Understanding Reactions to Human versus Algorithmic Bias

Lily Simon [redacted]

Department of Psychology, McGill University

PSYC 396: Undergraduate Research Project

Dr. Jordan Axt

December 10, 2021

Abstract

Algorithms have been advanced as a method to overcome the often biased, error-prone nature of human decision-making. However, the growing acceptance of algorithms has been met with resistance, out of fear that algorithms may perpetuate social inequalities. Recent findings support these concerns, as cases of algorithmic discrimination are unfolding in various contexts, including in healthcare and employment. In light of these discussions, it is important to understand how individuals may uniquely perceive biased algorithmic judgments as compared with biased human judgments. The current study (*N* = 364) explores this question within a hiring context, by investigating how people react to an algorithm versus a team of employees making decisions that result in discrimination against women or men. The findings suggest that people have stronger negative reactions and less trust in the hiring method when women were discriminated against; however, the nature of the evaluator did not have a significant effect on negative reactions or trust in the hiring method. The findings further indicate that people perceive significantly greater company responsibility when an algorithm is the cause of discrimination against men. Lastly, individuals were more likely to believe that the company would switch their hiring method when an algorithm was the evaluator. This research provides insight into how people perceive algorithmic versus human bias differently, while indicating that people react strongly to discrimination regardless of the decision agent. To avoid the potential of algorithms sustaining social disparities, it is important that governmental structures work to support the responsible use of algorithms.

Understanding Reactions to Human versus Algorithmic Bias

Biases pervade human cognition, influencing the judgments and decisions people make (Tversky & Kahneman, 1974). These judgment biases can have far-reaching consequences, including contributing to and sustaining intergroup discrimination (Bertrand & Mullainathan, 2004). To overcome human biases, many people have advocated for greater reliance on computer algorithms to aid or even fully complete judgments that used to be left entirely up to humans. For instance, research has advanced the use of responsible algorithms in organizational decision-making to attenuate the problems caused by biased, noisy human decision-making, since algorithms have the potential for making equitable decisions in hiring and promotion (Houser, 2019). Although algorithms, like humans, are susceptible to reproducing social category disparities, it may be more straightforward to fix biased machines than biased people (Mullainathan, 2019).

Despite their potential, algorithms may only reduce discrimination with proper regulation, which has yet to be established (Mullainathan, 2019). Indeed, regulation over the use of algorithms for determining important outcomes, such as those related to housing or admissions, may be necessary in order for such algorithms to avoid doing harm (Kantayya, 2020). These concerns emerge from the outlook that, if algorithms are simply used to reproduce the world as it is today, social progress will stagnate.

Empirical work has supported concerns over how bias can manifest in algorithms. Algorithms may reproduce racial and gender disparities, either through the biases of the people constructing them or in data used to train them (Manyika et al., 2019; Barocas & Selbst, 2016). For example, one recent study (Obermeyer et al., 2019) analyzed a widely used commercial prediction algorithm for allocating healthcare to patients with complex medical needs. The findings indicated that, even when Black and White patients were given the same risk score by the algorithm, Black patients had a worse objective health measure (Obermeyer at al., 2019). That is, although race was not a factor used by the algorithm for determining who is eligible for the program, racial disparities in access to healthcare persisted.

To explain how this effect could emerge, Obermeyer and colleagues (2019) discovered that the algorithm was using health costs to predict eligibility for the program despite the disparity between health costs– the amount a patient spends to maintain health– and objective health– a patient's need for health care. By treating health costs as a proxy for health needs, the algorithm introduced racial bias, as Black patients generate less healthcare costs than White patients, a difference that may be attributed to factors like distrust in the healthcare system or racial discrimination by healthcare providers (Obermeyer et al., 2019).

Algorithmic discrimination extends beyond healthcare into other domains, such as policing and employment. For instance, Amazon recently discarded a hiring algorithm designed to screen resumes for a software developer role because it was systematically biased against women (Dastin, 2018). The algorithm was built on resumes the company amassed over a decade, which were primarily from male applicants (Heilweil, 2020). As a result, the resume-screening tool factored in proxies for gender, including graduation from a women's college or terms such as "women's chess club captain" on resumes (Dastin, 2018). In other words, by using existing data to create an algorithmic screening tool, the algorithm produced gender biases already present in the company. These cases of algorithmic discrimination present a key issue identified in research; even systems designed to ignore social information like race or gender may still perpetuate such biases through the reliance on proxy variables.

As of yet, there is no clear path for addressing algorithmic bias and discrimination. This is in part because algorithmic systems operate in a "black box"; there's often a lack of visibility in terms of how an algorithm was created, the data used in building it, and how it functions (Heilweil, 2020). Nevertheless, algorithmic decision-making is on the rise. This is of particular concern in hiring, as organizations are increasingly relying on algorithms for identifying qualified candidates. This shift is in part due to efficient processing times and lower cost of algorithmic screening systems as compared with traditional methods (Tippins et al., 2021).

