‘International Journal of Communication 18(2024), 677–699 1932–8036/20240005

Questioning Artificial Intelligence: How Racial Identity Shapes the Perceptions of Algorithmic Bias

SOOJONG KIM♦
University of California Davis, USA

JOOMI LEE
University of Georgia, USA

POONG OH
Nanyang Technological University, Singapore

There is growing concern regarding the potential for automated decision making to discriminate against certain social groups. However, little is known about how the social identities of people influence their perceptions of biased automated decisions. Focusing on the context of racial disparity, this study examined if individuals’ social identities (White vs. people of color [POC]) and social contexts that entail discrimination (discrimination target: the self vs. the other) affect the perceptions of algorithm outcomes. A randomized controlled experiment (*N* = 604) demonstrated that a participant’s social identity significantly moderated the effects of the discrimination target on the perceptions. Among POC participants, algorithms that discriminate against the subject decreased their perceived fairness and trust, whereas among White participants, the opposite patterns were observed. The findings imply that social disparity and inequality and different social groups’ lived experiences of the existing discrimination and injustice should be at the center of understanding how people make sense of biased algorithms.

Keywords: automated decision making, artificial intelligence, race, discrimination, bias, fairness, trust, emotion

With advances in artificial intelligence (AI), the use of automated decision making (ADM) has been increasing in various domains, including news creation and recommendation (Diakopoulos & Koliska, 2017; Thurman, Moeller, Helberger, & Trilling, 2019), health care and medical diagnosis (Jha & Topol, 2016; Yu & Kohane, 2019), and policing and law enforcement (Kennedy, Caplan, & Piza, 2011; Nissan, 2017). Although people generally expect algorithms to outperform human decision making in some aspects (e.g., Grace, Salvatier, Dafoe, Zhang, & Evans, 2018; Grady, 2020; Kahng, 2021), there have been growing concerns

Soojong Kim: sjokim@ucdavis.edu
Joomi Lee: joomi.lee@uga.edu
Poong Oh: poongoh@ntu.edu.sg
Date submitted: 2022-11-19

about potentially biased and inaccurate outcomes produced by machines, which may contribute to perpetuating and exacerbating the existing inequality and discrimination in society (Koenecke et al., 2020; Obermeyer, Powers, Vogeli, & Mullainathan, 2019; Williams, Brooks, & Shmargad, 2018). For social scientists, engineers, policy makers, and journalists in this critical time of socio-technological changes, it is important to understand the reaction to and perception of the new technology in society and the public's concern and skepticism against it (Dietvorst, Simmons, & Massey, 2014; Dolata, Feurriegel, & Schwabe, 2021; Logg, Minson, & Moore, 2019).

Here, we investigate how the public perceives ADM processes that discriminate against users based on their social identities. Although recent studies have demonstrated that possibilities of algorithmic biases can undermine individuals' perceptions of fairness and trust (Araujo, Helberger, Kruikemeier, & de Vreese, 2020; Lee & Baykal, 2017), not much research has been done on how different social groups process biased algorithms cognitively and emotionally. The current research explores this issue focusing on racial identities. Because racial minorities undergo discrimination and prejudice against them that affect various aspects of their lives, including their economic status, social relationships, job performance, and educational achievement (e.g., Broman, Mavaddat, & Hsu, 2000; Brown et al., 2000; Greenhaus, Parasuraman, & Wormley, 1990), the way they perceive, react to, and cope with racial disparities is different from that of the majority group (Jacob et al., 2022; Sellers & Shelton, 2003). Although earlier findings indicate that beliefs in the fairness of algorithms lead racial minorities to prefer ADM over human decision making (Bigman, Yam, Marciano, Reynolds, & Gray, 2021; Bonezzi & Ostinelli, 2021), questions remain regarding social situations wherein people confront algorithms that produce discriminatory outcomes against certain social groups.

In this regard, the present research examines if people show significant differences when they face automated decisions that discriminate against themselves compared with those discriminating against others in a different social group. To understand how algorithmic biases impact different dimensions related to experiencing discrimination, we compared perceived fairness, trust, perceived pervasiveness, the tendency to question the ADM process, and negative emotional responses to an outcome across two social situations that entail different directions of bias: Discrimination against the self and others. We also tested whether cognitive and emotional responses to disparities in ADM differ across people's social identities (racial minority and majority) and different decision-making contexts. Finally, we discuss how algorithmic biases could heighten people's concerns and aversion to algorithms and what the implications of this dynamic are.

Literature Review

Perceptions of ADM

ADM refers to computational processes of decision making that involve the use of data and algorithms (Araujo et al., 2020; Newell & Marabelli, 2015). The concepts of ADM and algorithms are particularly useful for social science research because they allow researchers to capture and represent the perspectives of most technology users who may not fully comprehend detailed procedures and mechanisms within decision-making machines. This is one reason why these concepts are attracting a growing number of researchers in various fields (e.g., Araujo et al., 2020; Bigman et al., 2021; Lee, 2018; Newell & Marabelli,

2015). Additionally, these overarching concepts are beneficial for scientific research aiming to investigate the social applications and implications of machine learning and AI that are evolving rapidly and generating new concepts and terms at a remarkable pace.

Focusing on factors influencing people's preference and avoidance of ADM, previous research identified general appreciation and reliance on algorithmic advice among lay people (Logg et al., 2019; Thurman et al., 2019). These attitudes, however, can be easily converted to algorithm aversion and preference for human-made decisions after people experience algorithm errors (Dietvorst et al., 2014). Individuals with expertise, who tend to have higher confidence in their own judgments, also prefer their self-judgment over advice given by machines (Logg et al., 2019). Overall, individuals' perceptions of their self-characteristics as well as personal experiences of ADM are important determinants of algorithm appreciation and aversion.

Recent approaches have further inspected detailed cognitive dimensions that are closely linked to algorithm appreciation and avoidance. Perceived fairness of algorithms, trust in ADM, and emotional reactions to algorithm outcomes have been identified as some of the most critical aspects of human perceptions related to ADM (Araujo et al., 2020; Lee, 2018; Lee & Baykal, 2017; Lee & Rich, 2021; Wang, Harper, & Zhu, 2020). First, perceived fairness depends on whether the algorithm treats everyone equally, independent of biases or personal preferences (Lee, 2018). The concept of trust has been adapted from interpersonal contexts, which refers to a psychological state in which people have the intention to accept other people's behavior with positive expectations. For technological artifacts that possess human-like characteristics, trust is considered fundamental to human-machine relationships (Araujo et al., 2020; Lee, 2018; Mcknight, Carter, Thatcher, & Clay, 2011). For technologies that are expected to function fairly and accurately, trust is associated with functional aspects of the outcome, such as functionality, reliability, and usefulness of ADM (Choung, David, & Ross, 2022).

