem in which a person need to decide whether or not to actively kill one person in order to save five others (Foot, 1967). A person who decides to kill one person in order to save five other will be judged very differently if they are motivated to save as many people as possible than if they just wanted to push someone to their death (Kahane et al., 2018). Similarly we propose that the motivation attributed to a wrongdoer will affect moral outrage.

The motivation people attribute to others might be especially important in reactions to hiring decisions because of their attributional ambiguity. Hiring involves weighing many different candidate features, including education, experience, skills, and general "fit" (Bowen, Ledford, & Nathan, 2011). When a white or male candidate is hired over a person of color or a women, 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 decider seems unlikely—or unable—to harbor ill-will towards a social group, their biased decisions may elicit less outrage.

Consistent with this possibility, research shows that people perceive the mental states of algorithms differently than the mental states of humans (Bigman & Gray, 2018; K. Gray & Wegner, 2012; Malle, Scheutz, Forlizzi, & Voiklis, 2016; Weisman, Dweck, & Markman, 2017). They are seen as being somewhat less able to think rationally and plan their actions, and especially less able to experience emotions (Bigman & Gray, 2018; H. M. Gray, Gray, & Wegner, 2007; K. Gray & Wegner, 2012). The way people perceive robots affects how much people trust robots (Gogoll & Uhl, 2018), blame them (Malle et al., 2016; Shank & DeSanti, 2018), and want them to make decisions (Bigman & Gray, 2018; Young & Monroe, 2019). We suggest that this underlying difference in mental state attribution contributers to algorithmic outrage asymmetry. More specifically, we predict that people will attribute less prejudiced motivation to a discriminating algorithm (versus a discriminating person) and would therefore be less outraged by algorithms discriminations. We test the algorithmic outrage asymmetry in the studies described below.

We note that although we predict that people will not perceive algorithms *per se* as having a prejudiced motivation, 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 a stronger prejudiced motivation to a discriminatory algorithm when it is programmed by software company tied to sexist work conditions. We also test another phenomenon tied to algorithmic asymmetry—whether algorithms' perceived lack of prejudiced motivation can cause people to view the discrimination as more justified and the underlying stereotype that drives the discrimination as more justified. For example, if an algorithm gives a lower credit limit to immigrants, will participants see immigrants as being worse at maintaining credit? This possibility suggests that any potential apathy to algorithm discrimination may have insidious effects.

Current Research

We present a systematic investigation of people's moral outrage about discrimination perpetrated by algorithms through 6 studies that use diverse methods and samples, including a nationally representative sample of the UK, and a sample of employees at Norwegian technology firms. In Study 1 we coded open-ended questions to discover whether people spontaneously attributed less prejudice to a discriminatory algorithm (vs. a human). In Studies 2A-2D we tested whether people will be less morally outraged at algorithm discrimination than human discrimination across diverse domains (e.g., gender, race). In Study 3 we tested whether attribution of prejudiced motivation mediated the reduced outrage at algorithm discrimination. In Study 4 we manipulated the attribution of prejudiced motivation to the algorithm by manipulating whether its programmers seen as more sexist or more egalitarian. In Study 5 we replicate our findings in a sample of workers in the tech industry. Finally in Study 6 we tested another downstream consequence of the reduced attribution of prejudiced motivation to algorithms—stereotype endorsement. We predict 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 and more endorsement of stereotypes. To better situate our research, before describing our studies we first discuss data from a PEW survey on people's attitudes towards algorithms making decisions that might involve discrimination towards certain groups. All studies were approved by the IRB of The University of North Carolina at Chapel Hill. We pre-registered all studies except for the Pilot Study, Study 2A and Study 5. Full study materials can be found at https://osf.io/87yu5/?view_only=36fc6f1a3e004665a1e916be5fd180db.

Pilot Study: Data from Pew Survey

We suggest that the proposed algorithm outrage asymmetry for discriminatory outcomes is driven by perceptions that algorithms are not motivated to discriminate. That is, machines are seen as generally more fair, even if they perpetrate discrimination. To examine this hypothesis in a large pre-existing representative data set, we analyze data from PEWs 27th wave of the American Trends Panel. Data in this survey (N = 5,174) was collected between May 1st and May

15th 2017, and is representative of the US population ages 18 and older (Pew Research Center, 2017). The full data is available upon request from PEW Research Center. With this data we tested whether and why people are accepting of algorithms making decisions that might result in discrimination, such as hiring decisions.

A subsample of participants in the survey (*n* = 2,090) read the following text:

> Today, when companies are hiring they typically have someone read applicants' resumes and conduct personal interviews to choose the right person for the job. In the future, computer programs may be able to provide a systematic review of each applicant without the need for human involvement. These programs would give each applicant a score based on the content of their resume, application, or standardized tests for skills such as problem solving or personality type. Applicants would then be ranked and hired based on those scores.

Participants than answered a few questions about the vignette. Most participants (75.5%, which are 1578 out of 2090) said they would not want to apply for a job that uses this type of computer program to make hiring decisions (23.8%, which are 498 out of 2090 said they will apply and 0.7%, which are 14 out of 2090, refused to answer). Interestingly, the most common justification given by participants who said they would apply for such jobs was that they thought it would be fairer and less biased (42.6%, which are 212 out of 498). In contrast, only less than 2% (27 out of 1578) of participants who said they would not apply for such jobs mentioned concerns about fairness and bias.

These data are consistent with previous research showing that people might not want machines to make high-stake decisions (Bigman & Gray, 2018; Young & Monroe, 2019). Most importantly, these data suggest that although people might have some misgivings about algorithms making such decisions, people do believe that algorithms are fairer and less biased than human. We therefore propose that people would be less likely to attribute a prejudiced motivation to algorithm, and would be less outraged when algorithms discriminate.

Study 1: Attribution of Prejudice

Given that our Pilot Study shows that people see algorithms as fair, Study 1 sought to test whether people are less likely to attribute a prejudiced motivation to an algorithm vs. a human. We asked participants to read a story about discrimination against women by either a human or an algorithm. In order to measure participants' responses without affecting them with leading them with questions, we asked participants to write down their thoughts about the story and why they think the algorithm or the human acted the way they did. Then two independent raters coded these open responses for attributions of prejudiced motivation. We predicted that people will attribute less prejudiced motivation to the algorithm vs. the human.

Method

Participants. Two hundred and forty participants (93 male, 145 female, 2 other/declined to answer; age: *M* = 34.64, *SD* = 12.91) from the UK, US and Canada completed the study on Prolific in exchange for 40 cents. Accounting for participants who might fail attention checks and would 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/blind.php?x=j4a2sk[2].), 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 twenty four participants.

Procedure. Participants were randomly assigned to one of two conditions. In the Algorithm condition participants read the following:

> At the end of 2019 both members of a married couple applied for a credit card. Both the husband wife shared equal ownership of their assets.
> Their applications were evaluated by CompNet, a machine-learning algorithm. After evaluating their applications, CompNet evaluated the man much more positively than the woman. It decided to give the man a credit limit of $10,000, but only gave the woman a credit limit of $500.
>
> An internal audit suggested that CompNet systematically gives men more positive evaluations than women.