Considering the growing extent to which algorithms are determining consequential life outcomes, it is important to understand how individuals perceive algorithmic decision-making and algorithmic discrimination more specifically. This study seeks to answer the following questions: (1) How do people react to algorithms as compared with humans making biased decisions that result in discrimination? (2) How do people react to a historically disadvantaged group (women) as compared with a historically advantaged group (men) being discriminated against, and does this reaction change when the discrimination is due to human versus algorithmic bias?

Past Research on Perceptions of Algorithmic versus Human Decision-Making

Algorithm Aversion

Individuals may be resistant to using algorithms to begin with, even when discrimination is not the end result. This is known in research as algorithm aversion, which describes how people quickly lose confidence in algorithms after seeing them err (Dietvorst et al., 2015). Evidence for this phenomenon was found through a series of studies in which participants either observed an algorithm make forecasts, a human make forecasts, both, or neither (Dietvorst et al., 2015). Their findings indicate that people are especially averse to an algorithm seen to perform imperfectly, even when its performance surpasses that of a human (Dietvorst et al., 2015). Resistance to algorithms may then partly stem from a greater intolerance for error produced by algorithms as compared with humans; people are more likely to discard an algorithm than a human judge for making the same mistake.

Prior research has shown that algorithm aversion can be reduced when the algorithm is observed as being able to adapt or when individuals can modify an imperfect algorithm’s predictions (Dietvorst et al., 2018; Berger et al., 2020). For example, a recent study (Berger et al., 2020) investigated the value of showing an algorithm’s potential to learn as a way to counteract algorithm aversion. Through an online experiment, participants were asked to solve a business forecasting task and decide to what extent they would rely on an erring advisor to boost their chance of gaining a financial bonus. The experimental conditions differed in the type of the advisor (human versus algorithmic), its familiarity to participants (unfamiliar versus familiar), and its potential to learn (non-learning versus learning) to see how these aspects of the advisor would affect the participants’ reliance on the advice. The study findings reveal that familiarity with an erring algorithm reduces participants’ subsequent confidence in its advice; however, the demonstration of an algorithm’s potential to learn neutralizes this effect (Berger et al., 2020). When an algorithmic advisor does not meet one’s expectations, their reliance on it diminishes; thus, individuals may quickly overlook the value of information generated by algorithms (Berger et al., 2020). People not only appear to hold higher and more unrealistic expectations of algorithms than humans, but also perceive algorithms as lacking the ability to improve. Consequently, it would be expected that an imperfect algorithm (e.g., producing biased outcomes) may lead individuals to harbor doubt in the algorithm’s abilities and swiftly replace it.

Justice Perceptions of Algorithmic versus Human Decision-Making

Studies on algorithm aversion have focused on how people react to algorithmic versus human decision-making on nonhuman tasks (e.g., forecasting). To understand how the humanness of a task may change the way in which people react to algorithmic versus human decision-making, a study was conducted where participants were exposed to decisions made by algorithms or humans on a series of tasks perceived as requiring human skills (i.e., subjective judgment and emotional capability) or mechanical skills (i.e., ability to process quantitative data for objective measures; Lee, 2018). With tasks requiring mechanical skills (i.e., work scheduling and work assignment) algorithmic and human-made decisions did not differ in terms of participants' perceptions of fairness, trustworthiness, and elicited emotions; however, for tasks requiring human skills (i.e., hiring and work evaluation), decisions made by algorithms were perceived as less just and reliable and induced greater overall negative emotion than decisions made by humans. People's perceptions of algorithmic versus human decision-making appear to be dependent on the task characteristics (i.e., requiring human or mechanical skills). For algorithmic decision-making on a task requiring human skills, like hiring, it is clear that individuals may foster more negative sentiments about the algorithm's judgment.

Recent research has investigated justice perceptions of hiring in greater depth than explored in the study by Lee (2018; Noble et al., 2021). For instance, the work by Lee (2018) only included one item to measure overall decision fairness. In comparison, research by Noble and colleagues (2021) aimed to provide a multi-dimensional understanding of how organizational justice perceptions are affected by algorithmic screening methods. Participants were asked to read a vignette describing a scenario where they were applying for a job. The scenarios differed by condition such that participants were either told a "hiring representative will look through" or an "artificial intelligence (AI) technology" will scan their resume and that they either "meet the organization's qualifications" or "do not meet the organization's qualifications". The results indicate that participants rated algorithmic screening as less fair than traditional screening in all respects except for consistency. These perceived differences in fairness may partly come from a lack of confidence in the algorithm's ability to recognize occupational qualifications (i.e., job related skills, qualities, experiences; Noble et al., 2021). Their findings indicate that the shift from traditional to algorithmic decision-making in hiring has mostly negative effects on justice perceptions.