Perceptions of fairness and trust are also closely related to emotional responses to algorithm outcomes (Lee, 2018). In social situations, attribution to intentionality and agency in one's behavior is a key determinant of emotional reactions to that behavior (Betancourt & Blair, 1992). Violation of equality, in particular, tends to induce negative emotional responses. When equality is violated, individuals seek further explanations about the intention of the violator, and strong negative reactions (e.g., anger) can be evoked if the violation is deemed intentional (Shaver, 1985). When the violation is considered unintentional, trust in the violator can mitigate negative emotional reactions (Stouten, De Cremer, & van Dijk, 2006).

Similar attribution processes may occur during the evaluation of algorithms and their outcomes. Trust in AI was found to predict positive attitudes and emotional attachment toward AI technologies (Choung et al., 2022). On the other hand, Lee (2018) tested whether people would show less emotional reactions to ADM compared with human decision making due to the perceived lack of intentionality and agency in ADM. However, the result showed similar or even more negative reactions to ADM. These findings imply that cognitive appraisals of perceived fairness, trust, as well as emotional reactions may alter attributional processes involved in algorithm appreciation or aversion.

Social Disparities in ADM and Algorithmic Bias Perceptions

Although ADM aims to offer a cost-efficient, accurate, and objective alternative to potential inaccuracy and inconsistency in human decision making, it can still generate unfair outcomes that often reflect and even amplify existing inequality and injustice in society (Barocas, Hardt, & Narayanan, 2019; Hooker, 2021). While deliberate corrections and interventions are necessary to mitigate algorithmic biases, examining cognitive responses to the biases is also crucial in understanding the consequences of discrimination and injustice that machines could engender and in envisioning better socio-technological systems for the future (Benjamin, 2019; Noble, 2018; O'Neil, 2016).

Scholars have been paying increasing attention to the connection between the perception of ADM and racial disparity (Bigman et al., 2021; Parra, Gupta, & Dennehy, 2021). Past research suggests that existing racial inequality in society shapes the reactions to and perceptions of algorithms (Eubanks, 2018; Koenecke et al., 2020; Vincent & Viljoen, 2020). For example, a study found that the threat of racial inequality increases individuals' preference for ADM, particularly among minority populations (Bigman et al., 2021). It implies that the belief that nonhuman agents make fairer and less biased judgments can promote the acceptance of machine-driven decisions. Another study also found that people are less likely to recognize racial or gender disparities in ADM than in human decisions when assuming algorithms are fair (Bonezzi & Ostinelli, 2021). Empirical evidence also revealed that situations, where algorithm processes may neglect the unique characteristics of individuals and generate disadvantageous outcomes against them, can stimulate resistance to ADM (Longoni, Bonezzi, & Morewedge, 2019). These findings suggest that, without contextual information signaling potential algorithmic biases, people are less likely to cast doubt on ADM.

Social Identity, Social Disparity, and the Perception of ADM

The literature on social perception and attribution suggests that bias perceptions in social contexts are largely influenced by both the direction of bias (i.e., favorable or unfavorable to whom) and the perceiver's social identity (Major, Quinton, & McCoy, 2002; Phillips & Jun, 2022). Individuals make asymmetrical inferences about themselves versus others (Pronin, Gilovich, & Ross, 2004). It was found that when people make judgments about themselves relative to others, they view themselves as less biased and objective than others and more readily detect biases in others.

Importantly, the attribution to discrimination framework (Major et al., 2002) suggests that individuals attribute their negative life experiences to discrimination when (1) they (or their group) are treated unjustly/unequally, and (2) the disparity is based on social identity. Furthermore, attribution to discrimination can be reinforced by personal experiences of being targeted by discrimination and prejudice. When it comes to the perception of discrimination in ADM, the literature provides little knowledge about how social identities play a role during the attribution processes related to ADM. Even without the consideration of social identities, past research showed mixed findings on the effect of the favorability of the outcome, or perceptions of favorability. For example, some studies found a significant influence of favorable algorithms on perceived fairness (e.g., Wang et al., 2020), while others report marginal or null effects (e.g., Li & Xing, 2022).

The current study predicts that, in the social context of race, social identity shapes the reactions to racially biased automated decisions. Racial minorities experience discrimination and injustice in their daily lives, which in turn can lead to various deleterious outcomes, including economic disadvantages (Hangartner, Kopp, & Siegenthaler, 2021; Pager & Shepherd, 2008; Rosen, Garboden, & Cossyleon, 2021), physical and psychological distress (Berger & Sarnyai, 2015; Broman et al., 2000; Brown et al., 2000), and the degradation of job and academic performances (Greenhaus et al., 1990; Stevens, Liu, & Chen, 2018; Wong, Eccles, & Sameroff, 2003). How people experience and cope with racial discrimination depends on the social identities of individuals (Jacob et al., 2022; Sellers & Shelton, 2003).

The current study adopts the concept of "perceived pervasiveness" to capture how social groups perceive the degree to which discriminatory algorithm outcomes are pervasive and likely to occur in their everyday lives. Minority groups frequently experience negative events that involve discrimination and injustice against their social identities (Major et al., 2002; Sue, 2010). Consequently, these populations tend to be more vigilant and perceptive toward indications of biased decision making that may suggest discrimination against them (Major et al., 2002). It is possible that racial minorities would report greater increases in perceived pervasiveness when ADM produces biased decisions against them compared with members of the racial majority. Such perceptions can influence their attitudes toward algorithms and AI, their support for related policies, and other aspects of their lives (Schmitt, Branscombe, & Postmes, 2003; Stroebe, Dovidio, Barreto, Ellemers, & John, 2011).

Thus, we hypothesized that minority groups experiencing unfavorable automated decisions, compared with minorities facing decisions that discriminate against other social groups, would perceive that biased outcomes are more likely to happen and less fair, and the processes generating these discriminatory outcomes are less trustable. Similarly, minority groups may feel more negative emotions and be more likely to question ADM processes when automated decisions discriminate against their own social groups.

H1: Racial minorities will show greater reductions in (a) perceived fairness and (b) trust in ADM when ADM generates decisions biased against them compared with the racial majority.

H2: Racial minorities will show greater increases in (a) perceived pervasiveness, (b) negative emotion, and (c) tendency to question an outcome when ADM generates decisions biased against them compared with the racial majority.

Previous research investigated the public's ability to recognize and cast doubt on biased decisions made by algorithms (Parra et al., 2021), but questions remain regarding how the magnitudes of reactions vary depending on social situational contexts. People are generally averse to resorting to ADM regarding essential matters of everyday life, such as important health or financial issues (Longoni et al., 2019). Prior work suggests that situations involving explicit economic disadvantages or advantages (e.g., monetary incentives and produce prices) may increase the sensitivity to the fairness and reliability of a decision-making process (Esarey, Salmon, & Barrilleaux, 2012; Moliner, Martínez-Tur, Peiró, Ramos, & Cropanzano, 2013). Even if individuals experience the same situations involving ADM in which they are discriminated against their race, the interpretation of the incidences may differ between White people and people of color (POC).

