American Psychological Association

Journal of Experimental Psychology: Applied

ISSN: 1076-898X

2021, Vol. 27, No. 2, 447–459
https://doi.org/10.1037/xap0000294

Can Algorithms Legitimize Discrimination?

Andrea Bonezzi[1] and Massimiliano Ostinelli[2]
[1] Stern School of Business, New York University
[2] College of Business Administration, Winthrop University

Algorithms have been the subject of a heated debate regarding their potential to yield biased decisions. Prior research has focused on documenting algorithmic bias and discussing its origins from a technical standpoint. We look at algorithmic bias from a psychological perspective, raising a fundamental question that has received little attention: are people more or less likely to perceive decisions that yield disparities as biased, when such decisions stem from algorithms as opposed to humans? We find that algorithmic decisions that yield gender or racial disparities are less likely to be perceived as biased than human decisions. This occurs because people believe that algorithms, unlike humans, decontextualize decision-making by neglecting individual characteristics and blindly applying rules and procedures irrespective of whom they are judging. In situations that entail the potential for discrimination, this belief leads people to think that algorithms are more likely than humans to treat everyone equally, thus less likely to yield biased decisions. This asymmetrical perception of bias, which occurs both in the general population and among members of stigmatized groups, leads people to endorse stereotypical beliefs that fuel discrimination and reduces their willingness to act against potentially discriminatory outcomes.

Public Significance Statement
This research suggests that replacing human with algorithmic decision-making might contribute to legitimize discrimination. In situations that entail the potential for discrimination, algorithmic decisions that yield disparities are less likely than human decisions to be perceived as biased. The presumed objectivity of algorithms might foster stereotypical beliefs about stigmatized groups and make people less likely to take action against disparities that could be discriminatory.

Keywords: algorithmic bias, discrimination, disparities, inequality, algorithm aversion

Supplemental materials: https://doi.org/10.1037/xap0000294.supp

Imagine that a white man and a black man are arrested for stealing a car. After setting the defendants' court dates, a judge needs to decide whether they should be detained while awaiting trial. To make this decision, the judge assesses the risk that the two men will reoffend if released pre-trial. The defendants committed the same crime and have similar criminal histories. Yet, the judge deems the black man at higher risk to reoffend than the white man, so decides to detain the black man and release the white man. Most people would perceive such decision as biased. Now take a step back. Imagine that an algorithm, rather than a judge, had appraised the black defendant at higher risk to reoffend than the white defendant, dooming the black man to be detained and the white man to be released. Would people still perceive the decision as biased?

This article was published Online First March 22, 2021.

Andrea Bonezzi https://orcid.org/0000-0002-5624-8406
Massimiliano Ostinelli https://orcid.org/0000-0002-9515-9131

Data and materials for all studies are available on OSF: https://osf.io/276gm/

Correspondence concerning this article should be addressed to Andrea Bonezzi, Stern School of Business, New York University, 40 West 4th Street, New York, NY 10012, United States. Email: abonezzi@stern.nyu.edu

Algorithms are increasingly being used to make decisions that profoundly impact people's lives, such as who gets incarcerated (Dressel & Farid, 2018), hired (Bogen & Rieke, 2018), admitted to universities (Pangburn, 2019), provided access to healthcare (Bates et al., 2014), and extended financial credit (Gomber et al., 2018). The use of algorithms is predicated on the premise that statistical models can be more accurate and efficient than human decision-makers (Dawes et al., 1989; Grove et al., 2000; Meehl, 1954). Yet, recent evidence suggests that algorithms can also be biased, just like the human counterparts they are meant to replace (e.g., Eubanks, 2018; Noble, 2018; O'Neil, 2016; Schwemmer et al., 2020).

In its most general sense, algorithmic bias refers to systematic error. An algorithm is considered biased if it incorrectly assigns better outcomes to certain individuals or groups of individuals over others, yielding systematic disparities that have no valid grounds (Friedman & Nissenbaum, 1996). As such, the term algorithmic bias can have different connotations. Algorithmic bias assumes a discriminatory connotation when it yields systematic disparities associated with variables that are protected by law, such as gender or race. To illustrate, an algorithm used to screen job applicants was found to be biased because it consistently assigned higher employability scores to men over women with comparable qualifications (Dastin, 2018). Similarly, an algorithm used in the criminal justice system to assess defendants' risk of recidivism was considered

biased because it consistently classified black defendants at higher risk than they actually were, and white defendants at lower risk than they actually were (Angwin et al., 2016). Analogous instances of algorithmic bias have been documented in domains such as healthcare (Obermeyer et al., 2019), education (Schwartz, 2019), and credit lending (Bartlett et al., 2019).

Algorithmic bias, however, ought not necessarily imply discrimination. An algorithm can also be considered biased if it makes systematic errors that yield disparities that are not associated with variables that are protected by law. For example, consider an algorithm designed to screen job applicants that considers the analytical skills of the applicants, but ignores their qualitative skills, although both are important for the job. Such an algorithm will systematically favor candidates with stronger analytical skills over those with stronger qualitative skills. In this case, the algorithm would be considered biased because it introduces a systematic error in the selection process. Yet, to the extent that such error is not associated with variables that are protected by law, the bias does not assume a discriminatory connotation.

Algorithmic bias is pervasive and hard to eradicate because it can stem from many sources (Hao, 2019). Bias might stem from the data used to train the algorithm. When trained on historical data that reflect pre-existing biases, algorithms might learn to make predictions based on associations that reflect such biases (Hajian et al., 2016). For example, an algorithm that scans resumes to screen job applicants might mistakenly screen out female applicants if the data used to train the algorithm reflects biased decisions made by humans in the past that resulted in men being consistently preferred over equally qualified women. In addition, bias can stem from the way computer scientists program algorithms. Programmers can intentionally or unintentionally ingrain their own views and beliefs into an algorithm by deciding which variables the algorithm should consider versus ignore, or which objective function an algorithm should optimize (Kleinberg et al., 2018a). Bias can also stem from improper use of algorithms, such as when an algorithm calibrated on data that is representative of a specific and homogeneous population is deployed to make predictions about a wider and more heterogeneous set of audiences (Danks & London, 2017).

Prior research on algorithmic bias has, for the most part, assumed a technical connotation, in that it has focused on documenting and defining algorithmic bias from a statistical standpoint (e.g., Dieterich et al., 2016; Kleinberg et al., 2016; Larson et al., 2016). The present research aims to contribute a psychological perspective to the ongoing debate on algorithmic bias, raising a fundamental question that has thus far received little attention: are people more or less likely to perceive decisions that yield disparities as biased, when such decisions stem from algorithms rather than humans?

Conceptual Development

We propose that the answer to this question hinges on the fundamental belief that algorithms decontextualize decision-making because they neglect the unique characteristics of the individual being judged (Longoni et al., 2019; Newman et al., 2020; Sloan & Warner, 2018). People think of algorithms as reductionist tools that standardize decision-making by blindly applying predetermined rules and procedures irrespective of whom they are judging, because they lack the cognitive flexibility necessary to tailor decision-making to each individual (Haslam, 2006; Loughnan & Haslam, 2007; Nissenbaum & Walker, 1998). In contrast, human decision-makers possess cognitive flexibility and are therefore considered better able to recognize and consider the unique characteristics of the target they are judging and, consciously or unconsciously, can tailor decision-making to each individual.

We further propose that the belief that algorithms decontextualize decision-making can sway the perception of bias in opposite directions, because decontextualization can be either detrimental or beneficial to the judgment at hand. On the one hand, unique individual characteristics can provide information that is relevant and can improve judgment (e.g., individual motivations and mental states); neglecting such information can undermine the accuracy of a judgment, yielding decisions that are more biased. On the other hand, individual characteristics can provide information that introduces irrelevant elements that can distort judgment (e.g., race and gender); neglecting such information can improve the accuracy of a judgment, yielding decisions that are less biased. Based on this logic, we argue that the belief that algorithms decontextualize decision-making can foster two different inferences, which sway the perception of bias in opposite directions, depending on whether the situation entails the potential for discrimination or not.