Perceptions of Biased Algorithms versus Humans

In the studies by Lee (2018) and Noble and colleagues (2021), their studies operated on notion that algorithms and humans are performing without mistakes. However, how would individuals perceive algorithms as compared with humans when they make an "unjust" mistake (e.g., discrimination)? This question has been explored to some extent by prior research. Through a series of studies, people's moral outrage about discrimination in hiring and lending decisions caused by algorithms versus humans was examined (Bigman et al., 2020). Findings showed an "algorithmic outrage asymmetry", in which people are less morally outraged by algorithmic discrimination than by human discrimination for various forms of inequality (e.g., gender inequality, racial inequality). The studies further suggest that algorithmic outrage asymmetry occurs because people are less likely to attribute prejudice to an algorithm; thus, they would be less morally outraged when algorithms discriminate (Bigman et al., 2020).

Study Purpose

Prior literature indicates that people have more justice concerns when algorithms make decisions than when humans make decisions (Noble et al., 2021; Lee, 2018); however, people simultaneously perceive algorithms as having less prejudiced motivation than humans (Bigman et al., 2020). Less is known about how people perceive bias coming from algorithms versus humans when the target of discrimination changes (i.e., a historically advantaged group versus historically disadvantaged group is affected). Our research looks at reactions to discrimination perpetrated by algorithms versus humans and how these reactions may change as a result of the target of discrimination (i.e., women or men).

This study will contribute to the literature by synthesizing research on perceptions of algorithmic decision-making and algorithm aversion to understand how individuals perceive biased errors caused by algorithms that affect different social groups. Differences in how people react to algorithmic bias as compared with human bias may inform the process by which people adopt algorithms to replace human decision-making.

Summary of Procedure

For this study, participants are given a questionnaire with a scenario describing an instance of discrimination in hiring. We decided to use a hiring scenario, since it is seen as a task requiring human skills (Lee, 2018). The study will follow a 2 X 2 design, in which the evaluator (algorithm or team of employees) and target of discrimination (men or women) are manipulated in the scenario given to participants. Participant reactions were captured through a self-report questionnaire. Furthermore, we assessed a range of reactions beyond moral outrage (i.e., assessing whether people perceive algorithmic decision-making as unjust, wrong, or immoral), including trust in the hiring method, perceived company responsibility and global negative reaction.

Hypotheses

It is expected that people will report a greater negative reaction when the algorithm is the decision-maker because 1) algorithms are expected to perform perfectly and 2) the hiring task is considered to involve human skills (Lee, 2018). Support for this prediction comes from Lee (2018), which found that, on tasks that involve human skills, participants reacted more negatively to the decisions made by algorithms as compared with humans. In addition, we anticipate that participants will react more strongly to women being discriminated against as compared with men. It is expected that this result will emerge because events where women face discrimination in real-life may be more salient to participants than events where men face discrimination. As the scenario involves a technology-driven company and hiring for a highly gendered occupation (i.e., software engineering), the discrimination women face on the basis of gender will become especially pronounced. Since the algorithm is expected to perform without error, we would further expect participants to react more negatively when it reproduces real-world social category disparities (i.e., creating an interaction between evaluator type and form of discrimination). Seeing the algorithm err in a way that is perceived to be especially harmful in today's world– discriminating against women– may be even more upsetting.

Finally, we would expect that people may perceive the company as less responsible for algorithmic decisions, as perhaps algorithms are viewed as less subject to oversight. Greater trust in the hiring method is expected to result when evaluation is completed by the team of employees. We would anticipate this to happen because people quickly lose trust in algorithms when they see them make a mistake (Dietvorst et al., 2015). We would also expect that participants will perceive a greater likelihood of the company changing their hiring practices when the algorithm is the decision-maker because participants may not be aware of the potential for algorithmic improvement (Berger et al., 2020). Since people quickly lose confidence in algorithms when they err (Dietvorst et al., 2015), it is possible that participants will be more willing to scrap the algorithm hiring method.

Method

Participants

We recruited 364 participants (117 men, 239 women, 8 other/multiple identities; age: *M* = 32.8, *SD* = 11.62) from Prolific, an online tool for recruiting participants for research. This provided 80% power for detecting a medium effect size of *f* = 0.15. Participants were paid $1.75 US dollars for their time. For those who reported one race, participants were Black (3.57%), East Asian (1.37%), Southeast Asian (0.55%), Indigenous (1.37%), Latino (0.27%), South Asian (1.10%), and White (87.36%). Some participants (2.20%) chose multiple races.