H3: *People will show lower (a) perceived fairness and (b) trust in ADM when an ADM outcome is accompanied by explicit economic consequences compared with when the outcome is not connected with explicit economic consequences.*

H4: *People will show greater (a) negative emotion and (b) tendency to question the outcome when an ADM outcome is accompanied by explicit economic consequences compared with when the outcome is not connected with explicit economic consequences.*

Method

Sample

Participants were recruited online via Prolific, a survey recruitment platform. People who were at least 18 years old and located in the United States received a digital flyer about the experiment. People could click on a link in the flyer to access the online experiment created with Qualtrics, a survey managing system.

In total, 658 participants completed the experiment. Among these, 54 failed to follow the instruction or did not pass the attention check, and their responses were removed from the results. The remaining 604 participants were analyzed in this research. For analytical purposes, we categorized all participants into two social (race) groups: The White group (n = 442, 73.2%) and the POC group (n = 162, 26.8%). Participants who identified themselves as a man, woman, nonbinary, and others ("Prefer not to disclose" and "Prefer to self-describe") accounted for 42.7%, 53.8%, 3.0%, and 0.5%, respectively.

Procedure

A randomized controlled experiment was conducted to test the research hypotheses. Participants' informed consent was obtained on the first webpage of the experiment. On the next webpage, participants were instructed to think of one of their friends whose race is different from theirs. The instruction also stated that the friend should be a person of the same gender, age, education level, and economic status as the participant. They were asked to keep thinking of the friend they chose when considering the scenarios that followed. To check if participants followed the instruction correctly, a question was asked about the race of the chosen friend.

Nine scenarios were given to each participant, one scenario at a time. Multiple scenarios allowed us to compare and analyze their perspectives across different situations. The number of scenarios and questions was intentionally limited to keep the online experiment duration within an optimal range of 10 to 15 minutes, as recommended to prevent participant fatigue, ensure data quality, and meet participant expectations (Dynata, 2023; Revilla & Ochoa, 2017). Each scenario described a distinct situation in which the participants and their friends used an identical technology involving ADM, which resulted in a racially discriminatory outcome. The scenarios described realistic situations that have been observed in the real world and discussed in previous studies (Acikgoz, Davison, Compagnone, & Laske, 2020; Binns et al., 2018; Miller & Hosanagar, 2019; Parra et al., 2021). All scenario descriptions followed the narrative format proposed by Parra and colleagues (2021). The domains discussed in the scenarios included finance, the labor market, public service, media, and health and safety, as listed in Table 1.

Table 1. Scenarios Used in the Experiment.

A

Index	Domain	Scenario Given in the Self Condition
1	Finance	"You and your friend are both applying for the same financial product (such as a credit card, a personal loan, and a mortgage) on the same banking website. The website collects information about its users and automatically evaluates applications based on the information. "The products that are offered to you charge higher interest rates than those offered to your friend."
2	Labor market	"You and your friend are both looking for similar jobs on the same website. The website collects information about its users and automatically recommends job positions based on the information. "The positions that the website recommends to you offer lower salaries than those recommended to your friend."
3	Information technology	"You and your friend use the same smartphone model and the same voice-recognition system (e.g., Siri, Alexa, or Google Assistant). Both of you use the voice-recognition system routinely for everyday use. "You notice that the voice recognition system fails to recognize your voice more frequently than it fails with your friend's voice."
4	Labor market	"You and your friend applied for the same job position in which you both are seriously interested. You both have similar levels of experience, knowledge, and skills related to the position. "Each of you conducted an online interview with the company and was asked the same questions by an automated interview program. Applications and recorded responses are automatically rated by the program. You noticed that you and your friend did equally well in the interview and provided similar answers to the interview questions. "Two weeks after the interview, your friend is offered the position, but you do not receive any offer."
5	Public service	"You and your friend have the same nationality. You both are going through an automated immigration kiosk at an airport. The kiosk uses face recognition technology to verify travelers' identities. "The kiosk directs you to see an immigration officer and provide further information while your friend is cleared to go through."
6	Hospitality	"You and your friend are each booking a similar hotel room using the same travel booking website. The website collects information about its users and automatically recommends hotel rooms based on the information. "The website shows you fewer available rooms than it does for your friend."
7	Media	"You and your friend both regularly write posts on similar topics on the same online social network platform (e.g., Facebook, Instagram, and TikTok). The platform automatically examines posts created by users and identifies objectionable content that needs to be flagged or removed.

		"Your posts are found objectionable by the platform more frequently than those written by your friend."
8	Health and safety	"You and your friend both have similar diets and daily routines and are feeling just fine. Both of you are using the same health assessment app. The app collects information about its users and automatically estimates the risk of infectious diseases based on the information. "The app suggests that your risk of contracting infectious diseases is higher than your friend's."
9	Health and safety	"You and your friend live in the same neighborhood and have similar cars and driving patterns. Both of you are purchasing the same car insurance plan from an insurance company. "The company uses a program that automatically calculates each customer's insurance premium based on information about the customers they collect. "The company charges you more money than it charges your friend for the same insurance plan."

B

	Scenario 1 in the Self Condition	Scenario 1 in the Other Condition
Example (Scenario 1)	"The products that are offered to you charge higher interest rates than those offered to your friend."	"The products that are offered to your friend charge higher interest rates than those offered to you."

Participants were randomly assigned to one of the two conditions with different discrimination targets: Discrimination against the self (the "self" condition) and discrimination against the other (the "other" condition). Participants in both conditions read the same scenarios, and only the target of discrimination varied depending on the condition. The scenarios in the self condition explained situations where an outcome disadvantages the participant compared with the friend. The example scenario quoted above was for the self condition. In the other condition, participants were given scenarios in which an outcome disadvantages the friend. For example, Scenario 1 in Table 1 described the identical situation in the other condition but ended with a different statement: "The products that are offered to your friend charge higher interest rates than those offered to you." The order of presenting the scenarios was randomly determined for each participant. In each of the two conditions, 302 participants were assigned.

Participants answered a set of questions after reading each scenario. These questions measured perceived fairness, the tendency to question the outcome, and the likelihood of each situation in real life. Participants could start responding to these questions after reading a scenario for at least 10 seconds. An instructed response item was included as an attention check item in the middle of the experiment (Gummer, Roßmann, & Silber, 2021). At the end of the experiment, participants were asked about their demographic characteristics. The median duration of the entire experimental process was 672 seconds. Subjects received monetary compensation for their participation. The current research was exempted by the Institutional Review Board of the University of California Davis. The experiment was conducted in August 2022.

Measures

All variables were measured on a 7-point Likert scale ranging from 1 to 7. Perceived fairness (M = 3.1, SD = 1.1) was measured with a question ("How fair or unfair is this outcome for you?") in line with previous studies (Araujo et al., 2020; Lee, 2018). Trust in ADM (M = 2.7, SD = 1.0) was measured using a question ("How do you trust that this [website/technology/program/platform/app] makes a good-quality decision?") adapted from previous research (Lee, 2018). Negative emotion (M = 4.6, SD = 1.3, α = 0.93) was measured with three questions (Lee, 2018). Tendency to question the outcome (M = 5.0, SD = 1.1, α = 0.91) was operationalized by a measure of a participant's agreement with two statements ("This outcome is problematic" and "This outcome is questionable"). Perceived pervasiveness (M = 3.9, SD = 1.2) was measured with a question ("How likely is this outcome to happen in your everyday life?"). For the main outcome analysis, each dependent variable was averaged across all nine scenarios for each participant.