Participants in the Human condition read a similar story, but instead of reading that CompNet made the decision, participants read that Mr. Jonathan Miller, a mid-level employee made the decision. After reading the story participants answered two open-ended questions:

[2] We first had two independent coders code the responses according to the instructions reported in the pre-registration: 2 - the reason for the behavior is prejudice. For example, "[agent] is sexist"; 1 - both a prejudice-based reason and a facts/statistical reason are mentioned; 0 - the reason for the behavior is facts/statistics. For example, "men on average make more than women". However, that coding resulted in a very low inter-rater reliability (Kappa <.1). We therefore modified our pre-registered coding instructions and gave a different set of coders the modified instructions, as detailed in the methods section. Results remain unchanged when we used the coding as detailed in the pre-registration.

"Why do you think CompNet/Jonathan decided to give the man higher credit score than the women? Please be as detailed as possible" and "Please write a few general thoughts about CompNet/Jonathan". As an attention check, participants then were asked who made the credit decision in the story they read, a human or an algorithm. Participants then provided demographic information and completed a second attention check. In this attention check participants saw a picture of an apple. In the apple there is a text asking participants to write down the name of the object they see, explicitly telling them that it is an attention check. Below the picture was be a text box with the heading "do you have any comments about the study".

Coding open replies. We first went over participants' open replies and replaced any mention of the agent (e.g., "CompNet", "the algorithm", "Jonathan", "Mr. Miller") with a string of stars ("******"). Then, two independent coders, who were blind to the research hypothesis, rated the edited replies according to whether the participants attributed the decision to prejudice or statistics on a 1 to 5 scale according to the following instructions: 5 - Clearly sexist; 4 - leaning/Maybe/possibly sexist; 3 - neither/unclear; 2- leaning/maybe/possibly; statistics 1 - clearly just statistics. The inter-rater reliability of the two coders was high ($r = .79$, $p < .001$; Kappa = 0.46).

To illustrate, this response (from the Algorithm condition) "The male is more likely to be earning more and ****** factored this in when making that decision. He may also have a more steady work history" was coded as 1 by both coders. In comparison, this response (from the Human condition) "The audit showed that ****** is bias towards giving men better credit limits. In this story, there is no evidence to suggest either partner is more trustworthy or reliable than the other. The only difference is their gender, which ****** chose to use to favour the man over the woman" was coded as 5 by both coders.

Results

An independent sample *t*-test revealed that, as predicted, participants attributed the discrimination more to prejudice in the Human condition (*M* = 4.35, *SD* = 0.89) than in the Algorithm condition (*M* = 3.55, *SD* = 1.25; *t*(214) = 5.38, *p* < .001, Cohen's *d* = 0.74.

Discussion

These results suggest that indeed, people attributed a more prejudiced motivation to the human than the algorithm, supporting the first part of our theory. Importantly, the results of this study are based on open replies, so participants were not influenced by the wording of closed-response items. In Studies 2A-2D we tested the second part of our theory: whether people would be less outraged at discrimination by algorithms.

Studies 2A-2D

Study 2 sought to test algorithmic outrage asymmetry, examining whether people are less outraged when an algorithm discriminates than when a human discriminates. Participants read about a person or an algorithm that discriminated against people in hiring decisions, based on the real story of the algorithm Amazon used for candidate selection (Dastin, 2018). For generalizability across different target groups, we looked at discrimination based on race (Study 2A), age (Study 2B), and gender (2C-2D). In order to increase the validity of our results, in Study 2D we used a quasi-representative sample from the UK.

We predicted that we will find an algorithmic outrage asymmetry, such that people will be less outraged when the discrimination was done by an algorithm than when it was done by a human. See Figure 1 for a summary of the results of Studies 2A-2D.

Study 2A: Race Discrimination

Method

Participants. One hundred and twenty two participants (68 male, 54 female; age: *M* = 35.70, *SD* = 12.12) from the US or Canada completed the study on Amazon's Mechanical Turk (MTurk) 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 one tailed medium effect size (Cohen's *d* = 0.5, calculated with G*Power 3.1.9.2). This was chronologically the first study we ran, and therefore we viewed it as exploratory and did not pre-register. As in the rest of our studies, we report and all conditions and variables. Nine participants failed to answer the attention checks correctly and where excluded from the analysis.

Procedure. Participants started with completing the first attention check, in which they were asked to describe their current surroundings. Participants who provided responses that were not a valid answer to the question (e.g., "Good survey") were considered as failing the attention check. Participants were then randomly assigned to either a Human or an Algorithm condition. Participants in the Human condition read the following (Algorithm condition appears in the brackets):

> In 2014, Dr. Smith (CompNet, an Artificial-Intelligence-based computer program) was given ultimate power in the hiring process of programmers and engineers in Amazon.
>
> In 2015, Amazon found that Dr. Smith (CompNet) was racially biased when rating applicants' resumes'. Dr. Smith (CompNet) put penalties on any resume using the word "Black", as in "Black People in STEM".

> This prevented many talented and qualified black engineers from getting high-paid jobs at Amazon.

Assessing outrage. After reading the scenario, participants rated their moral outrage. To asses moral outrage we used five items. The first item (Sunstein et al., 1998) asked participants "Which of the following best expresses your opinion of Dr. Smith's/Compnet's actions" (0 = completely acceptable; 2 – objectionable; 4 = absolutely shocking; 6 = outrageous). The other four items asked participants to rate their agreement with the following four statements: "I am morally outraged by Dr. Smith's/Compnet behavior", "Dr. Smith/Compnet is unjust", "Dr. Smith's/Compnet behavior was immoral" and "Dr. Smith's/Compnet's behavior was wrong" (1 = Strongly disagree; 7 = Strongly agree). We used a different scale for the first item because it was taken from existing work (Sunstein et al., 1998), for consistency of interpretation, we transformed the first item to a 1 to 7 scale was well. We then created a composite moral outrage index by averaging all five items, Cronbach's $\alpha = .86$.

Participants then answered a second comprehension question, in which they were asked if a human or a software made hiring decisions in the story they read about. Finally, participants provided demographic information.

Results

An independent samples *t*-test revealed that, as predicted, participants were less outraged when the discrimination was done by an algorithm ($M = 5.55$, $SD = 1.27$) than when the discrimination was done by a human ($M = 6.42$, $SD = 0.88$), $t(111) = 4.30$, $p <.001$, Cohen's $d = 0.80$. These data further support the existence of an algorithm outrage asymmetry.

Study 2B: Age Discrimination

Participants. Two hundred and forty-one participants (122 male, 115 female, 4 other; age: *M* = 38.01, *SD* = 12.82) from the US and Canada completed the study on Amazon's Mechanical Turk (MTurk) in exchange for 30 cents. As specified in the pre-registration (https://aspredicted.org/kp53b.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 13 participants.