Procedure

The study procedure began with participants providing their informed consent electronically (see Appendix A). After doing so, they were randomly assigned to one of four conditions: (1) algorithmic bias against women, (2) algorithmic bias against men, (3) human bias against women, and (4) human bias against men. Participants read the following scenario (note: their assigned condition corresponds to which of the four versions of the scenario they saw):

> Imagine a technology company has an urgent need to find skilled software engineers. To help with the search, the company's leadership (designed a computer algorithm/put together a team of employees) to review the uploaded resumes from applicants who applied to the openings. The (algorithm/team of employees) then selected who (it/they) believed to be the top applicants. However, after a review of the recruitment process, the company found that their (algorithm/team) was discriminating against (women/men) in hiring. That is, the (algorithm's/team's) decisions were such that when given an equally qualified male and female applicant, the (algorithm/decision-makers) consistently

favored the (male/female) applicant. As a result, the (algorithm/team) may have excluded qualified (female/male) candidates for the role.

After reading the scenario, participants recorded their reactions to the scenario through a 6-item questionnaire. The questionnaire assessed their (1) emotional distress, (2) perceived harm to excluded applicants, (3) perceived company responsibility for the outcome, (4) perceived emotional distress of excluded applicants, (5) perceived flexibility of the company changing their practices, and (6) their trust in the hiring method (see Appendix B). In a Likert-scale format, question responses range from 1 to 5. Consistent with our pre-registration, the items on emotional distress, perceived harm to excluded applicants, and perceived emotional distress of excluded applicants were combined for overall negative reaction since the scores correlated with each other (all above $r = 0.50$). The items on perceived company responsibility for the outcome, perceived flexibility in the company changing their hiring practices, and trust in the hiring method were analyzed separately, as they were believed to capture different aspects of participant reaction. Following this questionnaire, participants were asked to provide their demographic information (i.e., gender, race, age, familiarity with algorithms, and political orientation; see Appendix C). Finally, all participants were debriefed and given more details about the study (see Appendix D).

Analysis

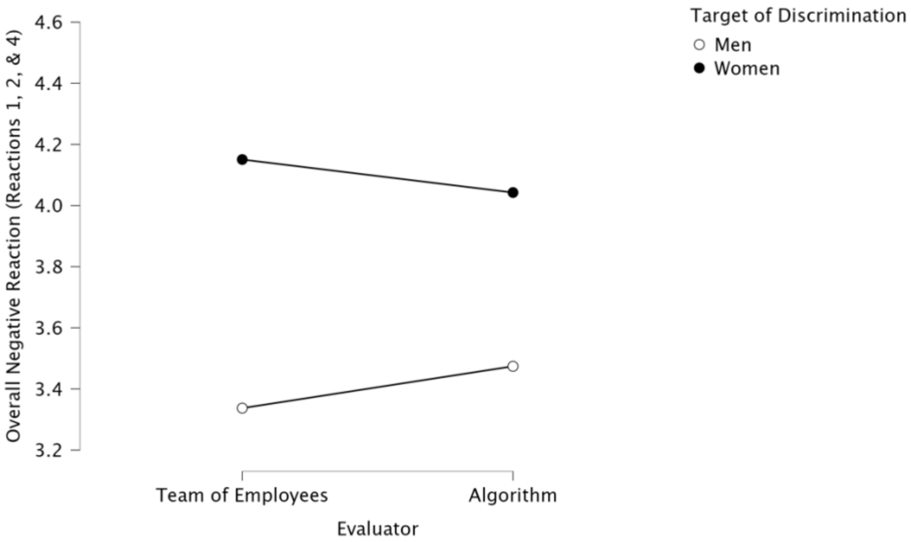
We conducted a series of between-subjects ANOVAs to assess the effects of the evaluator (algorithm versus team of employees) and target of discrimination (women versus men) on the dependent variables. As mentioned previously, emotional distress, perceived harm to excluded applicants, and perceived emotional distress of excluded applicants were aggregated into a sum score of negative reaction. ANOVA tests were run separately for the items concerning perceived company responsibility for the outcome, perceived flexibility of the company changing their hiring practices and trust in the hiring method.

Results

First, a two-way ANOVA was performed to analyze the effect of the evaluator and target of discrimination on overall emotional reaction. Simple main effects analysis showed that the evaluator did not have a statistically significant effect on overall negative reaction ($F(1, 360) = 0.027, p = 0.868$). A large main effect of target of discrimination on overall negative reaction was found, $F(1, 360) = 62.71, p < 0.001, \eta^2 = 0.15$. When the target of discrimination was women, there were, on average, greater overall negative emotional reactions (Team Evaluator Condition: $M = 4.15$, $SD = 0.79$; Algorithm Evaluator Condition: $M = 4.04$, $SD = 0.83$) than when men were the target of discrimination (Team Evaluator Condition: $M = 3.34$, $SD = 0.84$; Algorithm Evaluator Condition: $M = 3.47$, $SD = 0.87$; Figure 1). Our results indicate that there was not a statistically significant interaction between the effects of evaluator and target of discrimination ($F(1, 360) = 1.97, p = 0.16$).