Statistical Analyses

Multiple linear regressions examined the effect of the discrimination target on each dependent variable, moderated by a participant's race group. We first evaluated the coefficients and statistical significance of the interaction between the discrimination target and the race group. Additional subgroup analyses were also conducted to investigate the effect of the discrimination target in each race group (POC and White) for further investigations of moderation effects on each dependent variable. Although dividing participants into each subgroup inevitably limits the statistical power of analysis, subgroup analyses could still provide helpful insights into individuals' perceptions within each race group and their specific contexts.

In addition, we compared the marginal means of the outcome variables in each scenario and each condition. For this analysis, we estimated a new statistical model that includes the scenario as a predictor. The analysis estimated a random effect linear regression model that predicts a dependent variable as a function of the scenario, the discrimination target, the race group, and the interaction between the target and the race group, accounting for the correlation within a subject. The estimated marginal mean of a dependent variable was then calculated for each scenario in each condition based on the estimated random effect model.

All statistical analyses were conducted on an open-source statistical software, R (version 4.0.3).

Results

Discrimination Target × Race Interaction

Perceived Fairness

The influence of discrimination against the self on perceived fairness, compared with discrimination against the other, was significantly different between POC and White participants. As Table 2 presents, the interaction between the discrimination target and the race group on the average perceived fairness was significant (B = −0.531, SE = 0.200, p = .008). To further understand the nature of the interaction, we then analyzed responses from POC and White participants separately. The subgroup analysis indicated that

the reactions of POC and White participants were opposite. Among POC, the average perceived fairness was lower in the self condition than in the other condition (B = −0.483, SE = 0.168, p = .005), while the average perceived fairness among the White group was higher in the self condition than the other condition (B = 0.048, SE = 0.105, p = .645).

Even when perceived fairness was analyzed for each scenario, a consistent pattern was observed. As displayed in Figure 1, despite some variations across the scenarios, the signs of the discrimination target × the race group interactions were negative in all nine scenarios, and statistical significance was found for four of them. In Figure 1, a dependent variable is shown on the y-axis. Each dependent variable was measured for a single scenario (e.g., "Perceived fairness 5" is perceived fairness measured in Scenario 5). The x-axis represents experimental conditions. The red and blue lines represent POC and White participants, respectively. The dots indicate averages, and an error bar represents the standard error of a mean. The coefficient and the p value of the target × race interaction effect on each outcome are displayed inside the plot. Statistical significance was indicated with *** (p < .001), ** (p < .01), and * (p < .05).

Table 2. The Interaction Effect Between the Discrimination Target and the Race Group of the Subject on Experimental Outcomes.

	Avg. Perceived Fairness	Avg. Trust in ADM	Avg. Tendency to Question	Avg. Negative Emotion	Avg. Perceived Pervasiveness
All Participants	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)
Target self × race	−0.531	−0.674	1.155 (0.195)	1.021 (0.243)	0.928
POC	(0.200) **	(0.173) ***	***	***	(0.215) ***
Target self	0.048	0.278 (0.089)	−0.683	−0.231	−0.819
	(0.104)	**	(0.101) ***	(0.126)	(0.111) ***
Race POC	0.218	0.319 (0.121)	−0.496	−0.566	−0.353
	(0.140)	**	(0.137) ***	(0.170) ***	(0.150) *
Constant	3.091	2.566 (0.064)	5.373 (0.072)	4.756 (0.090)	4.317
	(0.074) ***	***	***	***	(0.079) ***
POC participants	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)
Target self	−0.485	−0.396	0.473 (0.163)	0.790 (0.209)	0.110
	(0.168) **	(0.147) **	**	***	(0.191)
Constant	3.310	2.885 (0.102)	4.888 (0.112)	4.190 (0.145)	3.964
	(0.117) ***	***	***	***	(0.133) ***
White participants	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)
Target self	0.048	0.278 (0.090)	−0.683	−0.231	−0.819
	(0.105)	**	(0.102) ***	(0.126)	(0.110) ***
Constant	3.091	2.566 (0.064)	5.373 (0.073)	4.756 (0.090)	4.317
	(0.074) ***	***	***	***	(0.078) ***

Note. *** p < .001, ** p < .01, * p < .05, The experimental outcomes were averaged across the nine scenarios. Coefficients *(B*) are unstandardized. Standard errors in parentheses.

Trust in ADM

The influence of discrimination against the self on trust in ADM, compared with discrimination against the other, also showed a significant difference between POC and White participants. The interaction between the discrimination target and the race group on trust in ADM was significant ($B = -0.674$, $SE = 0.173$, $p < .001$) as Table 2 presents. When the POC and White subgroups were analyzed separately, the reactions of the two groups were opposite. Among POC participants, trust was lower in the self condition than in the other condition ($B = -0.396$, $SE = 0.147$, $p = .008$), while trust among White people was higher in the self condition than in the other condition ($B = 0.278$, $SE = 0.090$, $p = .002$). When each scenario was analyzed separately, as visualized in Figure 1, the aforementioned pattern was consistent: A negative sign for the target × the race group interaction in all nine scenarios. Statistical significance was found for seven of the nine scenarios.

Tendency to Question the Outcome

Regarding the tendency to question ADM outcomes, the influence of the discrimination target varied depending on the race group. Table 2 shows that the interaction between the discrimination target and the race group on the average tendency to question ADM outcomes was significant ($B = 1.155$, $SE = 0.195$, $p < .001$). When we looked into the POC and White subgroups separately, the responses of POC and White participants displayed striking contrast. Among POC participants, the average tendency to question was greater in the self condition than in the other condition ($B = 0.473$, $SE = 0.163$, $p = .004$), while among White participants it was lower in the self condition than in the other condition ($B = -0.683$, $SE = 0.102$, $p < .001$). As another robustness check, the tendency to question ADM outcomes was calculated separately for each scenario. The result indicates that despite minor variations, the discrimination target × the race group interactions were positive and statistically significant in all nine scenarios.

Negative Emotion

There was a difference in the influence of the discrimination target on the average negative emotion between White and POC participants: The interaction between the discrimination target and the race group on the average negative emotion was significant ($B = 1.021$, $SE = 0.243$, $p < .001$). The subgroup analysis showed that the average negative emotion of the two groups, the White and the POC groups, shifted in opposite directions. Table 2 presents that while among POC participants discrimination against the self significantly increased the average negative emotion, compared with discrimination targeting others ($B = 0.790$, $SE = 0.209$, $p < .001$), discrimination against the self reduced the average negative emotion among White participants although it was only marginally significant ($B = -0.231$, $SE = 0.156$, $p = .067$). We also evaluated negative emotion in each scenario. The result supported the aforementioned finding: The discrimination target × the race group interactions were positive and statistically significant in eight of the nine scenarios.