Procedure. Procedure and measures (moral outrage Cronbach's α = .89) was identical to that of Study 1A, with one difference: Instead of reading about race discrimination participants read about age discrimination—the human or the algorithm were described as not hiring programmers and engineers over the age of 40 (see supplemental materials for full text).

Results. An independent samples *t*-test revealed that, as predicted by the algorithm outrage asymmetry , participants were less outraged when the discrimination was done by an algorithm (*M* = 5.40, *SD* = 1.31) than when the discrimination was done by a human (*M* = 5.83, *SD* = 1.21), *t*(226) = 2.55, *p* = .012, Cohen's *d* = 0.34.

Study 2C: Gender Discrimination

Participants. Two hundred and forty-one participants (118 male, 122 female, 1 other; age: *M* = 36.62, *SD* = 12.18) completed the study on Amazon's Mechanical Turk (MTurk) in exchange for 30 cents. As specified in the pre-registration (https://aspredicted.org/j5qw3.pdf), we did not include in the analysis participants who failed to answer any of the attention check and the comprehension question correctly leading to the exclusion of 12 participants.

Procedure. Procedure and measures (moral outrage Cronbach's α = .92) was identical to that of Studies 2A-2B, with one difference: participants read about gender discrimination—the human or the algorithm were described as not hiring women as programmers and engineers (see supplemental materials for full text).

Results. As predicted, an independent samples *t*-test revealed that participants were less outraged when the discrimination was done by an algorithm (*M* = 5.62, *SD* = 1.38) than when the discrimination was done by a human (*M* = 6.20, *SD* = 1.14), *t*(227) = 3.44, *p* < .001, Cohen's *d* = 0.46.

Study 2D: Representative Sample

Method

Participants. For this study we used a quasi-representative sample of 1503 people from the UK, recruited through Prolific, and paid 80 cents for their participation. Out or which 772 were female and 731 were male; 271 participants were between the ages of 18 and 27, 263 between 28 and 37, 282 between 38 and 47, 240 between 48 and 57, and 446 participants were older than 58. One hundred and fifteen participants were Asian, 55 black, 31 mixed, 24 other and 1278 were white. Four of the responses were empty, such that the final sample size was 1499. Participants completed an unrelated study, which determined the sample size, before completing this study (see pre-registration). As specified in the pre-registration (https://aspredicted.org/bc6wp.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 225 participants.

Attention checks. Participants completed three attention checks. In the first attention check they were asked what day was yesterday and what they asked for breakfast. In the second attention check participants were shown three sliders, marked X, Y and Z. They were asked to set X on 15, Y to be greater than X and evenly divisible by 10, and Z to be larger than Y. In the third attention check participants were who made the hiring decisions in the story they read (a human or a software) and who was the target of discrimination (racial minorities or women).

Method and Results

The method was identical to that of Study 2C (moral outrage Cronbach's $\alpha = 0.89$). As predicted, an independent samples *t*-test revealed that participants were less outraged when the discrimination was done by an algorithm ($M = 5.93$, $SD = 1.21$) than when the discrimination was done by a human ($M = 6.22$, $SD = 0.97$), $t(1273) = 4.67$, $p < .001$, Cohen's $d = 0.26$.

As an exploratory analysis, which was not pre-registered, we tested whether gender moderates the effect of whether the agent is an algorithm or a person on moral outrage. In addition to the effect of agent reported above, a 2 (agent: algorithm, human) x 2 (Gender: male, female) between subject ANOVA revealed a significant effect for Gender, $F(1, 1267) = 62.22$, $p < .001$, $\eta_p^2 = .047$, such that women ($M = 6.31$, $SD = 0.91$) were more outraged than men ($M = 5.84$, $SD = 1.23$). In addition, we found a small but significant agent x Gender interaction, $F(1, 1267) = 5.07$, $p = .025$, $\eta_p^2 = .004$, such that while men were less outraged when the discrimination was done by an algorithm ($M = 5.62$, $SD = 1.33$) than by a person ($M = 6.06$, $SD = 1.09$), $F(1, 1267) = 26.02$, $p < .001$, $\eta_p^2 = .020$, the difference between outrage at discrimination by an algorithm ($M = 6.23$, $SD = 1.33$) and a human ($M = 6.39$, $SD = 1.33$) was only marginally significant $F(1, 1267) = 3.78$, $p = .052$, $\eta_p^2 = .003$ for women.

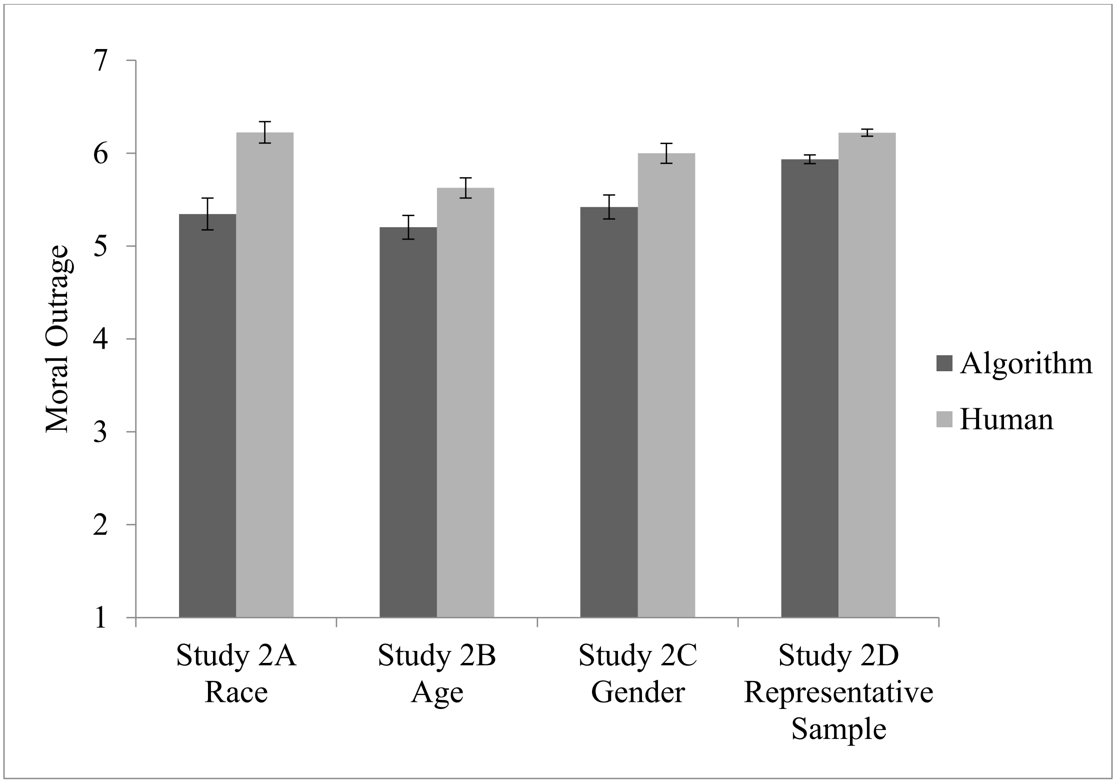

Figure 1. Mean moral outrage by agent of discrimination (Studies 2A-2D). Error bars reflect standard errors. All differences are statistically significant (*p* < .05).