Figure 1

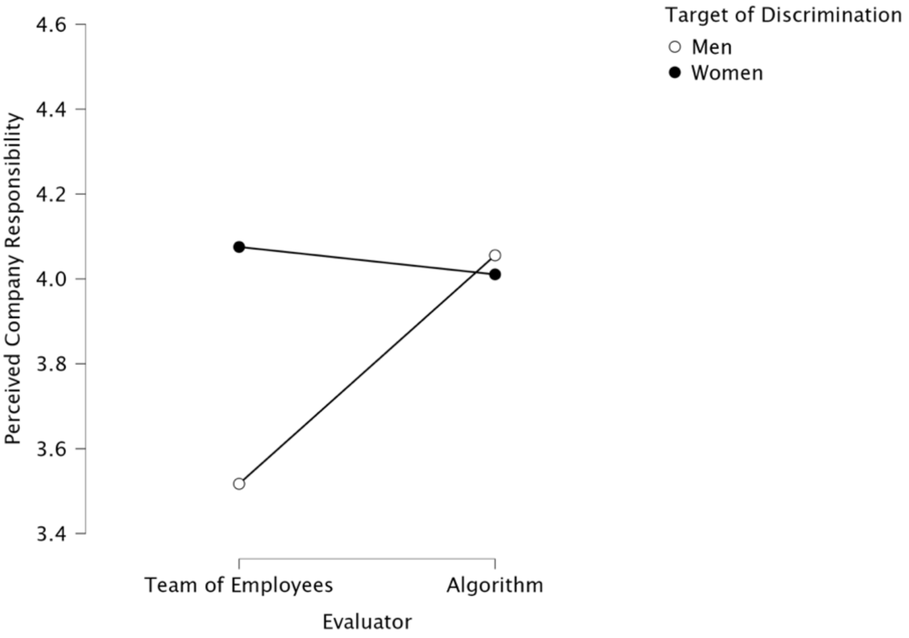
Plot for ANOVA on Overall Negative Emotional Reaction



In addition to overall emotional reaction, we ran a series of two-way ANOVAs to assess the effects of the evaluator and target of discrimination on 1) perceived company responsibility, 2) trust in the hiring method, and 3) perceived likelihood of the company changing their hiring method. A small main effect of evaluator on perceived company responsibility was found, *F*(1, 360) = 4.15, *p* = 0.04, $\eta^2 = 0.011$. The target of discrimination had a small main effect on perceived company responsibility as well, *F*(1, 360) = 5.98, *p* = 0.028, $\eta^2 = 0.013$. A small interaction between the effects of evaluator and target of discrimination was found for perceived company responsibility, *F*(1, 360) = 8.26, *p* = 0.010, $\eta^2 = 0.018$. When men were the target of discrimination, there was a significant increase in perceived company responsibility in the algorithm condition as compared with the team of employees condition (Figure 2). When women were the target of discrimination, there was no statistically significant difference between the evaluator conditions.

Figure 2

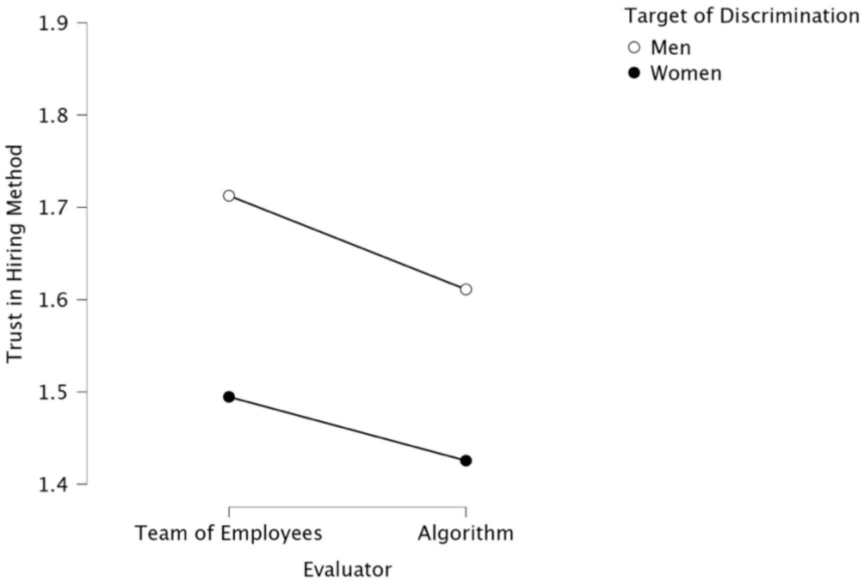
Plot for ANOVA on Perceived Company Responsibility



For trust in the hiring method, the simple main effect analysis showed that the evaluator did not have a statistically significant effect on trust in the hiring method (*F*(1, 360) = 1.04, *p* = 0.309). A small main effect of the target of discrimination on trust in the hiring method was found, *F*(1, 360) = 5.81, *p* = 0.016, $\eta^2 = 0.016$. When women were the target of discrimination, participants on average had less trust in the hiring method (Team Evaluator Condition: *M* = 1.49*,* *SD* = 0.67; Algorithm Evaluator Condition: *M* = 1.43, *SD* = 0.74) than when men were the target of discrimination (Team Evaluator Condition: *M* = 1.71, *SD* = 0.87; Algorithm Evaluator Condition: *M* = 1.61, *SD* = 0.90), regardless of whether the evaluator was an algorithm or team of employees (Figure 3). There was not a statistically significant interaction between the effects of the evaluator and target of discrimination (*F*(1, 360) = 0.038, *p* = 0.85).