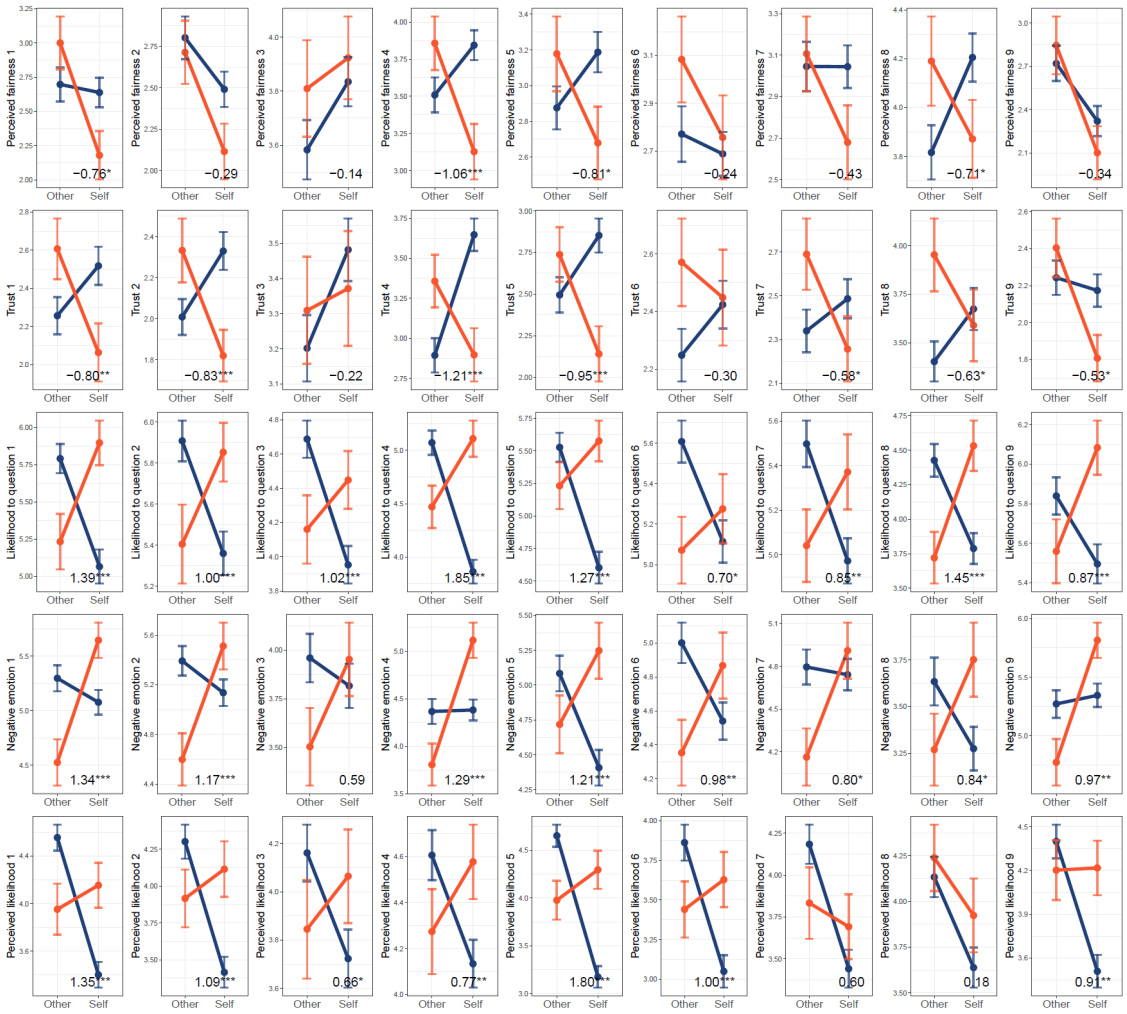

Figure 1. Responses by conditions, races, and scenarios.

Perceived Pervasiveness

In terms of the perceived pervasiveness of the outcomes in everyday life, the influence of the discrimination target was significantly different between White and POC participants. Table 2 presents that the interaction between the discrimination target and the race group on the average perceived pervasiveness of the outcome was significant (B = 0.928, SE = 0.215, p < .001). The subgroup analysis revealed what this significant interaction indicated. When the sample was divided into the two groups, the average perceived pervasiveness among POC participants did not show a significant difference between the self condition and the other condition (B = 0.110, SE = 0.191, p = .567), but the average perceived pervasiveness among White participants significantly decreased with discrimination targeting the self (B = −0.819, SE = 0.110, p < 0.001). The analysis of perceived pervasiveness in each scenario showed that the

discrimination target × the race group interactions were positive in all nine scenarios and statistically significant in seven of them.

Influence of the Friend's Race Among POC Participants

During the experiment, most POC participants (n = 94, 58.0%) chose a White friend for consideration, and the rest chose a non-White friend, for example, a Black participant chose an Asian friend. (All White participants chose a POC friend.) Thus, we conducted a robust check to examine the influence of the friend's race among POC participants.

For perceived fairness, trust in ADM, the tendency to question the outcome, and negative emotion, discrimination against the self changed the dependent variables in the same direction regardless of the friend's race, as visualized in Figure 2. In this figure, a dependent variable is shown on the y-axis. The x-axis represents experimental conditions. The green and blue lines represent participants who chose a POC friend and a White friend, respectively. The dots indicate averages, and an error bar represents the standard error of a mean. The coefficient and the p value of the target × race interaction effect on each outcome are displayed inside a plot, where statistical significance was indicated with *** (p < .001), ** (p < .01), and * (p < .05). As shown in Figure 2, regardless of their friend's race, POC participants reacted to the discriminatory situations targeting the self with less perceived fairness, less trust, a higher tendency to question the outcome, and higher negative emotion. The results also suggest that POC participants who considered a White friend might show stronger reactions to the discriminatory scenario than those who chose another non-White race. Regarding perceived pervasiveness, self-targeting discriminatory situations involving a White friend and a POC friend led to a higher and a lower perceived pervasiveness, respectively, compared with other-targeting discrimination.

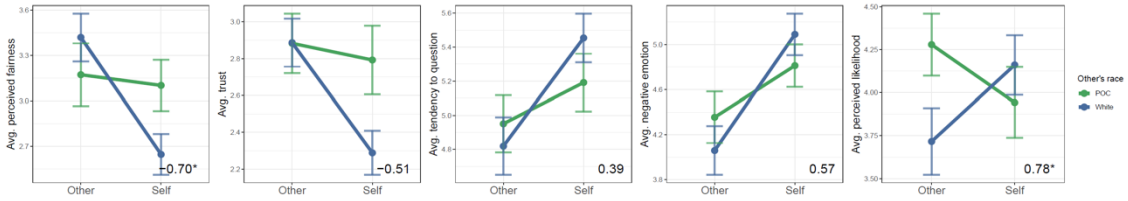

Figure 2. Average outcomes among POC participants by the friend's race (n = 162).