In situations that don't raise concerns about discrimination, the belief that algorithms decontextualize decision-making might foster the perception that algorithmic decisions are more biased than human decisions. This is because the belief that algorithms are unable to recognize the unique characteristics of an individual might lead people to think that algorithms are more likely than humans to ignore information about the individual that can be relevant to the judgment at hand, and doing so can lead to systematic errors (Sloan & Warner, 2018). For example, consider a situation that entails assessing a *white* defendant's risk of recidivism, a situation that is less likely to raise concerns about discrimination, as the defendant does not belong to a stigmatized population. In this situation, people might perceive an algorithmic assessment to be more biased than a judge's assessment because they might think that the algorithm does not consider, to the same extent that a judge would, information about the defendant that is relevant to make an accurate assessment (e.g., mitigating circumstances), therefore yielding systematic errors. This notion is consistent with prior literature on algorithm aversion (e.g., Dietvorst et al., 2015), according to which people often think that humans can make more accurate decisions than algorithms.

In situations that entail the potential for discrimination, in contrast, we propose that the belief that algorithms decontextualize decision-making might instead foster the perception that algorithmic decisions are less biased than human decisions. This is because the belief that algorithms are unable to recognize the unique characteristics of an individual might lead people to think that algorithms are more likely than humans to ignore information about the individual that could be grounds for discrimination, such as demographic information that denotes social group membership. As a consequence, people might think that algorithms are more likely than humans to treat everyone equally, avoiding systematic errors. For example, consider a situation that entails assessing a *black* defendant's risk of recidivism, a situation more likely to raise concerns about discrimination, as the defendant belongs to a stigmatized population. In this situation, people might perceive an algorithmic

assessment to be less biased than a judge's assessment because they might think that the algorithm, unlike the judge, does not consider information about the defendant that might be grounds for discrimination, such as the defendant's race.

In sum, we propose that the perception of bias for human as compared to algorithmic decisions is driven by the fundamental belief that algorithms, unlike humans, decontextualize decision-making because they are unable to recognize the unique characteristics of the individual being judged (Longoni et al., 2019; Newman et. al, 2020). We further argue that this fundamental belief can foster two different inferences, depending on whether the situation entails the potential for discrimination or not. In the absence of potential for discrimination (e.g., a white defendant), the belief that algorithms ignore the unique characteristics of the individual might lead people to infer that algorithms are more likely than humans to miss information that can be relevant to the judgment at hand, and doing so might lead to systematic errors (Sloan & Warner, 2018). As a consequence, algorithmic decisions could be perceived more biased than human decisions. In contrast, in the presence of potential for discrimination (e.g., a black defendant), the belief that algorithms ignore the unique characteristics of the individual might lead people to infer that algorithms are more likely than humans to treat everyone equally because they are more likely than humans to neglect information that might be grounds for discrimination, and doing so avoids systematic errors. As a consequence, algorithmic decisions should be perceived less biased than human decisions.

Our key hypothesis that, in situations that entail the potential for discrimination, algorithmic decisions that yield disparities are less likely than human decisions to be perceived as biased leads to three predictions about potential consequences of key societal relevance. First, perceiving that algorithmic decisions that yield disparities are less biased than human decisions can make people more likely to erroneously think that disparities stemming from algorithmic decisions are an accurate reflection of actual differences in dispositions and abilities, potentially reinforcing stereotypes that fuel discrimination. Second, perceiving that algorithmic decisions that yield disparities are less biased than human decisions can mislead members of stigmatized groups into preferring algorithmic over human evaluations, expecting that they will receive a more just treatment when decisions stem from algorithms rather than humans. Third, perceiving that algorithmic decisions that yield disparities are less biased than human decisions might thwart people's willingness to support actions against such disparities when they stem from algorithms rather than humans. In the remainder of the article we report nine studies that systematically test how people perceive algorithmic versus human decisions that yield disparities that entail the potential for discrimination, as well as the proposed psychological mechanism and downstream consequences.

Overview of the Studies

Studies 1a–c examine how people perceive algorithmic versus human decisions that yield disparities that entail the potential for discrimination. In particular, we test whether algorithmic decisions that yield gender or racial disparities are less likely than human decisions to be perceived as biased. Studies 2 and 3 test the proposed underlying mechanism. Study 2 tests the idea that the fundamental belief that algorithms, unlike humans, ignore the unique characteristics of the individual being judged fosters different inferences, which sway the perception of bias in opposite directions, depending on whether the situation entails the potential for discrimination or not. Study 3 aims to provide convergent evidence by testing whether, in a situation that entails the potential for discrimination, people think that algorithms are more likely than humans to ignore individual characteristics that could be grounds for discrimination. Studies 4, 5, and 6 explore societally relevant consequences. In particular, Study 4 tests the hypothesis that perceiving algorithmic decisions less biased than human decisions can make people more likely to erroneously think that disparities stemming from algorithms are a reflection of actual differences in dispositions and abilities. Studies 5a–b examine responses to algorithmic decisions by those who are most likely to be negatively impacted by algorithmic bias, namely, people who belong to stigmatized groups that are the target of discrimination. In particular, we test the hypothesis that members of stigmatized groups might prefer algorithmic over human evaluations in situations that entail the potential for discrimination. Finally, Study 6 explores people's propensity to take actions against disparities generated by algorithmic as opposed to human decisions. In particular, we test the hypothesis that people might be less likely to support actions aimed to remove disparities when decisions stem from algorithms rather than humans.

With respect to data practices, we report all conditions, manipulations, measures, and data exclusions. Unless reported, no participant was excluded. An attention check was included at the beginning of each Study. Respondents who failed the attention check did not qualify for the Study. In all studies, the sample size was predetermined, and we analyzed the data only after all responses were collected. A sensitivity power analysis (Faul et al., 2009) indicated that the studies had the power to detect effects of size considered to be of practical relevance (Ferguson, 2009), with a significance level α of .05 and a power (1–β) of .80 (Studies 1s: $d = .46$; Study 2 and Study 3: $\eta_p^2 = .02$; Study 4: $d = .46$; Study 5a: $d = .42$; Study 5b: $w = .20$; Study 6: $d = .46$). Experimental stimuli are provided in the Supplementary Materials.

Studies 1a–c

We open our empirical investigation by examining how people perceive algorithmic versus human decisions that yield disparities that entail the potential for discrimination. We examine decisions that yield gender and racial disparities across three domains where algorithms are increasingly being used to replace human decision-making, yet have been shown to have the potential to perpetrate bias: education, hiring, and criminal justice. Specifically, the first three studies test our basic hypothesis that algorithmic decisions that yield gender or racial disparities are less likely than human decisions to be perceived as biased.

Study 1a: University Admission Decisions

Method

Study 1a had a one-factor, between-subjects design: algorithm versus human decision-maker. We aimed to recruit 150 respondents. One hundred fifty respondents (59% females; age: $M = 39$,

SD = 12) recruited on Mturk read about a university where admission decisions were based on an analysis of applicants conducted either by an admission team or by an artificial intelligence (AI) software. To ensure that respondents understood the information presented, they were asked to indicate who/what conducted the analysis of the applicants, before moving on to the next screen. All but two respondents answered this question correctly. Removing these observations does not change the conclusion of the analysis reported below. Respondents were then presented with information about the university's acceptance rates for different ethnic groups, which showed a clear racial disparity, in that white candidates were three times more likely to be accepted than black and hispanic candidates (acceptance rate: 47% white applicants, 15% black applicants, 15% hispanic applicants). Respondents then rated the likelihood that the decisions made by the admission team/AI software were biased (1 = *very unlikely* 7 = *very likely*).

Results

As predicted, respondents perceived admission decisions that yielded a racial disparity less biased were they were made by an algorithm ($M = 3.97$, $SD = 2.06$) than by a human ($M = 4.82$, $SD = 1.49$), $t(148) = 2.88$, $p < .01$, $d = .47$.