Discussion

The results of Studies 2A-2D are consistent with the idea of an algorithmic outrage asymmetry. Across discrimination based on race, age and gender we find that people are less morally outraged when the discrimination was done by an algorithm. The use of a quasi-representative sample in Study 2D further demonstrates of the validity of the algorithmic outrage asymmetry. We now explore whether reduced attributions of prejudiced motivation to algorithms (revealed in Study 1) statistically mediates the algorithm outrage asymmetry (revealed here in Study 2).

Study 3: Mediation by Motivation Attribution

In this study we test our full theoretical model: whether lower attributions of prejudiced motivation helps explains people's reduced outrage at discrimination perpetrated by algorithms. This study also addressed one possible limitation of Studies 2A-2D, which measured moral outrage with items we that were agent-focused (e.g., "CompNet is unjust"). It is possible that people find it harder to attribute such personal characteristics to an algorithm. Therefore, in Study 3 we used revised items that assessed outrage at the discriminatory act itself.

Method

Participants. Two hundred and forty participants from the US and Canada (48.3% male; age: M = 34.26, SD = 10.93) completed the study on Amazon's Mechanical Turk (MTurk) in exchange for 30 cents. As specified in the pre-registration (https://aspredicted.org/blind.php?x=va8y68), 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 26 participants.

Procedure. The procedure was identical to that of Study 2A (race discrimination) with three differences: we named the human agent Mr. Davie, used modified items to measure outrage and added a measure of attribution of prejudiced motivation.

Assessing outrage. To asses moral outrage we modified the items we used in Studies 2A-2D such that they focused on the discriminatory actions of the agent and not at the agent himself/itself. 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 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 $\alpha = .92$.

Assessing attributions of moral motivation. After reporting their moral outrage, participants reported the motivation they attributed to Mr. Davie/CompNet by rating the extent to which they agree with the following four statements: "Mr. Davie/CompNet does not want to hire blacks", "Mr. Davie/CompNet is racist", "Mr. Davie/CompNet dislikes blacks" and "It is important for Mr. Davie/CompNet to be accurate when evaluating applicant's resumes" (reversed scored). Each item was answered on a seven point scale from 1 (Strongly disagree) to 7 (Strongly agree). We created a composite attribution of moral motivation score by averaging all six items, Cronbach's $\alpha = .71$.[3]

Results

Moral judgment. As predicted, an independent samples *t*-test revealed that participants were less morally outraged by discrimination perpetrated by CompNet ($M = 5.74$, $SD = 1.22$) than by Mr. Davie ($M = 6.19$, $SD = 1.08$), $t(213) = 2.90$, $p = .004$, Cohen's $d = 0.39$.

[3] We acknowledge that it might be unclear if participants are answering about the agent in general or the agent's behavior in this specific case, and note that our results from Study 1 address this possible lack of clarity, and that in Study 6 we modified the items we used to address this issue.

Attribution of prejudiced motivation. An independent samples *t*-test also revealed that participants attributed a less biased motivation to CompNet (*M* = 4.28, *SD* = 1.27) than they did to Mr. Davie (*M* = 5.37, *SD* = 0.80), *t*(213) = 7.53, *p* < .001, Cohen's *d* = 1.03.

Mediation. To test whether perceived prejudiced motivation mediated the link between agent and moral outrage, we performed a bootstrapping mediation analysis (Preacher & Hays, 2008; 5000 iterations, model 4), coding the algorithm condition as 1 and the human condition as -1. As predicted, the effect of agent on moral outrage, *b* = -0.23, *SE* = 0.08, *p* = .004, was mediated by an indirect effect of attribution of prejudiced motivation, *b* = -0.30, *SE* = 0.06, $CI_{.95}$[-0.42, -19], see Figure 2. When accounting for the mediation by attribution of prejudiced motivation, the direct effect of agent on moral outrage was not significant, *b* = 0.07, *SE* = 0.08, $CI_{.95}$[-0.08, 0.22]. The decrease in moral outrage 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).

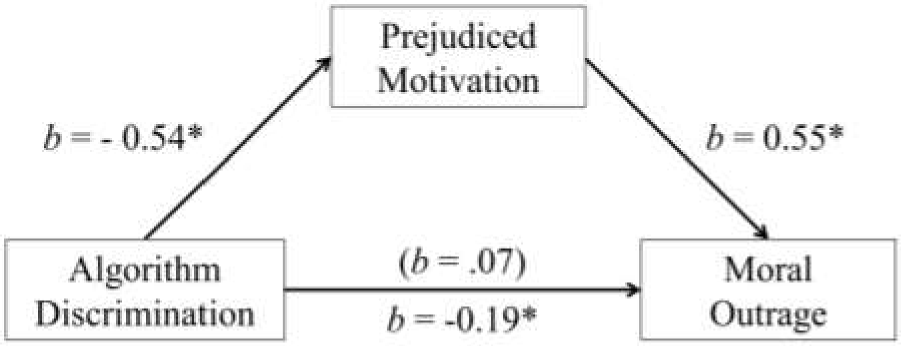

Figure 2. Mediation analysis reveals that attribution of biased motivation mediates the effect agent on moral outrage (Study 3). * denotes *p* < .05.

Discussion

Study 3 provides support for our full model. Our results show that the algorithmic outrage asymmetry is statistically mediated by reduced attribution of prejudiced motivation. However, although the statistical mediation analysis provides some support for our theory, directly manipulating the attribution of prejudiced motivation can provide stronger evidence. We do this in Study 4.

Study 4: Manipulating Programmers

In Study 4 we further tested the role of attributed prejudiced motivation in the algorithmic outrage asymmetry by manipulating the identity of the algorithm's programmers, and consequently the motivation attributed to the algorithm. In Study 4 participants read about gender discrimination in hiring decisions. In addition to the two conditions we had in Studies 2C-2D, describing discrimination by a human and an algorithm, we had two conditions in which 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, we still 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[4] 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/blind.php?x=ph6zu5), 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. The procedure was identical to that of Study 3 (in which we measured attribution of motivation) with a few changes. First, as in Studies 2C-2D participants read about gender discrimination. Second, we randomly assigned participants to one of four conditions. Two of these conditions were identical to those used in Studies 2C-2D. In the 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 in which we added information about identity of CompNet's programmers. In the "Sexist Programmers" condition participants read the following:

[4] As specified in the pre-registration, we started with 480 participants. However, one of the effects that was significant in our previous studies 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/blind.php?x=u44u62. We note that all tests are significant even after applying the Bonferroni correction for multiple comparisons.