Figure 3

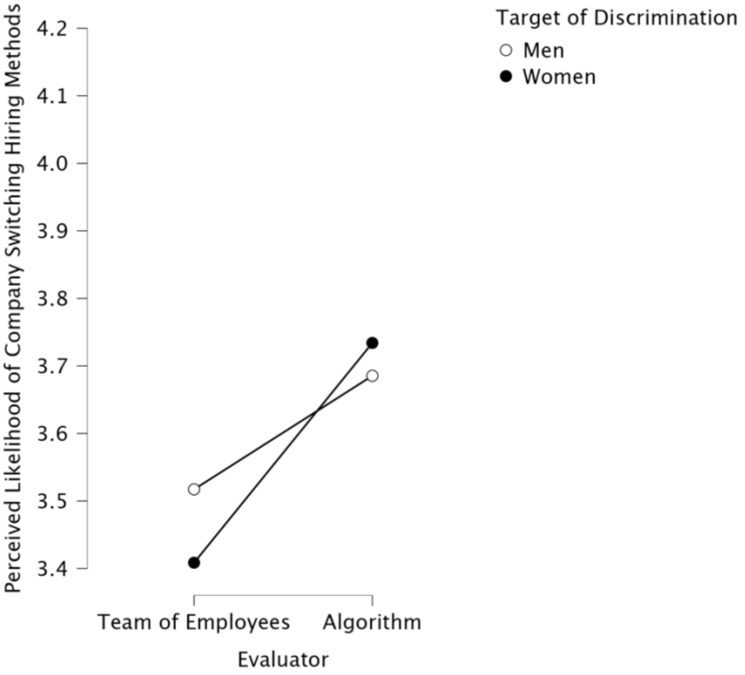
Plot for ANOVA on Trust in Hiring Method



Finally, for perceived likelihood of the company changing their hiring method, the simple main effect analysis showed a small main effect of the evaluator on the perceived likelihood of the company switching hiring methods, *F*(1, 359) = 4.86, *p* = 0.028, $\eta^2 = 0.013$. Participants were, on average, more likely to see the company changing their hiring method when the algorithm (Women Condition: *M* = 3.73, *SD* = 1.04; Men Condition: *M* = 3.68, *SD* = 1.05) was the evaluator than when the team of employees was the evaluator (Women Condition: *M* = 3.41, *SD* = 1.08; Men Condition: *M* = 3.52, *SD* = 1.10; Figure 4). Simple main effect analysis showed that the target of discrimination did not have a statistically significant effect on perceived likelihood of the company switching hiring methods (*F*(1, 359) = 0.072, *p* = 0.79). There was not a statistically significant interaction between the effects of evaluator and target of discrimination (*F*(1, 359) = 0.494, *p* = 0.48).

Figure 4

Plot for ANOVA on Perceived Likelihood of Company Switching Hiring Methods



Discussion

This study investigated people's reactions to discrimination caused by algorithms as compared with humans, and how reactions may change depending on the social group affected (i.e., women versus men). We hypothesized that algorithmic decision-making would result in stronger and more negative emotional reactions. The results do not support this hypothesis, as there was no significant difference between the evaluator conditions (algorithm versus team of employees) in emotional reaction. This suggests that people will react negatively to discrimination regardless of who– or what– is causing discrimination. This finding is inconsistent with previous research suggesting that people respond more negatively to algorithms as compared with humans completing human tasks (Lee, 2018). One potential way of reconciling these different results is through considering error. Whereas the study by Lee (2018) did not involve an erring evaluator, the current study investigates a more consequential mistake in discrimination. The finding suggests that judgment errors resulting in discrimination diminishes any perceived differences in emotional reaction between human and algorithmic evaluators.

We further anticipated that participants would report greater emotional reactions when women are the target of discrimination (compared to men). Indeed, our results indicate that participants were significantly more upset when women were discriminated against as compared with men. This may be a result of how women are the target of discrimination in the real-world where they remain vastly underrepresented in the tech industry (Kennedy et al., 2021), and so participants were less bothered by an instance of "reverse discrimination" that did not reinforce existing disparities.