Study 1b: Hiring Decisions

Method

Study 1b had a one-factor, between-subjects design: algorithm versus human decision-maker. We aimed to recruit 150 respondents. One hundred fifty-one respondents (51% females; age: $M = 38$, $SD = 11$) recruited on Mturk read that a local company had recently posted a job announcement for four store manager positions at four local restaurants. Respondents further read that the hiring decision was based on the results of an analysis of the candidates conducted by a recruiter/AI software. In both cases, it was specified that, to predict the success of a candidate as a store manager, the recruiter/AI software took into consideration background information about the candidates, along with answers to attitudinal questions. To ensure that respondents understood the information presented, they were asked to indicate who/what conducted the analysis of the candidates, before moving on to the next screen. All but one participant answered this question correctly. Removing the observations from this participant does not change the conclusion of the analysis reported below. Respondents then read that 100 people applied for the four positions and that the demographics of the applicants were as follows: 45 white men, 23 white females, 20 black men, and 12 black females. Respondents then learned that four white males were hired, an outcome indicating both racial and gender disparities. Respondents then rated the likelihood that the decision made by the recruiter/AI software was biased (1 = *very unlikely*; 7 = *very likely*).

Results

As predicted, respondents perceived hiring decisions that yielded a racial and gender disparity less biased when they were made by an algorithm ($M = 4.18$, $SD = 1.96$) rather than by a human ($M = 4.89$, $SD = 1.85$), $t(149) = 2.29$, $p = .02$, $d = .37$.

Study 1c: Parole Decisions

Method

Study 1c had a one-factor, between-subjects design: algorithm versus human decision-maker. We aimed to recruit 150 respondents. One hundred forty-nine respondents (59% females; age: $M = 39$, $SD = 12$) recruited on Mturk read that parole decisions often rely on an assessment of the risk that a defendant will reoffend, made by a judge/AI software. They further read that risk assessments are expressed with a numerical score ranging from 1 to 10, where lower scores indicate lower risk. To ensure that respondents understood the information presented, they were asked to indicate who/what performs the risk assessments, before moving on to the next screen. Eight respondents did not answer this question correctly. Removing these observations does not change the conclusion of the analysis reported below. Respondents were then presented with the average risk assessment score for two ethnic groups, which showed a clear racial disparity: the average risk score for black men defendants (8 out of 10) was 60% higher than the average risk score for white men defendants (5 out of 10). Respondents then rated the likelihood that the risk assessments made by the judge/AI software were biased (1 = *very unlikely*; 7 = *very likely*).

Results

As predicted, respondents perceived judicial risk assessments that yielded a racial disparity less biased when they were made by an algorithm ($M = 4.14$, $SD = 1.92$) rather than by a human ($M = 4.95$, $SD = 1.47$), $t(147) = 2.85$, $p < .01$, $d = .47$.

Discussion

Overall, Studies 1a–c show that decisions that yield racial and gender disparities are less likely to be perceived as biased when they stem from algorithms rather than humans. This result was robust across three domains where algorithms are increasingly used to replace human decision-makers and algorithmic bias has been documented: education, hiring, and criminal justice. Our findings provide preliminary evidence that, in situations that entail the potential for discrimination, people might not perceive algorithmic bias to the same extent they perceive human bias. In the next studies, we explore the psychological mechanism that drives such differential perception of bias (Studies 2 and 3), as well as downstream consequences of societal relevance (Studies 4–6).

Study 2

We proposed that differences in perception of bias for human versus algorithmic decisions are driven by the belief that algorithms, unlike humans, decontextualize decision-making because they are unable to recognize the unique characteristics of the individual being judged and blindly apply predetermined rules and procedures in a rigid way, irrespective of whom they are judging (Haslam, 2006; Loughnan & Haslam, 2007). We further argued that this fundamental belief can foster two inferences, which sway the perception of bias in opposite directions, depending on whether the situation entails the potential for discrimination or not.

When the situation entails the potential for discrimination (e.g., a *black* defendant), the belief that algorithms ignore the

unique characteristics of the individual being judged might lead people to infer that algorithms are more likely than humans to treat everyone equally because they ignore information about the individual that might be grounds for discrimination. As a consequence, algorithmic decisions should be perceived less biased than human decisions.

In contrast, when the situation does not entail the potential for discrimination (e.g., a *white* defendant), the belief that algorithms ignore the unique characteristics of the individual being judged might lead people to infer that algorithms are more likely than humans to miss information about the defendant that can be relevant to the judgment at hand, and doing so might be detrimental to the accuracy of an assessment (Sloan & Warner, 2018). As a consequence, algorithmic decisions should be perceived more biased than human decisions.

Method

Study 2 had a 2(decision-maker: human vs. algorithm) × 2(defendant race: white vs. black) between-subjects design. We aimed to recruit 400 respondents. Three hundred ninety-nine respondents (58% females; age: $M = 40$, $SD = 13$) recruited on Mturk were presented with a news article about a man being sentenced to 5 years in prison for stealing a car, based on a risk assessment that deemed him at a high risk of recidivism. Between subjects, we manipulated the potential for discrimination by changing the race of the defendant (white vs. black). Moreover, we manipulated whether the risk assessment was conducted by a judge or by an algorithm. Respondents rated the likelihood that the risk assessment was biased (bias: 1 = *very unlikely*; 7 = *very likely*); based on our proposed theoretical account, we expected to observe a Race × Decision-maker interaction on this variable. Then, they rated the extent to which they thought that the judge/algorithm was blind to the unique characteristics of the defendant (blindness to individual characteristics: 1 = *strongly disagree*; 7 = *strongly agree*), the judge/algorithm missed information that was relevant to make an accurate assessment (missing relevant information: 1 = strongly disagree; 7 = strongly agree), and the judge/algorithm treated the defendant like any other defendant (treating everyone equally: 1 = strongly disagree; 7 = strongly agree). Based on our proposed theoretical account, we expected to observe a main effect of decision-maker on blindness to individual characteristics, and a Race × Decision-maker interaction on missing relevant information and treating everyone equally.

Results

A 2 × 2 ANOVA on bias revealed a significant main effect of race [$F(1, 395) = 32.76, p < .001, \eta_p^2 = .08$], a non-significant main effect of decision-maker [$F(1,395) < 1$], and a Significant race × Decision-maker interaction ($F(1, 395) = 20.54$, $p < .001$, $\eta_p^2 = .05$]. As predicted, when the defendant was black, respondents considered the risk assessment less biased when it was performed by an algorithm ($M = 3.64$, $SD = 1.97$) rather than by a judge ($M = 4.55$, $SD = 1.74, F(1, 395) = 14.73, p < .001, \eta_p^2 = .04$). When the defendant was white, the effect reversed, and respondents considered the risk assessment more biased when performed by an algorithm ($M = 3.44$, $SD = 1.53$) rather than by a judge ($M = 2.83$, $SD = 1.42$, $F(1, 395) = 6.62$, $p = .01$, $\eta_p^2 = .02$).

A 2 × 2 ANOVA on blindness to individual characteristics revealed a significant main effect of decision-maker [$F(1, 395) = 46.28$, $p < .001$, $\eta_p^2 = .11$], a significant main effect of race [$F(1, 395) = 5.73$, $p = .02$, $\eta_p^2 = .01$], and a non-significant race × Decision-maker interaction [$F(1, 395) < 1$]. As predicted, respondents perceived an algorithm to be more blind to the individual characteristics of the defendant than a judge, both when the defendant was black ($M_{algorithm} = 4.63$, $SD_{algorithm} = 1.80$; $M_{judge} = 3.66$, $SD_{judge} = 1.39$, $F(1, 395) = 19.84$, $p < .001$, $\eta_p^2 = .05$), and when the defendant was white ($M_{algorithm} = 4.34$, $SD_{algorithm} = 1.51$; $M_{judge} = 3.21$, $SD_{judge} = 1.43$, $F(1, 395) = 26.69$, $p < .001$, $\eta_p^2 = .06$).