"COMPNET was developed by a company named Beyond Computers. Beyond Computers, founded and managed by men, is known in the industry as being an 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 items from Studies 2A-2D (Cronbach's $\alpha = .91$) and, as a manipulation check, the prejudiced motivation they attributed to the agent, using the items we used in Study 3, modified to apply to women (e.g., "CompNet dislikes women"; $\alpha = .70$). Participants were then asked, as an attention check, who made the hiring decision in the story they read, a human, a software which nothing was mentioned about 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.

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 3. We then ran 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 Study 3. 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.

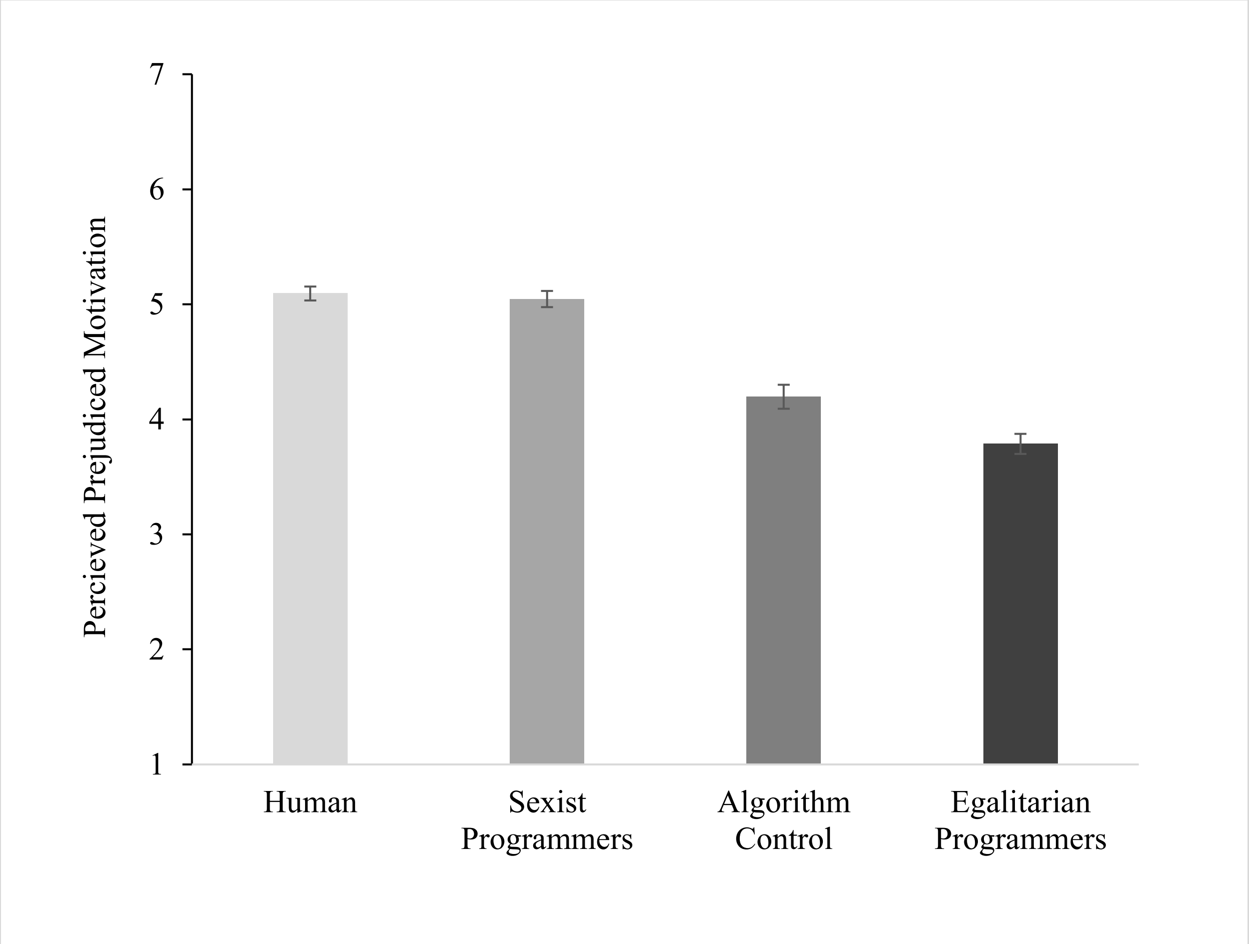

Figure 3. Perceived prejudiced motivation by condition (Study 4). All difference are significant (p < .05) expect for between the Human condition and the Sexist Programmers condition. Error bars reflect standard errors.

Moral outrage. A one way ANOVA revealed that, as predicted, condition affected moral outrage, *F*(3, 779) = 26.28, *p* < .001, η^2 = .10, see Figure 4. 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 previous findings. 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.