In addition to emotional reaction, we investigated people's perceptions of company responsibility for the outcome, their trust in the hiring method, and perceptions of the company's likelihood in changing the hiring method. We explored whether people would see the company as less accountable for the outcome when the algorithm is the evaluator because it may be more difficult for individuals to identify who is culpable in this circumstance (e.g., the company, creators of the algorithm). The results indicate that there was a significant difference between evaluator conditions when men were the target of discrimination, such that people perceived greater company responsibility when algorithms were biased against men. This finding is inconsistent with our hypothesis. This result may be attributed to the wording of the scenario, which makes it clear that the company designed the algorithm; thus, there is little to no ambiguity concerning who is responsible for the discriminatory outcome in both evaluator conditions. However, this does not specifically account for why, when the algorithm discriminated against men, greater perceived company responsibility resulted.

A possible interpretation of this finding may be that, when women are the target of discrimination, individuals are insensitive to the nature of the evaluator. Individuals may see that the company needs to be held accountable for the perpetuation of existing gender-based disparities, regardless of the cause. However, when men are the target of discrimination, individuals may believe that the human decision-makers are proactively trying to hire more women to ensure that existing biases against women do not creep into the hiring process. The company may then be seen as less accountable for the discrimination against men if they are attempting to override any potential biases against female applicants. Conversely, algorithms may be seen as incapable of having these intentions, and so the outcome may be seen as an error rather than a company choice. Further research is necessary to understand why this difference in evaluator conditions emerged when men are the target of discrimination.

In perceptions of trust, we anticipated that participants would report less trust in the algorithm hiring method since previous research has indicated that people rapidly lose confidence in algorithms after seeing them err (Dietvorst et al., 2015). The results were not consistent with this hypothesis, as there was no statistically significant effect of the evaluator on trust in the hiring method. This may be attributed to the error being discrimination in this study, rather than incorrect predictions in forecasting seen in previous studies (Dietvorst et al., 2015;

Berger et al., 2020). Whereas forecasting mistakes do not necessarily cause harm, discrimination is harmful to those affected. Perhaps trust in the hiring method was low across evaluator conditions because individuals were focused on the outcome (i.e., discrimination), rather than the nature of the evaluator. It is expected and more understandable that humans make mistakes in forecasting decisions, whereas algorithms are held to much higher standards of performance (Dietvorst et al., 2015). A mistake resulting in discrimination may be seen as intolerable, regardless of the responsible agent; thus, participants may be more outcome-focused rather than process-focused when determining how much they trust the hiring method. Indeed, the target of discrimination had an impact on people's trust in the hiring method, as participants expressed less trust in the hiring method when women were biased against. This finding may be explained by the notion that the hiring method is reinforcing real-world disparities.

We further expected that participants would perceive a greater likelihood of the company changing their hiring practices when the algorithm is making decisions due to research indicating that individuals do not see the potential for algorithmic improvement unless shown (Berger et al., 2020). After seeing the algorithm err by generating bias, participants may be more willing to replace this hiring method (Dietvorst et al., 2015; Berger et al., 2020). The results are consistent with this hypothesis; when the algorithm was the evaluator, participants reported a greater likelihood of the company changing their hiring practices.

Implications

This research can serve as an initial step towards understanding how people may perceive algorithmic versus human bias, thereby contributing to an emerging area of research on perceptions of flawed algorithmic decision-making. The findings suggest that people react adversely to discrimination, regardless of the responsible agent. Consequently, companies that adopt algorithms for hiring decisions will not face greater backlash in the event that discrimination is detected in their hiring process. When further considering the efficiency of algorithmic hiring systems (Tippins et al., 2021) and the relative ease at which bias may be fixed (Mullainathan, 2019), algorithms may indeed be better– or at least no worse – than humans for hiring judgments. However, it is critical that proper regulation and practices are implemented to ensure the responsible use of algorithms. For instance, the presence of bias in algorithmic hiring systems should be routinely monitored for and eliminated (Turner Lee, 2019). Once proper regulation and practices are established, companies can harness algorithmic decision-making to progress closer towards equity in hiring.

Limitations

Several limitations should be considered regarding our study sample and design. The sample recruited from Prolific is not representative of any identifiable population. Instead, the sample overrepresents young, White, and female participants. As a result, it is possible that the results may lack generalizability, especially to racial/ethnic minority groups since they are more likely to face the ramifications of algorithmic decision-making (Kantayya, 2020). A more representative sample may have led to stronger negative emotional reactions to algorithmic discrimination than what was found in the present study.

In addition to the sampling issues, the study was not immersive and therefore lacked psychological realism. Since participants were only asked to read a brief, general scenario describing discrimination in hiring, this may not have fully captured how algorithmic and human discrimination is uniquely experienced and perceived in everyday life. That is, there were no real stakes for study participants. As a result, one way to increase the realism of the study would be to make the target of discrimination the participant or someone they know (e.g., friend, family member). Another potential change to make the study more immersive would be to describe real instances of algorithmic and human discrimination, such as the case of bias in Amazon's hiring algorithm (Dastin, 2018). By making the study more real and immersive for participants, this may have led to different findings. More specifically, a more immersive study may have perhaps resulted in reports of stronger negative emotional reactions, regardless of the nature of the evaluator (i.e., algorithm versus human).