Marginal Means by Scenarios

Figure 3 visualizes and compares the estimated marginal means (EMMs) of the outcome variables. In this figure, the red dots indicate EMMs. The gray bars represent the 95% confidence intervals of the EMMs. A difference between two EMMs is statistically significant at α = 0.05 with the Bonferroni correction if their respective arrows ("comparison arrows") do not overlap (Lenth, 2022).

Figure 3 depicts a pattern consistent across the two experimental conditions. In both conditions, perceived fairness and trust in ADM were at their lowest level in the three scenarios about insurance premiums (Scenario 9), salaries of recommended jobs (Scenario 2), and interest rates of financial products

(Scenario 1). These three scenarios were the highest in the pervasiveness to question the outcome and negative emotion. Contrarily, the three scenarios that induced the highest perceived fairness and trust discussed the calculation of disease risk (Scenario 8), the evaluation of job interviews (Scenario 4), and failure rates in voice recognition (Scenario 3). These three scenarios were the lowest in terms of the tendency to question the outcome and negative emotion.

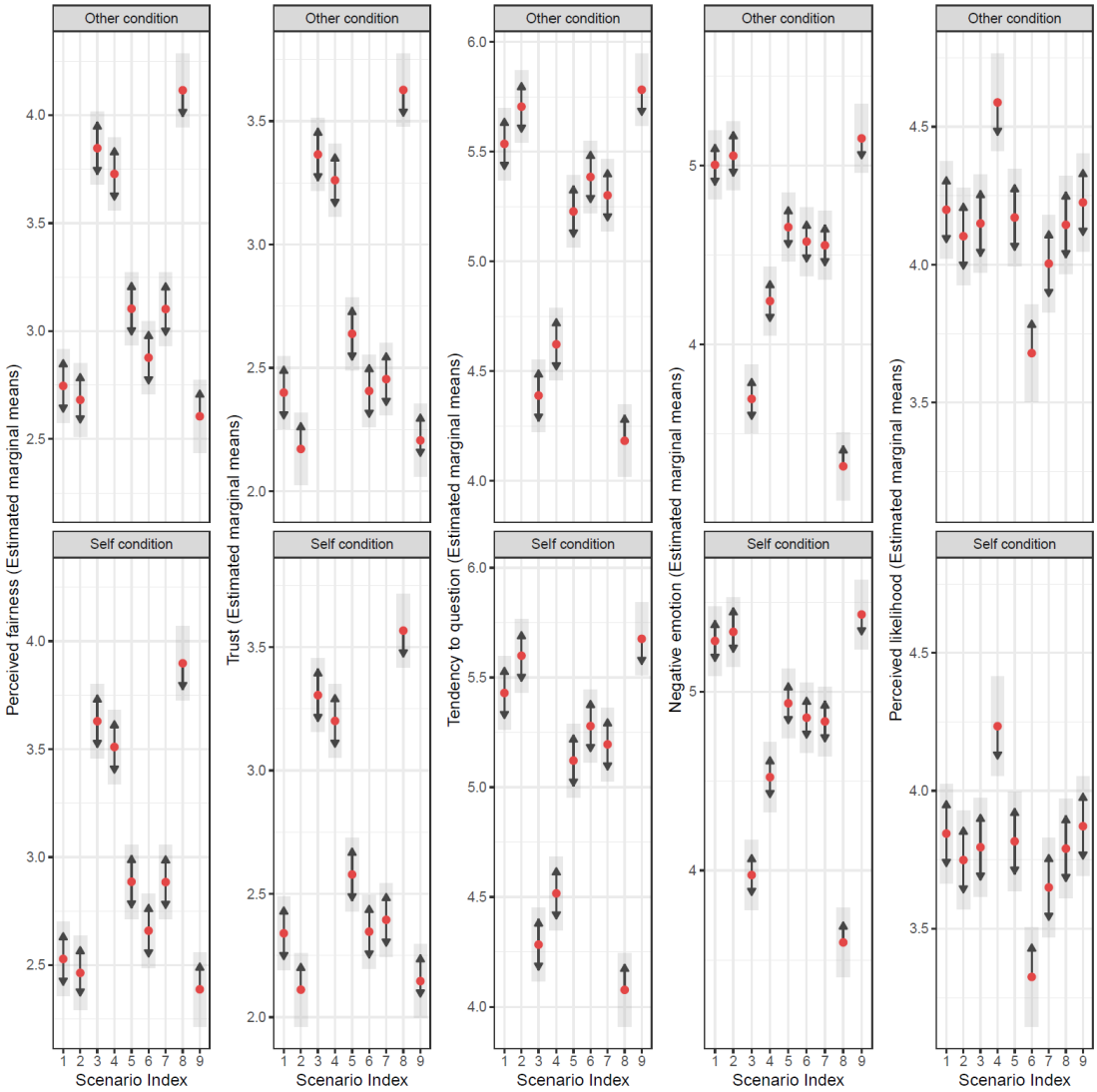

Figure 3. Estimated marginal means of outcomes by scenario and condition.

Perceived pervasiveness showed a pattern distinguished from the other variables. While most scenarios produced similar levels of perceived pervasiveness, the scenario about job interview evaluation (Scenario 4) exhibited the highest perceived pervasiveness, which is significantly different from other

scenarios' results. The perceived pervasiveness of the scenario about availability on a booking website (Scenario 6) was significantly lower than in other scenarios.

Discussion

This study examined how social identities influence the perceptions of automated decisions that discriminate against certain social groups. The results demonstrate the significant impacts of social identities that shape cognitive and affective responses to algorithmic biases. First, lending support to H1, discrimination against the self decreased POC's perceived fairness and trust in ADM compared with discrimination against the other, whereas White individuals exhibited the opposite pattern. Furthermore, as predicted in H2, discrimination unfavorable to the subject significantly increased the tendency to question the outcome, negative emotion, and perceived pervasiveness of the incidence among racial minorities, but these outcomes were lower with discrimination targeting the other among the racial majority group. The significant interactions indicate that the effects of discrimination targets were significantly different between the two racial groups. These results provide compelling new evidence that social identity plays an essential role in shaping people's cognitive and affective reactions to algorithmic bias. In contrast to the previous reports on marginal or null effects of social/group identities (Wang et al., 2020), this finding highlights the importance of social groups' innate characteristics in people's experiences of algorithms and AI (Lee & Rich, 2021).

It is worth emphasizing that these results clearly reflect what POC and White people experience and observe in real-world situations. POC participants reported that discrimination targeting themselves was more likely to happen in the real world than discrimination targeting other races. White people also responded that ADM situations were more likely to discriminate against POC people than themselves, implying that people in the majority group are also aware of discrimination against racial minorities. These results align with previous reports on racially biased automated decisions (Benjamin, 2019; Buolamwini & Gebru, 2018; Obermeyer et al., 2019; Rosen et al., 2021). Also, the results suggest that social disparity and inequality and the lived experience of the existing discrimination and injustice are at the center of understanding the perception of automated decisions impacting a large number of people. An excessive focus on mathematical fairness or one-size-fits-all approaches, without taking into account diversity, inclusion, sociocultural reality, and public perception in designing, deploying, and evaluating algorithms may lead to a "false equality" that actually exacerbates existing inequalities. A deep understanding of sociocultural factors is also essential to inform potential solutions that have been proposed to address biased AI, such as Explainable AI, which, while not a cure-all solution, could constitute one aspect of a larger approach to address the broader issue of algorithmic fairness and accountability (de Bruijn, Warnier, & Janssen, 2022).