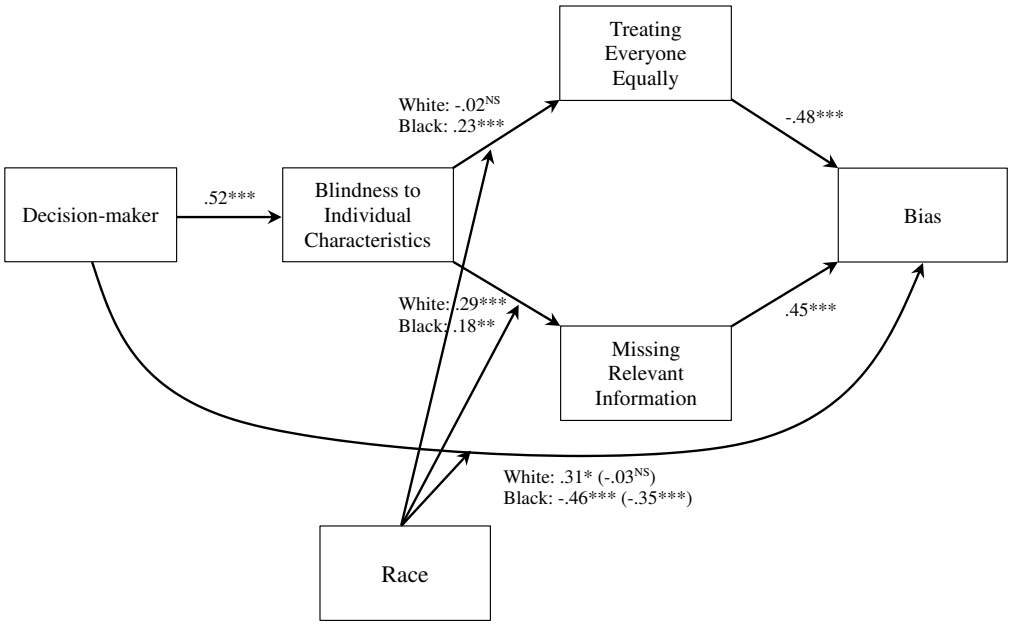
A 2 × 2 ANOVA on missing relevant information revealed a significant main effect of decision-maker [$F(1, 395) = 26.29$, $p < .001$, $\eta_p^2 = .06$], a non-significant main effect of race [$F(1, 395) = 2.24$, $p = .14$, $\eta_p^2 = .01$], and a Significant race × Decision-maker interaction [$F(1, 395) = 13.19$, $p < .001$, $\eta_p^2 = .03$]. As predicted, when the defendant was white, respondents perceived that the algorithm ($M = 4.25$, $SD = 1.45$) missed information that was relevant to make an accurate assessment to a greater extent than the judge did ($M = 3.02$, $SD = 1.20$, $F(1, 395) = 38.26$, $p < .001$, $\eta_p^2 = .09$), whereas, when the defendant was black, there was no difference between algorithm ($M = 3.95$, $SD = 1.55$) and judge ($M = 3.74$, $SD = 1.38$, $F(1, 395) = 1.12$, $p = .29$, $\eta_p^2 = .00$).

A 2 × 2 ANOVA on treating everyone equally revealed a significant main effect of race [$F(1, 395) = 14.68$, $p < .001$, $\eta_p^2 = .04$], a non-significant main effect of decision-maker [$F(1, 395) = 2.00$, $p = .16$, $\eta_p^2 = .01$], and a Significant race × Decision-maker interaction [$F(1, 395) = 9.08$, $p < .01$, $\eta_p^2 = .02$]. As predicted, when the defendant was black, respondents perceived that the defendant was treated like any other defendant more by an algorithm ($M = 5.11$, $SD = 1.71$) than by a judge ($M = 4.46$, $SD = 1.59$, $F(1, 395) = 9.83$, $p < .01$, $\eta_p^2 = .02$), whereas, when the defendant was white, there was no difference between algorithm ($M = 5.23$, $SD = 1.25$) and judge ($M = 5.46$, $SD = 1.25$, $F(1, 395) = 1.27$, $p = .26$, $\eta_p^2 = .00$).

To test our proposed mechanism, we conducted a moderated serial mediation analysis using the custom model depicted in Figure 1, with 5,000 bootstraps (Hayes, 2018). Decision-maker served as the independent variable (−1 = *judge* and 1 = *algorithm*), bias as the dependent variable, blindness to individual characteristics as the first mediator, equal treatment and missing relevant information as competing subsequent mediators, and race as the moderator (−1 = *white* and 1 = *black*).

In line with our theorizing, when the defendant was black, the results showed a significant indirect effect via blindness to individual characteristics and equal treatment (i.e., decision-maker → blindness to individual characteristics → equal treatment → bias; $b = -.06$, 95% CI: −.11 to −.02), whereas there was no evidence for a significant indirect effect via blindness to individual characteristics and missing relevant information (i.e., decision-maker →blindness to individual characteristics → missing relevant information → bias), as the confidence interval included zero ($b = .04$, 95% CI: .00–.09). This result is consistent with our hypothesis that, when the situation entails the potential for discrimination (i.e., a black defendant), the belief that algorithms are more blind than humans to individual characteristics leads to the perception that algorithms are more likely than humans to treat everyone equally. As a consequence, algorithmic decisions are perceived less biased than human decisions.

Figure 1
Moderated Serial Mediation Model in Study 2



Note. * $p < .05$. ** $p < .01$. *** $p < .001$.

In contrast, when the defendant was white, the results showed a significant indirect effect via blindness to individual characteristics and missing relevant information (i.e., decision-maker → blindness to individual characteristics → missing relevant information → bias; $b = .07$, 95% CI: .02–.12), whereas there was no evidence for a significant indirect effect via blindness to individual characteristics and equal treatment, as the confidence interval included zero (i.e., decision-maker → blindness to individual characteristics → equal treatment → bias; $b = .01$, 95% CI: −.02–.04). This result is consistent with our hypothesis that, when the situation does not entail the potential for discrimination (i.e., a white defendant), the belief that algorithms are more blind than humans to individual characteristics leads to the perception that algorithms are more likely than humans to miss information that is relevant to the judgment at hand. As a consequence, algorithmic decisions are perceived more biased than human decisions.

Discussion

While we acknowledge the correlational nature of this analysis (Fiedler et al., 2018), these findings provide support for our proposed psychological mechanism. Consistent with our theorizing, the results suggest that people believe that algorithms are more likely than humans to ignore the unique characteristics of the individual being judged. This fundamental belief fosters different inferences, depending on whether individual characteristics entail the potential for discrimination or not. When a target's individual characteristics entail the potential for discrimination (i.e., a black defendant), the belief that algorithms ignore the unique characteristics of the individual leads people to infer that algorithms are more likely than humans to treat everyone equally. Consequently, algorithmic decisions are perceived less biased than human decisions. In contrast, when a target's individual characteristics do not entail the potential for discrimination (i.e., a white defendant), the belief that algorithms ignore the unique characteristics of the individual leads people to infer that algorithms are more likely than humans to miss information that is relevant to the judgment at hand. Consequently, algorithmic decisions are perceived more biased than human decisions.

In Study 3, we aim to provide convergent evidence for the proposed underlying mechanism via moderation. In particular, we aim to show that, when individual characteristics entail the potential for discrimination (e.g., a black defendant), the belief that algorithms ignore the unique characteristics of the individual being judged leads people to think that algorithms are more likely than humans to ignore information that might be grounds for discrimination.

Study 3

The results of Study 2 show that people believe that algorithms are more likely than humans to ignore the individual characteristics of the target being judged. We argued that, in situations that entail the potential for discrimination, this belief leads people to think that algorithms are more likely than humans to ignore individual characteristics that could be grounds for discrimination. In Study 3, we test this hypothesis by directly manipulating whether or not respondents were informed that a decision-maker (i.e., human versus

algorithm) considered information about social group membership that can be grounds for discrimination (i.e., the demographic profile of an individual). If our hypothesis is correct, in a situation that entails the potential for discrimination (e.g., a black defendant), human decisions should be perceived more biased than algorithmic decisions, unless people are informed that the algorithm considered the demographic profile of an individual, thus information about social group membership that can be grounds for discrimination.