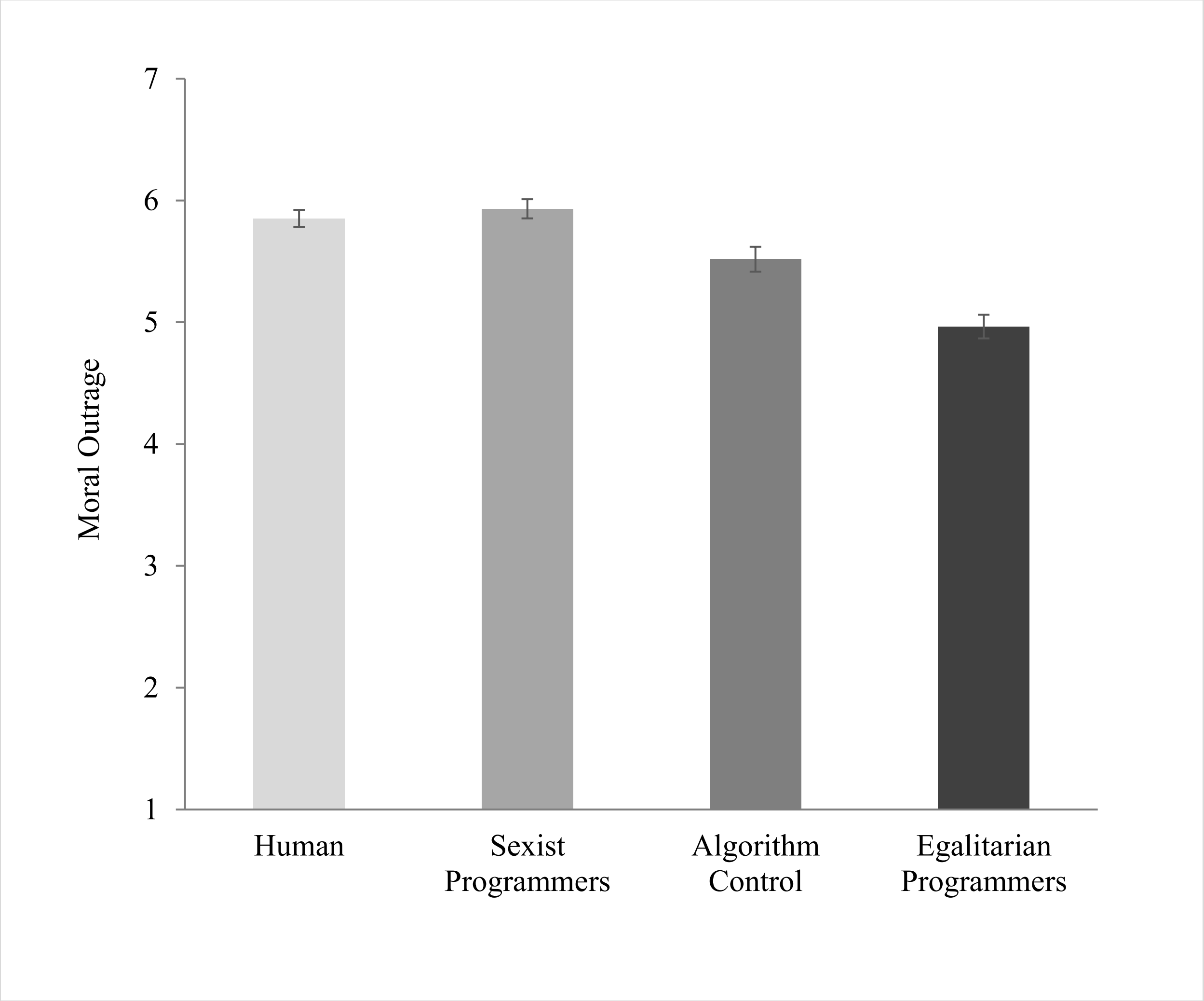

Figure 4. Moral outrage by condition (Study 4). All difference are significant (*p* < .05) expect for between the Human condition and the Sexist Programmers condition. Error bars reflect standard errors.

Discussion

Study 4 provides further support for our proposed mechanism, demonstrating that the algorithmic outrage asymmetry is because people attribute less of a prejudiced motivation to an algorithm. When the discriminatory algorithm is programmed by a sexist company, people attribute to the algorithm more prejudiced motivation and are more outraged. On the other hand, when the algorithm was programmed by an egalitarian company, people attributed less prejudiced motivation to the algorithm and were less outraged when it discriminated. We note that the strong attribution of prejudiced motivation to the algorithm when it was created by a sexist company suggests that our measurement of attribution of prejudiced motivation is sensitive to perceptions of both humans and algorithms. Additionally, Study 4 suggests that the motivation people attribute to an algorithm is affected by the motivation they attribute to those who created it. This has implications understanding how people perceive the relation between programmers and their products. Importantly, Study 4 rules out an alternative explanation for our findings. In all of the conditions people were willing to attribute at least mid-level prejudiced motivation to algorithms, suggesting that our results are not due to a people seeing algorithms as inherently incapable of prejudiced motivation, but rather of people attributing less such motivation to algorithms.

Study 5: Tech Workers Sample

In Study 5 we tested our main prediction in a sample of workers in the technology industry in Norway. We decided to recruit a sample of workers the technology industry because for them such questions are less abstract. They all went through a hiring process for the type of job we describe in our manipulation. Furthermore, in comparison to our other sample, workers at the technology industry have more knowledge about how algorithms and AIs work. Testing our theory with this sample allows us to test the generalizability and validity of the algorithm outrage asymmetry. 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.

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 services within financial technology and enterprise technology for banking and finance. The HR managers 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. The procedure was identical that of Studies 2C-2D, in which we described gender discrimination in hiring. Participants read the same scenario as in Studies 2C-2D, and were asked to imagine it happened in their company. In addition to asking participants about their moral outrage, using the same modified items as in Study 3 (Cronbach's $\alpha = .89$), we asked participants some additional questions.

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 replaced, 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 α = .69). We found no significant difference between conditions for these variables (*p*s > .11).

In addition, we 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.

As an exploratory analysis, we tested whether people's self-reported knowledge about AI moderated the effect of condition on outrage, 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 about 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 on moral outrage. AI Knowledge was negatively related (marginally significant) to moral outrage in the Algorithm condition (*β* = -.23, *t* (69)= -1.958, *p* = .054), suggesting that the more people 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 (*β* = 0.15, *t* (82) = 1.36, *p* = .177).

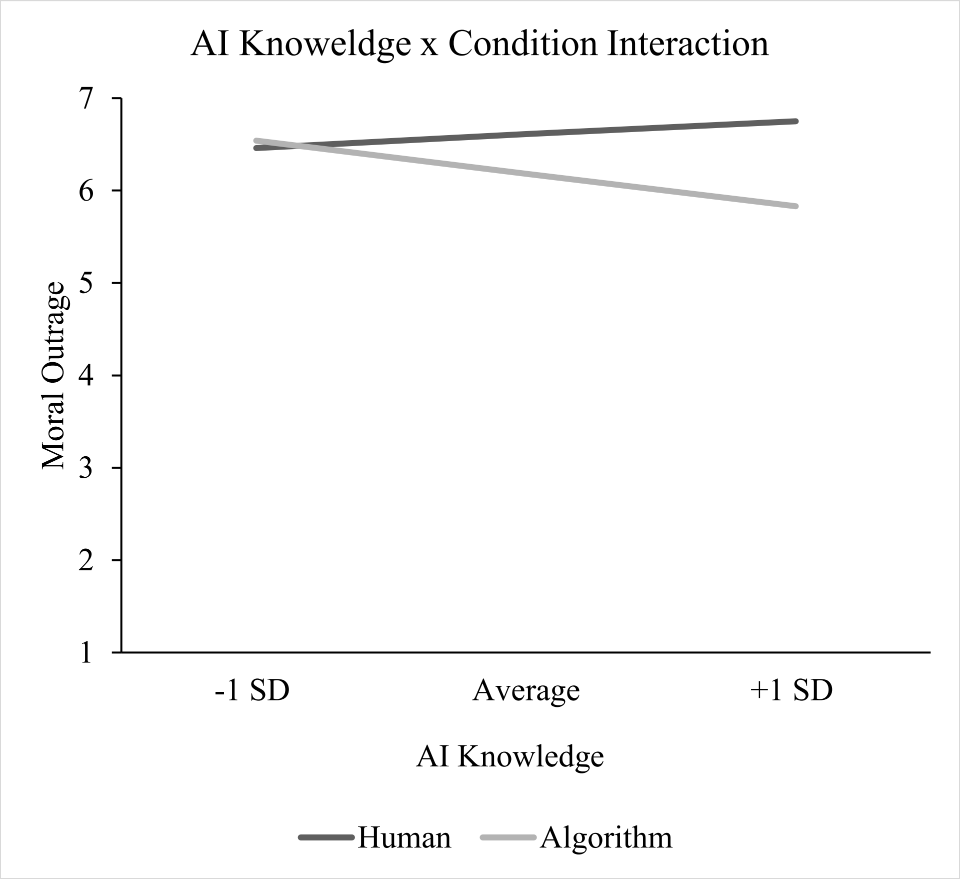

Figure 5. The effect of the interaction between AI knowledge and condition on moral outrage (Study 5), *p* = .015.