Furthermore, this research discovered that perceived fairness and trust are lower when discriminatory outcomes accompany economic disadvantages compared with outcomes not involving explicit economic penalties, supporting H3. Contrarily, the findings also indicated that the tendency to question the outcome and negative emotion are higher in discriminatory situations involving economic disadvantages imposed on the self, supporting H4. Combined with the results presented earlier, these outputs enable us to reconcile two seemingly conflicting arguments: People experiencing biased algorithms react to economic

advantages and disadvantages, but the importance of social identities emerges when we start considering the detailed context of discrimination, such as which social group is being discriminated by algorithms.

Implications of Findings

The present findings have several important implications. First, this study reveals that social identity shapes the perceptions of ADM. While knowledge has been thin about the linkages between racial identity and racially biased automated decisions, this study provides new empirical evidence that discriminatory situations caused by machines could be perceived very differently depending on people's racial identity. It also means that research frameworks that neglect existing social inequality and injustice may provide only a partial explanation of how people process AI-augmented discrimination and how they navigate through the increasingly complex socio-technological system. Likewise, algorithmic adjustments focusing only on offering myopic economic incentives to mitigate biased outcomes will not encompass systematic disparities across different social identity groups and their situational contexts. Third, the results of this study evince people's clear reaction when one of the core principles of ADM, fairness, is violated. Our finding shows that people clearly recognize the target of discrimination and respond with sharp changes in their perceptions and emotions. It implies the possibility that repeated and large-scale exposure to algorithms that discriminate against a significant portion of a population (e.g., Buolamwini & Gebru, 2018; Koenecke et al., 2020; Obermeyer et al., 2019) could corrode public trust and credence in AI and produce considerable confusion, conflict, and disruption in society. Last, this study reveals that the context of discrimination also influences the magnitude of public reactions. We found that the perception of ADM is affected more when ADM results in an economic (dis)advantage, but it is worth noticing that reactions to noneconomic discrimination are also not ignorable.

Limitations

The present research is not without limitations. First, the participants were not a representative sample of the U.S. population, and one should be cautious in generalizing the current findings. Also, since the experiment was conducted with U.S. residents, the specific social and racial contexts of this study should be carefully interpreted while considering the current findings in other social and cultural contexts. Second, we observed individuals' reactions when they compared themselves with one of their friends, and future research should investigate the perception of biased ADM on various types of social networks, such as how algorithm bias affects connections in working or learning environments. Third, although our study categorized participants into POC and White groups, research with larger samples and more detailed analyses will be more useful in gaining a more contextualized understanding of the subgroups within the POC group, including Black, American Native, and Asian.

Conclusion

Scholars have warned that the increasing use of algorithms and AI might increase the risk of racial discrimination and exacerbate the existing disparities and inequalities in society (Koenecke et al., 2020; Miller, 2020; Obermeyer et al., 2019; Williams et al., 2018). Systematic social disparity and inequality differentially affect various aspects of individuals' daily lives depending on their social identity. The current study incorporates

individuals' social identities to understand their cognitive and affective responses to racially biased automated decisions and demonstrates that people are able to recognize when their beliefs in fairness and equality of ADM are violated, discriminating against racial minorities. Individuals' social identity is a crucial predictor that shapes the magnitude of their resistance and aversion to algorithms. Overall, findings from this study contribute empirical evidence of public perceptions and reactions related to potential algorithmic biases that are important for researchers, policy makers, and practitioners to understand the negative outcomes of algorithmic bias in society and develop fairer and equitable socio-technological systems.

References

Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment, 28*(4), 399–416. doi:10.1111/ijsa.12306

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society, 35*(3), 611–623. doi:10.1007/s00146-019-00931-w

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. Retrieved from https://fairmlbook.org

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Medford, MA: Polity Press.

Berger, M., & Sarnyai, Z. (2015). "More than skin deep": Stress neurobiology and mental health consequences of racial discrimination. *Stress, 18*(1), 1–10. doi:10.3109/10253890.2014.989204

Betancourt, H., & Blair, I. (1992). A cognition (attribution)-emotion model of violence in conflict situations. *Personality and Social Psychology Bulletin, 18*(3), 343–350. doi:10.1177/0146167292183011

Bigman, Y. E., Yam, K. C., Marciano, D., Reynolds, S. J., & Gray, K. (2021). Threat of racial and economic inequality increases preference for algorithm decision-making. *Computers in Human Behavior, 122*, 106859. doi:10.1016/j.chb.2021.106859

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. doi:10.1145/3173574.3173951

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied, 27*(2), 447–459. doi:10.1037/xap0000294

Broman, C. L., Mavaddat, R., & Hsu, S.-Y. (2000). The experience and consequences of perceived racial discrimination: A study of African Americans. *Journal of Black Psychology, 26*(2), 165–180. doi:10.1177/0095798400026002003

Brown, T. N., Williams, D. R., Jackson, J. S., Neighbors, H. W., Torres, M., Sellers, S. L., & Brown, K. T. (2000). "Being black and feeling blue": The mental health consequences of racial discrimination. *Race and Society, 2*(2), 117–131. doi:10.1016/S1090-9524(00)00010-3

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91.

Choung, H., David, P., & Ross, A. (2022). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction, 39*(9), 1727–1739. doi:10.1080/10447318.2022.2050543

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly, 39*(2), 101666. doi:10.1016/j.giq.2021.101666

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism, 5*(7), 809–828. doi:10.1080/21670811.2016.1208053

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. doi:10.1037/xge0000033

Dolata, M., Feuerriegel, S., & Schwabe, G. (2021). A sociotechnical view of algorithmic fairness. *Information Systems Journal, 32*(4), 754–818. doi:10.1111/isj.12370

Dynata. (2023). *Survey length best practices: Are shorter surveys better?* Retrieved from https://www.dynata.com/survey-length-best-practices-are-shorter-surveys-better/

Esarey, J., Salmon, T. C., & Barrilleaux, C. (2012). What motivates political preferences? Self-interest, ideology, and fairness in a laboratory democracy. *Economic Inquiry, 50*(3), 604–624. doi:10.1111/j.1465-7295.2011.00394.x

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

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research, 62*(1), 729–754. doi:10.1613/jair.1.11222

Grady, D. (2020, January 1). A.I. is learning to read mammograms. *The New York Times*. Retrieved from https://www.nytimes.com/2020/01/01/health/breast-cancer-mammogram-artificial-intelligence.html

Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. *The Academy of Management Journal, 33*(1), 64–86. doi:10.2307/256352

Gummer, T., Roßmann, J., & Silber, H. (2021). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research, 50*(1), 238–264. doi:10.1177/0049124118769083