Method

Study 3 had a 2(decision-maker: human vs. algorithm) × 2(demographics: control vs. considered) between-subjects design. We aimed to recruit 400 respondents. Four hundred respondents (54% females; age: $M = 40$, $SD = 13$) recruited on Mturk were presented with the same news article used in Study 2, about a black man who was sentenced to five years in prison for stealing a car, based on a risk assessment that deemed him at a high risk of recidivism. Between subjects, we manipulated whether the risk assessment was conducted by a judge or by an algorithm, and whether respondents were informed that the judge/algorithm considered the defendant's demographic profile (e.g., information about race, gender, and age). Respondents rated the likelihood that the risk assessment was biased (1 = *very unlikely*; 7 = *very likely*).

Results

A 2 × 2 ANOVA on likelihood of bias revealed a significant main effect of decision-maker [$F(1, 396) = 4.16$, $p = .04$, $\eta_p^2 = .01$], a non-significant main effect of demographics [$F(1, 396) = 3.14$, $p = .08$, $\eta_p^2 = .01$], and a Significant decision maker × Demographics interaction [$F(1, 396) = 6.12$, $p = .01$, $\eta_p^2 = .02$]. As predicted, when the decision-maker was not specified to have considered the demographic profile of the defendant, respondents perceived the risk assessment to be less biased when performed by an algorithm ($M = 3.44$, $SD = 1.82$) than by a judge ($M = 4.29$, $SD = 1.73$, $F(1, 396) = 10.13$, $p < .01$, $\eta_p^2 = .03$), thus replicating the results of Study 2. In contrast, when the decision-maker was specified to have considered the demographic profile of the defendant, respondents perceived the risk assessment performed by the algorithm ($M = 4.24$, $SD = 2.05$) as biased as the risk assessment performed by the judge ($M = 4.16$, $SD = 1.89$, $F(1, 396) < 1$).

From a different angle, when the risk assessment was performed by the judge, respondent's perception of bias did not differ as a function of whether the judge was specified to have considered the demographic profile of the defendant ($M = 4.16$, $SD = 1.89$) or not ($M = 4.29$, $SD = 1.73$, $F(1, 396) < 1$). This finding suggests that respondents assumed that the judged considered the demographic profile of the defendant, even when such information was not provided. In contrast, when the risk assessment was performed by the algorithm, respondent's perceived it more biased when the algorithm was specified to have considered the demographic profile of the defendant ($M = 4.24$, $SD = 2.05$) than when such information was not provided ($M = 3.44$, $SD = 1.82$, $F(1, 396) = 8.97$, $p < .01$, $\eta_p^2 = .02$). This finding suggests that respondents assumed that the algorithm did not consider the demographic profile of the defendant, when such information was not explicitly provided.

Discussion

The results of Study 3 provide further evidence for our proposed underlying mechanism by showing that, in situations that entail the potential for discrimination, people perceive algorithmic decisions to be less biased than human decisions because they think that algorithms are more likely than humans to ignore individual characteristics that could be grounds for discrimination. Taken together, the first three studies provide support for our hypothesis that people perceive decisions that yield outcomes that entail the potential for discrimination (e.g., racial or gender disparities) less biased when they stem from an algorithm rather than a human. In the next three studies, we test three consequences that should stem from this asymmetrical perception of bias, and that have societal relevance.

Study 4

We proposed that, if decisions that yield disparities are less likely to be perceived as biased when they stem from algorithms rather than humans, people might then be more likely to think that such disparities reflect actual differences in dispositions and abilities. In Study 4, we test this hypothesis in the context of gender disparities in hiring. Specifically, we expect that people might be more likely to rationalize gender disparities as a reflection of actual differences in dispositions and abilities between genders, when such disparities stem from algorithmic rather than human decisions. If supported, this prediction implies that replacing human with algorithmic decision-making might contribute to fostering stereotypical beliefs about stigmatized groups.

Method

Study 4 had a one-factor, between-subjects design: algorithm versus human decision-maker. We aimed to recruit 150 respondents. One hundred fifty respondents (56% females; age: $M = 40$, $SD = 13$) recruited on Mturk read that a company was looking to hire an executive director, and that the company had used either a recruiter or an AI software to screen applicants and select a few candidates to interview for the position. To ensure that respondents understood the information presented, they were asked to indicate who/what conducted the analysis of the candidates, before moving on to the next screen. Four respondents did not answer this question correctly. Removing these observations does not change the conclusions of the analysis reported below. Respondents then read that 68 males and 32 females applied for the position and that five males were selected to be interviewed, an outcome indicative of a gender disparity. Respondents were then asked to rate the likelihood that the decision made by the recruiter/AI software was biased (1 = *very unlikely*; 7 = *very likely*). Then, respondents indicated their agreement with four items, adapted from Cundiff and Vescio (2016), that explained the decision in terms of dispositional differences between genders (e.g., the women who applied possessed skills that are better suited for other positions; $\alpha = .89$).

Results

As predicted, respondents perceived a hiring decision that yielded a gender disparity less biased when it was made by an algorithm ($M = 4.28$, $SD = 1.98$) rather than by a human ($M = 5.04$, $SD = 1.60$), $t(148) = 2.59$, $p = .01$, $d = .42$. Moreover,

respondents were more likely to explain the decision in terms of dispositional differences among genders when it was made by an algorithm ($M = 4.33$, $SD = 1.37$) rather than by a recruiter ($M = 3.70$, $SD = 1.47$), $t(146) = 2.68$, $p = .01$, $d = .44$ (two missing values on this measure). This result suggests that respondents considered female candidates less qualified for the executive position when the gender disparity stem from an algorithmic rather than a human decision.

A mediation analysis (Hayes, 2018; Model 4) revealed a significant indirect effect via perception of bias (i.e., decision-maker → bias → dispositional differences; $b = .28$, 95% CI: .08–.52). Moreover, when controlling for perception of bias, the main effect of decision-maker on dispositional differences became non-significant [$t(145) = 1.60$, $p = .11$], thus indicating full mediation.

Discussion

The results of Study 4 support our hypothesis that, because algorithmic decisions that yield disparities are perceived less biased than human decisions, people are more likely to rationalize such disparities as a reflection of actual differences in dispositions and abilities when they stem from algorithms rather than humans. In our Study, when an algorithm, as opposed to a recruiter, made a hiring decision that favored men over women, respondents rationalized the decision by thinking that female applicants were actually less qualified than male applicants. This result has important societal implications because such rationalizations can reinforce stereotypes that inhibit efforts aimed to combat discrimination (Reyna, 2000), thus contributing to perpetuate inequalities (Cundiff & Vescio, 2016). We elaborate further on this point in the general discussion.

Studies 5a–b

In the first four studies, respondents judged situations where disparities affected others. In Study 5a–b, we test whether our findings hold even when members of stigmatized groups are directly affected. It is possible that members of stigmatized groups, who are more often subject to discrimination (Operario & Fiske, 2001), might be more sensitive to the presence of disparities, and thus more likely to perceive decisions that yield disparities as biased, regardless of whether they stem from algorithms or from humans. However, members of stigmatized groups might also share the same fundamental beliefs about how algorithms operate as those who do not belong to stigmatized groups. Consequently, even members of stigmatized groups should be more likely to perceive algorithmic decisions that yield disparities less biased than human decisions. In studies 5a–b, we test this hypothesis by asking female participants to evaluate two situations that entail the potential for gender discrimination in hiring. Specifically, In Study 5a, we examine to what extent female participants perceive as biased an allegedly discriminatory decision that leads to the rejection of a job application, as a function of whether the decision is made by a human versus an algorithm. We expected that women would perceive an allegedly discriminatory decision less biased when made by an algorithm rather than a human. In Study 5b, we explore whether female participants prefer to be evaluated by a human versus an algorithm, when concerns about gender discrimination are more versus less salient. We expected that women would prefer to be evaluated by an algorithm rather than a human when concerns about discrimination are more versus less salient.