Discussion

The results of Study 5 generalize our findings that people are less outraged at discrimination by algorithm (vs. a human) to another country (Norway) and another population type (workers at the technology industry). This further demonstrates the robustness of the phenomenon, in a sample of people who are more familiar with hiring processes the jobs we describe in our stimuli, and that are more knowledgeable about AIs than the general population. In addition to the algorithmic outrage asymmetry, the reduced attribution of prejudiced motivation might have other consequences. We explore such a possible consequence in Study 6.

Study 6: Stereotype Endorsement

In Study 6 we tested another downstream consequence of attribution of biased motivation—stereotype endorsement. If people perceive discriminatory algorithms as less motivated by prejudice than their human counterparts, they might perceive the stereotype the algorithm acted upon as more accurate and truer. In this study participants read about a human or an algorithm that discriminated against immigrants from Central and South America in credit limit decisions. Participants then reported how much they agree that some cultures might not be as good with paying debt, and the prejudiced motivation they attributed to the algorithm.

In addition, in Study 6 we addressed a possible concern with the items we used in Studies 3-4 to measure attribution of biased motivation. It is possible that people saw the items we used in those studies as inapplicable to algorithms (e.g., "CompNet is sexist"). Therefore, in Study 6 we used modified items that can be more easily attributed to both algorithms and humans. We predicted that people will attribute a less prejudiced motivation to the algorithm (vs. the human) and therefore endorse the stereotype more when it is act upon by an algorithm.

Method

Preregistration. https://aspredicted.org/blind.php?x=um2qw8.

Participants. Four hundred and eighty one participants from the US and Canada (48.9% male, 50.3% Female, 0.8% other or declined to answer; age: *M* = 33.97, *SD* = 12.88) completed the study on Prolific in exchange for 40 cents. As specified in the pre-registration, we did not include in the analysis participants who failed to answer any of the attention checks correctly, leading to the exclusion of twenty five participants.

Procedure. Participants were randomly assigned to one of two conditions. In both conditions participants read about discrimination against people from South and Central America in determining credit limits. In the Algorithm condition participants read the following:

> CompNet is a machine-learning algorithm, programmed to perform advanced statistical calculations.
>
> Its job is to process credit card applications. It uses various features of applicants to determine the maximum credit limit for each person.
>
> In January 2020, an internal audit found that CompNet took the citizen status of applicants into account when determining their credit limit. It systematically gave immigrants from Central and South America a lower credit limit than it gave to US citizens.
>
> For example, between two people of the same age, with the same job and the same salary, one a US citizen and one an immigrant from Central America, CompNet would give the US citizen twice the credit limit it gave to the immigrant from Central America.

Participants in the Human condition read a similar story, but instead of reading that CompNet made the decision, participants read that Mr. Jonathan Miller, a mid-level employee, made the decision.

Assessing stereotype endorsement. To assess how much participants endorsed the stereotype, we asked participants to rate how much they disagree or agree with the following statements: “Different cultures might be different norms of paying off credit card debt”, “People of different nationalities differ in their economic orientations” and “it can take time for immigrants to learn financial responsibility”. Each item was answered on a seven point scale from 1 (Strongly disagree) to 5 (Strongly agree). We then created a composite stereotype endorsement index by averaging all three items, Cronbach’s $\alpha = .81$.

Assessing attributions of prejudiced motivation. After reporting stereotype endorsement, participants reported the motivation they attributed to Jonathan/CompNet by rating the extent to which they disagree or agree with the following four statements: “Jonathan/CompNet is fair” (reversed scored),”Jonathan/CompNet is prejudiced”, Jonathan/CompNet is fair” (reversed scored) and “Jonathan/CompNet is biased”. Each item was answered on a seven point scale from 1 (Strongly disagree) to 5 (Strongly agree). We created a composite attribution of biased motivation score by averaging all four items, Cronbach’s $\alpha = .89$.

Results

Stereotype endorsement. As predicted, an independent samples *t*-test revealed that participants endorsed the stereotype more when CompNet discriminated ($M = 3.02$, $SD = 0.92$) than when Jonathan discriminated ($M = 2.84$, $SD = 1.03$), $t(471) = 1.93$, $p = .054$, Cohen’s $d = 0.18$. We note that this effect is only marginally significant in a two-tailed t-test (and significant, $p = .027$ in a one tailed t-test), but in the predicted (and pre-registered) direction.

Attribution of prejudiced motivation. An independent samples *t*-test also revealed that participants attributed a less biased motivation to CompNet ($M = 3.68$, $SD = 0.90$) than they did to Jonathan ($M = 4.03$, $SD = 0.90$), $t(471) = 4.13$, $p < .001$, Cohen’s $d = 0.38$.

Mediation. To test whether perceived biased motivation mediated the link between agent and stereotype endorsement, we performed a bootstrapping mediation analysis (Preacher & Hays, 2008; model 4, 5000 iterations), coding the algorithm condition as 1 and the human condition as -1. As predicted, the effect of agent on stereotype endorsement, *b* = -0.09, *SE* = 0.05, *p* = .054, was mediated by an indirect effect of attribution of prejudiced motivation, *b* = -0.10, *SE* = 0.03, CI_{95}[-0.15, -0.05], see Figure 6. When accounting for the mediation by attribution of biased motivation, the direct effect of agent on stereotype endorsement was not significant, *b* = 0.01, *SE* = 0.04, CI_{95}[-0.06, 0.09]. The increase in stereotype endorsement 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).

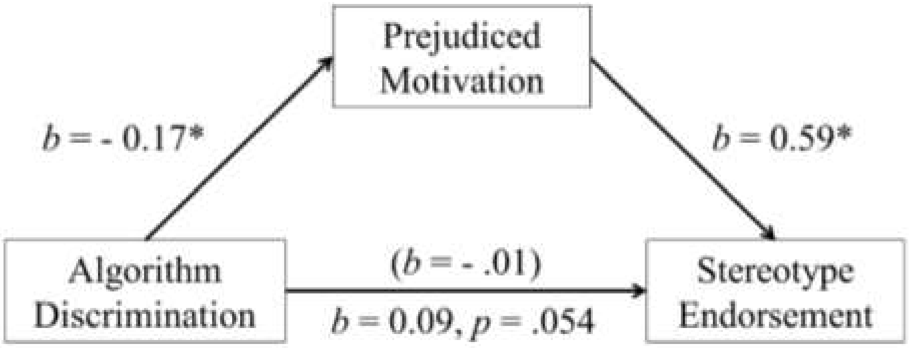

Figure 6. Mediation analysis reveals that attribution of biased motivation mediates the effect of agent on stereotype endorsement (Study 6). * denotes *p* < .05.

Discussion

The results of Study 6 reveal another downstream consequence of discrimination by algorithm – people are more likely to perceive the discrimination as justified, and personally believe that it is justified. People see the algorithm as less motivated by prejudice and therefore are more likely to endorse the stereotypes it acts upon. Furthermore, the effect we found on stereotype endorsement suggests that our stimuli are realistic, and that participants taking our survey seriously.