Hangartner, D., Kopp, D., & Siegenthaler, M. (2021). Monitoring hiring discrimination through online recruitment platforms. *Nature, 589*(7843), 572–576. doi:10.1038/s41586-020-03136-0

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem." *Patterns, 2*(4), 1–4. doi:10.1016/j.patter.2021.100241

Jacob, G., Faber, S. C., Faber, N., Bartlett, A., Ouimet, A. J., & Williams, M. T. (2022). A systematic review of black people coping with racism: Approaches, analysis, and empowerment. *Perspectives on Psychological Science, 18*(2), 392–415. doi:10.1177/17456916221100509

Jha, S., & Topol, E. J. (2016). Adapting to artificial intelligence: Radiologists and pathologists as information specialists. *JAMA, 316*(22), 2353–2354. doi:10.1001/jama.2016.17438

Kahng, A. B. (2021). AI system outperforms humans in designing floorplans for microchips. *Nature, 594*(7862), 183–185. doi:10.1038/d41586-021-01515-9

Kennedy, L. W., Caplan, J. M., & Piza, E. (2011). Risk clusters, hotspots, and spatial intelligence: Risk terrain modeling as an algorithm for police resource allocation strategies. *Journal of Quantitative Criminology, 27*(3), 339–362. doi:10.1007/s10940-010-9126-2

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., . . . Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America, 117*(14), 7684–7689. doi:10.1073/pnas.1915768117

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1), 1–16. doi:10.1177/2053951718756684

Lee, M. K., & Baykal, S. (2017). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1035–1048. doi:10.1145/2998181.2998230

Lee, M. K., & Rich, K. (2021). Who is included in human perceptions of AI? Trust and perceived fairness around healthcare AI and cultural mistrust. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. doi:10.1145/3411764.3445570

Lenth, R. V. (2022). *Emmeans: Estimated marginal means, aka least-squares means*. Retrieved from https://CRAN.R-project.org/package=emmeans

Li, C., & Xing, W. (2022). Revealing factors influencing students' perceived fairness: A case with a predictive system for math learning. *Proceedings of the Ninth ACM Conference on Learning @ Scale*, 409–412. doi:10.1145/3491140.3528293

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. doi:10.1016/j.obhdp.2018.12.005

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

Major, B., Quinton, W. J., & McCoy, S. K. (2002). Antecedents and consequences of attributions to discrimination: Theoretical and empirical advances. *Advances in Experimental Social Psychology, 34*, 251–330. doi:10.1016/S0065-2601(02)80007-7

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems, 2*(2), 1–25. doi:10.1145/1985347.1985353

Miller, A. P., & Hosanagar, K. (2019, November 8). How targeted ads and dynamic pricing can perpetuate bias. *Harvard Business Review*. Retrieved from https://hbr.org/2019/11/how-targeted-ads-and-dynamic-pricing-can-perpetuate-bias

Miller, J. (2020, September 18). Is an algorithm less racist than a loan officer? *The New York Times*. Retrieved from https://www.nytimes.com/2020/09/18/business/digital-mortgages.html

Moliner, C., Martínez-Tur, V., Peiró, J. M., Ramos, J., & Cropanzano, R. (2013). Perceived reciprocity and well-being at work in non-professional employees: Fairness or self-interest? *Stress and Health, 29*(1), 31–39. doi:10.1002/smi.2421

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of "datification." *Journal of Strategic Information Systems, 24*(1), 3–14. doi:10.1016/j.jsis.2015.02.001

Nissan, E. (2017). Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement. *AI & Society, 32*(3), 441–464. doi:10.1007/s00146-015-0596-5

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York: New York University Press.

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. doi:10.1126/science.aax2342

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Crown Publishing Group.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual Review of Sociology, 34*(1), 181–209. doi:10.1146/annurev.soc.33.040406.131740

Parra, C. M., Gupta, M., & Dennehy, D. (2021). Likelihood of questioning AI-based recommendations due to perceived racial/gender bias. *IEEE Transactions on Technology and Society, 3*(1), 41–45. doi:10.1109/TTS.2021.3120303

Phillips, L. T., & Jun, S. (2022). Why benefiting from discrimination is less recognized as discrimination. *Journal of Personality and Social Psychology, 122*(5), 825–852. doi:10.1037/pspi0000298

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review, 111*(3), 781–799. doi:10.1037/0033-295X.111.3.781

Revilla, M., & Ochoa, C. (2017). Ideal and maximum length for a web survey. *International Journal of Market Research, 59*(5), 557–565. doi:10.2501/ijmr-2017-039

Rosen, E., Garboden, P. M. E., & Cossyleon, J. E. (2021). Racial discrimination in housing: How landlords use algorithms and home visits to screen tenants. *American Sociological Review, 86*(5), 787–822. doi:10.1177/00031224211029618

Schmitt, M. T., Branscombe, N. R., & Postmes, T. (2003). Women's emotional responses to the pervasiveness of gender discrimination. *European Journal of Social Psychology, 33*(3), 297–312. doi:10.1002/ejsp.147

Sellers, R. M., & Shelton, J. N. (2003). The role of racial identity in perceived racial discrimination. *Journal of Personality and Social Psychology, 84*(5), 1079–1092. doi:10.1037/0022-3514.84.5.1079

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York, NY: Springer-Verlag.

Stevens, C., Liu, C. H., & Chen, J. A. (2018). Racial/ethnic disparities in US college students' experience: Discrimination as an impediment to academic performance. *Journal of American College Health, 66*(7), 665–673. doi:10.1080/07448481.2018.1452745

Stouten, J., De Cremer, D., & van Dijk, E. (2006). Violating equality in social dilemmas: Emotional and retributive reactions as a function of trust, attribution, and honesty. *Personality and Social Psychology Bulletin, 32*(7), 894–906. doi:10.1177/0146167206287538

Stroebe, K., Dovidio, J. F., Barreto, M., Ellemers, N., & John, M.-S. (2011). Is the world a just place? Countering the negative consequences of pervasive discrimination by affirming the world as just: Negative consequences of discrimination. *British Journal of Social Psychology, 50*(3), 484–500. doi:10.1348/014466610x523057

Sue, D. W. (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. Hoboken, NJ: John Wiley & Sons.

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism, 7*(4), 447–469. doi:10.1080/21670811.2018.1493936

Vincent, G. M., & Viljoen, J. L. (2020). Racist algorithms or systemic problems? Risk assessments and racial disparities. *Criminal Justice and Behavior, 47*(12), 1576–1584. doi:10.1177/0093854820954501

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. doi:10.1145/3313831.3376813

Williams, B., Brooks, C., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy, 8*, 78–115. doi:10.5325/jinfopoli.8.2018.0078

Wong, C. A., Eccles, J. S., & Sameroff, A. (2003). The influence of ethnic discrimination and ethnic identification on African American adolescents' school and socioemotional adjustment. *Journal of Personality, 71*(6), 1197–1232. doi:10.1111/1467-6494.7106012

Yu, K.-H., & Kohane, I. S. (2019). Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety, 28*(3), 238–241. doi:10.1136/bmjqs-2018-008551