Study 5a: Perception of Bias About a Job Rejection

Method

Study 5a had one-factor, between-subjects design: algorithm versus human decision-maker. We aimed to recruit 150 female respondents. The final sample size was determined by the following procedure. We recruited 300 respondents on Mturk. At the beginning of the Study, respondents were asked to indicate their gender (male, female, and prefer not to answer). Only female respondents were redirected to our Study. We expected that at least 50% of respondents would be females and that this procedure would therefore yield the desired sample size. The final sample consisted of 180 female respondents (age: $M = 38$, $SD = 12$).

Respondents were asked to imagine that they had applied for a job at a company that was rumored to discriminate against female applicants. Respondents were then presented with an extract from an email from the company stating that they had not been selected for a follow-up interview. The email mentioned that the review process was conducted by a recruiter/algorithm. To ensure that respondents understood the information presented, they were asked to indicate who/what screened the application before moving on to the next screen. All but two respondents answered the question correctly. Removing these observations does not change the conclusion of the analysis reported below. Respondents then rated the likelihood that the decision made by the recruiter/algorithm was biased (1 = *very unlikely*; 7 = *very likely*).

Results

As predicted, respondents perceived the decision less biased when it was made by an algorithm ($M = 4.28$, $SD = 1.85$) than a human ($M = 5.10$, $SD = 1.54$), $t(178) = 3.23$, $p = .002$, $d = .48$.

Study 5b: Choosing How to be Evaluated

Method

Study 5b had a one-factor, between-subjects design: potential for discrimination more versus less salient. We aimed to recruit 200 female respondents. For this Study, we recruited only female respondents by specifying gender as a recuiting criterion on Turk-Prime (Litman et al., 2017). Two hundred female respondents (age: $M = 40$, $SD = 12$) were asked to imagine that they were applying for a job in an industry where women are underrepresented, for which they knew they were well-qualified. They further read that they had the option to have their application screened either by an algorithm or by a recruiter. In one condition, we made the potential for discrimination more salient by drawing respondents' attention to the possibility that they could be discriminated against because of their gender. In the other condition, no concern about gender discrimination was raised, thus the potential for discrimination was less salient. Respondents indicated whether they would prefer to have their application screened by a recruiter or by an algorithm. To confirm that our manipulation worked as intended, we then asked respondents to rate the extent to which, when making their decision,

they were concerned about the possibility of being discriminated because of their gender (1 = *not at all*; 7 = *very much*).

Results

Respondents were more likely to be concerned about the possibility of being discriminated because of their gender when the potential for discrimination was more salient ($M = 5.25$, $SD = 1.70$) than when it was less salient ($M = 4.60$, $SD = 2.07$), $t(198) = 2.43, p = .02, d = .34$, confirming the effectiveness of our manipulation. More importantly, as predicted, participants choice differed significantly between conditions, $\chi2(1) = 15.69, p < .001$, $w = .28$. When the potential for discrimination was less salient, only 37% of respondents chose to be evaluated by the algorithm, a choice share significantly below the point of indifference (i.e., 50%), $Z = -2.6$. $p = .01$. In contrast, when the potential for discrimination was more salient, 65% of respondents chose to be evaluated by the algorithm, a choice share significantly above the point of indifference, $Z = 3.00$, $p < .01$.

Discussion

Overall, Studies 5a–b show that even members of stigmatized groups might perceive algorithmic decisions that entail the potential for discrimination less biased than human decisions. In our studies, women considered an allegedly discriminatory employment decision less biased when it stemmed from an algorithm rather than a recruiter. And they preferred to be evaluated by an algorithm rather than a recruiter when the potential for discrimination was salient. These findings contribute to understanding how individuals that are more susceptible to the negative consequences of discrimination perceive algorithmic as opposed to human decisions. In so doing, these findings warn again the risk that systematic discrimination might go more undetected when perpetrated by algorithms rather than humans, even by those who stand to be most impacted by it. In Study 6, we aim to explore an additional societal implication of our findings, namely, whether algorithmic, as opposed by human decisions, affect people's willingness to take actions against disparities that might be discriminatory.

Study 6

Prior research suggests that perceiving decisions that yield disparities as biased is a necessary condition to mobilize people to support actions aimed to remove unjust inequalities in our society (Corcoran et al., 2015; Earl, 2004, 2006). If people perceive algorithmic decisions that yield disparities less biased than human decisions, they might be less likely to support actions aimed to remove such disparities when they stem from algorithms rather than humans. We test this hypothesis in Study 6.

Method

The Study had a one-factor, between-subjects design: algorithm versus human decision-maker. We aimed to recruit 150 respondents. One hundred fifty respondents (55% females; age: $M = 40$, $SD = 13$) recruited on Mturk read that a reporter investigated a local company and discovered that female employees were less likely to receive bonus payments than male employees. They then read that the company publicly responded that bonus payments were determined by a department manager/algorithm based on the performance and qualifications of the employees, and not on their gender. Finally, respondents were read that a group of workers started a petition against the company for gender discrimination, and were asked whether they would consider signing the petition (1 = *definitely not*; 6 = *definitely yes*).

Results

As predicted, respondents were less willing to sign the petition when disparities in bonus payments were determined by an algorithm ($M = 3.41$, $SD = 1.65$) than by a human ($M = 4.03$, $SD = 1.43$), $t(148) = 2.43$, $p = .02$, $d = .40$.

Discussion

The results of Study 6 suggest that people might be less likely to take action against decisions that yield disparities when such decisions stem from algorithms rather than humans. This result has important societal implications, in that it suggests that replacing humans with algorithmic decision-making might make people less likely to take action against decisions that might, in fact, be discriminatory.

General Discussion

Across nine studies, we document a tendency to perceive algorithmic decisions that yield disparities that entail the potential for discrimination less biased than human decisions. In Study 1, algorithmic decisions that yielded gender and racial disparities were perceived less biased than human decisions. This result was robust across three domains of key societal importance, where algorithms are increasingly replacing human decision-makers. Studies 2 and 3 showed that this differential perception of bias is driven by the fundamental belief that algorithms, unlike humans, decontextualize decision-making because they ignore the unique characteristics of the individual being judged. This fundamental belief fosters different inferences, which sway the perception of bias in opposite directions, depending on whether the situation entails the potential for discrimination or not. Studies 4, 5, and 6 showed that this differential perception of bias has societally relevant consequences. Perceiving algorithmic decisions less biased than human decisions can (a) induce people to erroneously think that disparities that stem from algorithms are an accurate reflection of actual differences in dispositions and abilities (Study 4), (b) lead members of stigmatized groups into being more accepting of algorithmic decisions that might, in fact, be discriminatory (Study 5a), (c) preferring algorithmic over human evaluations (Study 5b), and (4) reduce people's propensity to take actions against disparities that might be discriminatory, when such disparities are perpetrated by algorithms (Study 6). Altogether, these findings have both theoretical and practical implications.

Theoretical Contributions

Our findings contribute to research on algorithmic bias. Prior investigations have predominantly focused on documenting the *existence* of algorithmic bias and defining it from a technical

standpoint (e.g., Friedman & Nissenbaum, 1996; Kleinberg et al., 2016). We add to this line of research by contributing a psychological perspective, exploring whether people *perceive* algorithmic bias. Our research provides one of the first attempts to understand people's perception of algorithmic bias, unpacking the duality inherent in the notion of bias as systematic error that can have a discriminatory or nondiscriminatory connotation. Our research shows that when confronted with decisions that yield outcomes that are not discriminatory, people may be prone to overestimate algorithmic bias, a tendency that might induce people to over rely on human judgment. Yet, when confronted with decisions that yield outcomes that can be discriminatory, people may be prone to underestimate algorithmic bias, a tendency that might induce people to over rely on algorithmic judgment. In doing so, our work enriches the current debate on algorithmic bias by broadening its scope and highlighting the importance of understanding people's perception of bias and its consequences. We contend that, although the statistical detection of bias is of paramount importance, understanding people's perceptions of algorithmic bias is equally important, as efforts aimed to remove bias in society are often driven by whether or not people perceive such bias (Corcoran et al., 2015; Earl, 2004).