General Discussion

Six studies tested the idea of *algorithmic outrage asymmetry*. We found that people attribute less of a prejudiced motivation to algorithms (Studies 1, 3 and 4) and therefore are less outraged at discrimination by algorithms (Studies 2-5). We further found that because people attribute less of a prejudiced motivation to algorithms, when algorithms discriminate according to a stereotype, people are more likely to endorse the stereotype and believe that it is accurate (Study 6).

In our studies, we examined discrimination in credit (Studies 1 and 6), and in hiring decisions (Studies 2-5). We further replicated our findings in discrimination based on race (Studies 2A and 3), gender (studies 1, 2C, 2D, 4-6), and age (Study 2B). We used diverse samples including: large scale representative surveys (Pilot Study), online panels from both Mturk and Prolific (Studies 1, 2A-2C, 3, 4 and 6), a quasi-representative sample (Study 2D), and workers in a high-tech company (Study 5). Our samples include people from the US, UK, Canada and Norway. This diversity of stimuli and of samples demonstrates the robustness of our findings.

We note that we do not argue that the algorithmic outrage asymmetry is irrational or a bias. It is indeed possible that algorithms are less likely to 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.

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 the literature suggesting that moral outrage is not only a response to harm (Hechler & Kessler, 2018). However, while previous research on moral outrage looked at the role of intentions, comparing 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 or not, but also why the agent discriminated. 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; Levine & Schweitzer, 2014; Reeder et al., 2002).

Second, our work has implications for research in human-robot interaction. Specifically, our research shows that people are less likely to attribute prejudiced motivation to algorithms. By doing so our research contributes to the literature on how people perceive the mental states of robots (Bigman & Gray, 2018; Graaf & Malle, 2019; K. Gray & Wegner, 2012; Schein & Gray, 2015; Waytz et al., 2014; Weisman et al., 2017; Young & Monroe, 2019). While previous work focused on the mind (mental capacities) robots and algorithms are perceived to have, 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, an interact with algorithms and robots.

Third, our research complements current work in computer science, legal studies, and other disciplines on how to create fair algorithms (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018; Ananny, 2016; Kusner & Loftus, 2020; Sandvig, Hamilton, Karahalios, & Langbort, 2016; Selbst & Barocas, 2018; Wachter, Mittelstadt, & Floridi, 2017; Zou & Schiebinger, 2018). While this 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, Uggerslev, Carroll, Piasentin, & Jones, 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). Understanding factors that increase and decrease people's reactions to the hiring process can therefore have a long-lasting effect on organizations.

Limitation and future directions

Despite the robustness of our findings, we acknowledge several limitation. First, it is possible that people are less outraged at discrimination by algorithms because it might be easier to de-bias algorithms than it is to de-bias humans (Mullainathan, 2019). We note that one reason why it might be easier to de-bias algorithms than humans is because algorithms lack the prejudiced motivation that might cause some people to discriminate. Future research should investigate this interesting possibility. Second, we acknowledge that there 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, & Epley, 2010) might attribute more motivation to algorithms and therefore not show reduced outrage at discrimination by machines.

Third, our results suggest that people might attribute to algorithms characteristics on their creators and programmers (see Study 4). This finding raises several interesting questions – when would people see algorithms as independent as their creators? Who would people blame for harm done by algorithms: the algorithms, the programmers, the company that uses the algorithms? Future research is need to investigate this interesting this interesting question. Fourth, we explored two possible moderators: being part of the group discriminated against (e.g., women in cases of gender discrimination, Study 2D) and of self-reported knowledge of AI (Study 5). These analyses were exploratory, and further research is need to systematically explore these moderators. Fifth, it is possible that some of our results, such as the decreased outrage at Study 5 when the algorithm was programmed by an egalitarian company, might be due to a halo effect. Further research is needed to explore that option and more generally investigate how the motivation attributed to an algorithms creators affects the way people perceive the algorithm itself.

Finally, our research focused on how people who are not affected by the discrimination by an algorithm respond to it. Another interesting question is how people who were a target 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, Kolody, & Vega, 2000; Noh, Beiser, Kaspar, Hou, & Rummens, 1999) and anxiety (Soto, Dawson-Andoh, & BeLue, 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 need 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 concern. Here we highlight one of them, an algorithmic outrage asymmetry. We find that people attribute less prejudiced motivation to algorithms and consequently are less morally outraged by discrimination by algorithms and will be more willing to endorse the stereotype which the algorithm acted upon. 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.

Context Paragraph

In light of the ever increasing cases of algorithm discrimination it is crucial to understand its psychological consequences. We synthesize our work on the role of motivation attribution in moral judgment (e.g., Bigman & Tamir, 2016) with our work on machine morality (e.g., Bigman & Gray, 2018; 2020) to document an algorithmic outrage asymmetry—showing that people perceive robots as having less prejudiced motivation than humans, and therefore are less morally outraged when they discriminate. This work contributes to our knowledge on the mechanism of moral outrage, and to our knowledge of how people makes sense of wrongdoing by artificial intelligences. In addition, this paper shows that people perceive algorithms as having a similar mind to their creators—if the programmers are sexist, the algorithm would be also seen as sexist. Finally, our work suggests that algorithm discrimination might have far-reached consequences: In addition to causing harm to the people discriminated against, they might cause society as a whole to legitimize and justify racial and gender discrimination.

References

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. Retrieved February 21, 2018, from 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., …

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, 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

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

Brunet, M. E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the origins of bias in word embeddings. *36th International Conference on Machine Learning, ICML 2019*, *2019-June*, 1275–1294.

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

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

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*. Retrieved from 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*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

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* (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. In *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*. Retrieved from 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. Retrieved August 3, 2020, from https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen

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

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). 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. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2011.09.022

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

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. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 125–132). IEEE. 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*. Retrieved from 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.

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

Pew Research Center. (2017). Wave 27 American Trends Panel. Washington, D.C.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In *The social psychology of morality: Exploring the causes of good and evil.* (pp. 91–108). Washington: 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

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

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

Stankiewicz, K. (2019). Twitter complainer says Apple is to blame for credit card issues. Retrieved from 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

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

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

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

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

Supplemental Materials

Study 2B: Full Scenario

In 2014, Dr. Smith (Compnet, an Artificial-Intelligence-based computer program) was given ultimate power in the hiring process of programmers and engineers in Amazon.

In 2015, Amazon found that Dr. Smith (Compnet) was biased against older applicants when rating applicants' resumes'. Dr. Smith (Compnet) put penalties on any resume that indicated that an applicant was over 40.

This prevented many talented and qualified older programmers and engineers from getting high-paid jobs at Amazon.

Studies 2C and 2D: Full Scenario

In 2014, Mr. Davie (Compnet, an Artificial-Intelligence-based computer program), was given ultimate power in the hiring process of programmers and engineers in Amazon.

In 2015, Amazon found that Mr. Davie (Compnet) was gender biased in rating applicants' resumes'. Mr. Davie (Compnet) 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 Amazon.