Our research also contributes to the literature on clinical versus statistical judgments, in two distinct ways. After Meehl's seminal contribution (Meehl, 1954), a growing body of evidence has shown that people tend to trust human more than algorithmic decisions (Grove & Meehl, 1996), a phenomenon referred to as algorithm aversion (Dietvorst et al., 2015). This behavior can be suboptimal in light of evidence suggesting that algorithms can outperform human intuition (Dawes et al., 1989; Grove et al., 2000). Our research suggests that algorithm aversion might not manifest, and might even reverse, in situations that entail the potential for discrimination. For example, in Study 5b, female respondents preferred to be evaluated by a human rather than an algorithm *only* when the potential for discrimination was not salient. When the potential for discrimination was salient, the effect reversed, such that female respondents preferred to be evaluated by an algorithm rather than a human. Thus, our findings add to recent research that suggests that algorithm aversion might not be as universal as previously thought (Castelo et al., 2019; Logg et al., 2019).

Our research further contributes to the literature on clinical versus statistical judgments by providing a more nuanced understanding of a key psychological driver of algorithm aversion, namely, algorithms' perceived inability to contextualize decision-making (Newman et al., 2020; Sloan & Warner, 2018). Prior research argues that algorithms' perceived inability to contextualize decision-making leads people to infer that algorithms are more prone to error than humans, because they are more likely than humans to neglect information that is relevant to the judgment at hand, a belief that drives algorithm aversion (Longoni et al., 2019). In contrast, our results suggest that, in situations that entail the potential for discrimination, algorithms' perceived inability to contextualize decision-making triggers different inferences, leading to a reversal of algorithm aversion. In these situations, this fundamental belief can lead people to infer that algorithms are less prone to error than humans, because they are more likely than humans to neglect information that might be grounds for discrimination. Thus, our findings add to prior research by showing that perceived decontextualization can trigger different inferential processes. Depending on the situation, decontextualization can be perceived to lead to a loss of information that is detrimental rather than beneficial to the judgment at hand, potentially resulting in algorithm appreciation rather than aversion.

Finally, our findings contribute to research on social inequalities. Past research (Cundiff & Vescio, 2016; Yzerbyt et al., 1997) has examined how stereotypical beliefs affect how people perceive disparities. Specifically, Cundiff and Vescio (2016) showed that those who endorse gender stereotypes are more likely to attribute gender disparities to dispositional differences between men and women, and less to discrimination. We contribute to this literature by identifying a novel factor that sways how people make sense of disparities, namely, whether disparities stem from decisions made by humans versus algorithms. Our research shows that people are more likely to rationalize gender disparities as a reflection of actual differences in dispositions and abilities between genders when such disparities stem from algorithmic rather than human decisions, a conclusion that can be dangerously erroneous.

Practical Implications

If people are less likely to perceive decisions that yield disparities as biased when they stem from algorithms rather than humans, then replacing human with algorithmic decision-making can potentially contribute to legitimize discrimination. For example, the results of Study 4 suggest that algorithmic decision-making may increase the risk that disparities might be erroneously interpreted as an accurate reflection of differences in people's dispositions and abilities. As a result, algorithms might reinforce stereotypical beliefs that contribute to perpetuating discrimination. Indeed, a society's first line of defense against discrimination lays in people's ability to recognize that disparities might stem from biased decisions (Crosby, 1993; Spring et al., 2018). Perceiving bias is a necessary condition to mobilize efforts aimed to remove inequalities in society (Corcoran et al., 2015; Earl, 2004). The results of Study 6 suggest that algorithmic decision-making may make people less likely to take actions against decisions that might, in fact, be discriminatory. We believe that algorithms have tremendous potential to unveil and eventually rid decision-making from human biases (Kleinberg et al., 2020), leading to more accurate and equitable results (Gates et al., 2002). Yet, to the extent that algorithmic bias exists and is hard to detect and eradicate (Hao, 2019), the use of algorithms might have societal implications that should not be ignored. More research is needed to explore these implications.

By showing that differential perceptions of bias for human versus algorithmic decision-making are driven by erroneous beliefs about how algorithms operate, our research highlights the importance of fostering algorithmic literacy. Recent investigations show that algorithmic bias can go undetected by the public, even when it affects the well-being of millions of people (Obermeyer et al., 2019). Our findings suggest that education can alleviate the danger that the general public might be deceived by the apparent objectivity of algorithmic decision-making. For example, the results of Study 3 show that the asymmetric perception of bias was eliminated when participants were informed that an algorithm considered information about the demographic profile of an individual. Yet, organizations that use these algorithms often reassure the general public and policymakers that their tools do not rely on such information (Puente, 2019). We argue that such claims may exploit people's misconceptions about algorithms. Most instances of algorithmic

bias are not due to the direct use of protected variables as inputs into algorithmic decision-making, but to the use of non-protected "proxies" that are highly associated to protected variables (Barocas & Selbst, 2016; Hajian et al., 2016). For example, an algorithm that uses SAT scores to screen students' applications might introduce bias in the selection process even if it ignores race as a variable, because of the historical association between SAT scores and race (Geiser, 2015). Thus, the fact that an algorithm does not use protected variables does not guarantee that the algorithm is unbiased, yet people might erroneously conclude that it is. In fact, technical debates about how to debias algorithms have pointed to the need for algorithms to actually use protected variables in their predictions as a way to detect and correct for bias (Gillis & Spiess, 2019; Kleinberg et al., 2018b). Our research calls to establishing checks and balances aimed not only at detecting and alleviating algorithmic bias, but also at educating the general public to think critically about the way algorithms operate.

Limitation and Future Research

Our research provides robust evidence that perceptions of bias differ as a function of whether decisions that yield disparities stem from algorithms versus humans. Yet, our research also has limitations that open avenues for future research. We focus our discussion on six key areas that we believe offer particularly fruitful opportunities to extend the current work.

First, our investigation was by nature limited in scope, and as such, it does not provide an exhaustive analysis of all the possible instances of algorithmic bias. For example, we focused only on gender and racial disparities. Additional research is needed to examine perceptions of bias for decisions that yield disparities related to other protected variables, such as age, sexual orientation, or religion. Moreover, the disparities investigated in our studies are not representative of the full range of disparities people may be confronted with. Facing disparities that are more versus less extreme might either thwart or magnify perceptions of bias. Indeed, in our studies, respondents were relatively ambivalent about the perception of algorithmic bias, as suggested by responses around the midpoint of the measure of bias. Thus, future research is needed to generalize our findings to a broader range of situations, where disparities might be more versus less pronounced.

Second, it is possible that differential perceptions of bias for algorithmic versus human decisions might be, at least in part, contingent on the general public being misinformed about the existence of algorithmic bias. That is, based on everyday experience, people might be more familiar with instances of human rather than algorithmic bias. As such, the phenomenon we document might progressively attenuate as people become more educated about instances of algorithmic bias. Although we agree that awareness of algorithmic bias is key to mitigating the erroneous perception that algorithms are unbiased decision-makers, our results might point to a more fundamental issue than simple misinformation. In this regard, it is important to mention that our empirical investigation was conducted at a time of an extensive media coverage of instances of algorithmic bias. More research is needed to investigate what kind of educational interventions might be most effective at eradicating the erroneous beliefs that drive the phenomenon we document.

Fourth, to probe the generalizability of our results, we tested our hypothesis across different domains where algorithms are widely used to replace human decision-making, and modeled our stimuli after documented cases of algorithmic bias. For example, in the hiring domain, where algorithms are increasingly being used at all stages of the hiring process (for a review, see Bogen & Rieke, 2018), our stimuli were based on a case in which an algorithm developed to screen applicants' resumes was found to systematically favor male over female applicants (Dastin, 2018). Similarly, in the criminal justice domain, where algorithms are extensively deployed to assess offenders' risk of recidivism (for a review, see Desmarais et al., 2016), our stimuli were based on a case where an algorithm used to predict the risk of recidivism was allegedly found to systematically assess white defendants at a lower risk than black defendants (Angwin et al., 2016; for a rebuttal see Dieterich et al., 2016). Yet, we acknowledge that in our studies, respondents were exposed to simple scenarios aimed to resemble what people may be exposed to via everyday news. Future research is needed to explore the robustness of our results in field settings.

Fifth, we suggested that our findings point to an important societal implication, namely, the risk that by hindering perceptions of bias, algorithmic decision-making might reinforce stereotypes and make people less likely to take actions against decisions that might be discriminatory. For example, the results of Study 4 suggest that algorithmic decisions may foster stereotypical beliefs, in that people might be more likely to rationalize disparities as a reflection of actual differences in dispositions and abilities, when disparities stem from algorithmic rather than human decisions. Similarly, the results of Study 6 suggest that people might be less willing to take action against alleged discrimination when algorithms, rather than humans, make decisions. However, we acknowledge that our empirical investigation is only the first step toward investigating these important societal implications. More research is needed to explore these and other responses to disparities generated by algorithmic as opposed to human decisions.

Finally, it is important to note that even if people perceive algorithmic decisions to be less biased than human decisions, they might not discount the role of institutional discrimination. To illustrate, even if people think that algorithms are less biased than humans in making university admissions decisions, they may still believe that institutional discrimination makes it more difficult for marginalized students to achieve the standards set for admission. In fact, the perceived impartiality of algorithms might make institutional discrimination even more salient. Future research is needed to explore whether algorithmic decision-making impairs or magnifies the role of institutional discrimination, and whether this impacts support for affirmative action.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica. www.propublica.org/article/machine-bias-risk-assessments-in-criminalsentencing

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671–732.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). *Consumer-lending discrimination in the FinTech Era*. National Bureau of Economic Research, Working Paper No 25943. https://doi.org/10.3386/w25943

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk

and high-cost patients. *Health Affairs*, *33*, 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041

Bogen, M., & Rieke, A. (2018). *Help wanted: An exploration of hiring algorithms, equity, and bias*. Technical report. Upturn.

Bonezzi, A., & Ostinelli, M. (2020). *Can algorithms legitimize discrimination?* https://doi.org/10.17605/OSF.IO/276GM

Castelo, N., Bos, M., & Lehmann, D. (2019). Task-dependent algorithm aversion. *JMR, Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Corcoran, K., Pettinicchio, D., & Young, J. (2015). Perceptions of structural injustice and efficacy: Participation in low/moderate/high cost forms of collective action. *Sociological Inquiry*, *85*, 429–461. https://doi.org/10.1111/soin.12082

Crosby, F. J. (1993). Why complain? *Journal of Social Issues*, *49*, 169–184. https://doi.org/10.1111/j.1540-4560.1993.tb00916.x

Cundiff, J. L., & Vescio, T. K. (2016). Gender stereotypes influence how people explain gender disparities in the workplace. *Sex Roles: A Journal of Research*, *75*, 126–138. https://doi.org/10.1007/s11199-016-0593-2

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, *17*, 4691–4697.

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

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674. https://doi.org/10.1126/science.2648573

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments in US correctional settings. *Psychological Services*, *13*(3), 206–222. https://doi.org/10.1037/ser0000075

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Technical report. Northpointe.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Earl, J. (2006). Pursuing social change online: The use of four protest tactics on the internet. *Social Science Computer Review*, *24*, 362–377. https://doi.org/10.1177/0894439305284627

Earl, J. (2004). The cultural consequences of social movements. In D. A. Snow, S. A. Soule, & H. Kriesi (Eds.), *The Blackwell companion to social movements* (pp. 508–530). Blackwell.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Ferguson, C. J. (2009), An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538.

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests—An analysis of articles published in 2015. *Journal of Experimental Social Psychology*, *75*, 95–102. https://doi.org/10.1016/j.jesp.2017.11.008

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Gates, S. W., Perry, V. G., & Zorn, P. M. (2002). Automated underwriting in mortgage lending: Good news for the underserved? *Housing Policy Debate*, *13*(2), 369–391. https://doi.org/10.1080/10511482.2002.9521447

Geiser, S. (2015). *The growing correlation between race and SAT scores: New findings from California*. University of California, Berkeley Center for Studies in Higher Education.

Gillis, T., & Spiess, J. (2019). Big data and discrimination. *The University of Chicago Law Review. University of Chicago. Law School*, *89*(2), 459–488.

Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. *Journal of Management Information Systems*, *35*, 220–265. https://doi.org/10.1080/07421222.2018.1440766

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *16*, 2125–2126. https://doi.org/10.1145/2939672.2945386

Hao, K. (2019). *This is how AI bias really happens—and why it's so hard to fix*. MIT Technology Review. www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*, 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Hayes, A. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. The Guilford Press.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018b). Algorithmic fairness. *American Economic Association Papers and Proceedings*, *108*, 22–27.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018a). Discrimination in The age of algorithms. *The Journal of Legal Analysis*, *10*, 113–174. https://doi.org/10.1093/jla/Laz001

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2020). Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30096–30100. https://doi.org/10.1073/pnas.1912790117

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. https://arxiv.org/abs/1609.05807v2

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How we analyzed the COMPAS recidivism algorithm*. ProPublica. www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., Bonezzi, A., & Morewedge, C. (2019). Resistance to medical artificial intelligence. *The Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Loughnan, S., & Haslam, N. (2007). Animals and Androids: Implicit Associations between Social Categories and Nonhumans. *Psychological*

Science, *18*(2), 116–121. https://doi.org/10.1111/j.1467-9280.2007.01858.x

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000

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

Nissenbaum, H., & Walker, D. (1998). Will computers dehumanize education? A grounded approach to values at risk. *Technology in Society*, *20*, 237–273. https://doi.org/10.1016/S0160-791X(98)00011-6

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press. https://doi.org/10.2307/j.ctt1pwt9w5

O'Neil, C. (2016). *Weapons of math destruction: How Big data increases inequality and threatens democracy*. Crown Publishers.

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

Operario, D., & Fiske, S. T. (2001). Ethnic identity moderates perceptions of prejudice: Judgments of personal versus group discrimination and subtle versus blatant bias. *Personality and Social Psychology Bulletin*, *27*, 550–561. https://doi.org/10.1177/0146167201275004

Pangburn, D. J. (2019). *Schools are using software to help pick who gets in. What could go wrong?* Fast Company. www.fastcompany.com/90342596/schools-are-quietly-turning-toai-to-help-pick-who-gets-in-what-could-go-wrong

Puente, M. (2019). *LAPD official behind controversial data programs to retire after winning lucrative contract*. https://www.baltimoresun.com/La-me-sean-malinowski-predictive-policing-20190508-story.html

Reyna, C. (2000). Lazy, dumb, or industrious: When stereotypes convey attributional information in the classroom. *Educational Psychology Review*, *12*, 85–110. https://doi.org/10.1023/A:1009037101170

Schwartz, O. (2019). *Untold history of AI: Algorithmic bias was born in the 1980s*. IEEE Spectrum. https://spectrum.ieee.org/tech-talk/tech-history/dawn-of-electronics/untold-history-of-ai-the-birth-of-machine-bias

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., & Lockhart, J. W. (2020). Diagnosing gender bias in image recognition systems. *Socius: Sociological Research for a Dynamic World*, *6*, 1–17. https://doi.org/10.1177/2378023120967171

Sloan, R. H., & Warner, R. (2018). When is an algorithm transparent? Predictive analytics, privacy, and public policy. *IEEE Security and Privacy*, *16*(3), 18–25. https://doi.org/10.1109/MSP.2018.2701166

Spring, V. L., Cameron, C. D., & Cikara, M. (2018). The upside of outrage. *Trends in Cognitive Sciences*, *22*, 1067–1069. https://doi.org/10.1016/j.tics.2018.09.006

Yzerbyt, V., Rocher, S., & Schadron, G. (1997). Stereotypes as explanations: A subjective essentialistic view of group perception. In R. Spears, P. J. Oakes, N. Ellemers, & S. A. Haslam (Eds.), *The social psychology of stereotyping and group life* (pp. 20–50). Blackwell.

Received April 20, 2020
Revision received December 14, 2020
Accepted December 24, 2020 ■