Future Directions

An important direction for future research is to explore whether these findings hold when the evaluator is shown to improve. Previous research has shown that people see algorithms as lacking the ability to improve (Berger et al., 2020); thus, mistakes and inaccuracies in algorithmic decision-making are weighed more severely as compared with human decision-making. When it comes to a consequential mistake (e.g., discrimination), it is possible that people see algorithms as incapable of change as well. An extension of this study can explore how, when bias is shown to be routinely extracted from algorithms, people may react to discrimination perpetrated by algorithms in terms of their trust in the hiring method and perceived likelihood of the company changing their hiring practices. This would be a valuable extension of the current research by understanding the circumstances under which people feel more comfortable with the use of algorithms in hiring.

Another worthwhile direction of future research would be to explore how people perceive hybrid hiring processes in which both algorithms and humans have a role in decision-making. This better reflects real-world hiring practices, which are often made up of both algorithms and humans (Harris, 2018). While algorithms often play a role in the initial screening of job candidates, hiring experts often, if not always, make the final hiring decision. In the case of discrimination within this hiring process, it would be interesting to see where individuals attribute discrimination in such cases.

A third direction of future research would be to look at how this research may extend to other forms of discrimination. It would be especially important to look at racial discrimination since racial/ethnic minorities often face the negative implications of algorithmic decision-making (Kantayya, 2020; Obermeyer, 2019). For instance, predictive policing tools have been shown to preserve systemic racism, as Black people are disproportionately forecasted to have a high probability of engaging in criminal activity (Richardson et al., 2019). Due to rising concerns over the perpetuation of racial discrimination through algorithmic decision-making, individuals may react more negatively to algorithmic discrimination on the basis of race than gender. This may be especially true if the context is one in which racial discrimination is particularly salient (e.g., policing, healthcare).

A final direction of future research would be to increase the personal relevance of the discrimination described in the scenario. As previously mentioned, this would increase the immersiveness of the study. Greater relevance of materials may also allow for a better understanding of how individuals react in the face of discrimination, as reactions found in the current study could be even further amplified. A potential way to investigate this idea would be to create an immersive scenario where participants apply for a job using their own resume, which could be evaluated by either an algorithm or human. They could then be informed that their application was rejected. For some participants, it could be revealed that this rejection was on the basis of discrimination. This extension of the current study may allow for a better understanding of how individuals perceive discrimination when they are on the receiving end.

Conclusion

This research indicates that people react more negatively to discrimination against women than men, regardless of the decision agent (i.e., team of employees or algorithm). The results further show that people perceive greater company responsibility when an algorithm, as compared with the team of employees, was discriminating against men. People had less trust in the hiring method when women were the target of discrimination irrespective of the nature of the evaluator. Finally, people were more likely to believe that the company would discard the biased algorithm than the team of employees. The findings from the current study can inform future research on how people perceive flawed algorithmic performance as compared with flawed human performance in hiring. In light of the potential for algorithmic decision-making to perpetuate social biases, it is crucial that proper regulation and procedures are established to ensure the responsible use of algorithms by companies.

References

Barocas, S. & Selbst, A. D. (2016). Big data’s disparate impact. *California Law Review, 104*(3), 671-732. http://dx.doi.org/10.15779/Z38BG31

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2020). Watch me improve– Algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering, 63,* 55-68. https://doi.org/10.1007/s12599-020-00678-5

Bertrand, M. & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991-1013. https://doi.org/10.1257/000282804200256

Bigman, Y., Gray, K., Waytz, A., Arnestad, M., & Wilson, D. (2020). *Algorithmic discrimination causes less moral outrage than human discrimination.* PsyArXiv. https://doi.org/10.31234/osf.io/m3nrp

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters. https://www.reuters.com/article/us-amazon-comjobs-automation-insight/amazon-scraps-secret-airecruiting-tool-that-showed-bias-against-womenidUSKCN1MK08G

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

Dietvorst, B. J., Simmons, J. P., & Massey C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3),1155-1170. https://doi.org/10.1287/mnsc.2016.2643

Harris, C. G. (2018). Making better job hiring decisions using "human in the loop" techniques. *HumL@ISWC2018*. http://ceur-ws.org/Vol-2169/paper-03.pdf

Heilweil, R. (2020, February 18). *Why algorithms can be racist and sexist.* Vox. https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency

Houser, K. (2019). Can AI solve the diversity problem in the tech industry? Mitigating noise and bias in employment decision-making. *Stanford Technology Law Review.* https://ssrn.com/abstract=3344751

Kennedy, B., Fry, R., & Funk, C. (2021, April 14). *6 facts about America's STEM workforce and those training for it*. Pew Research Center. https://www.pewresearch.org/fact-tank/2021/04/14/6-facts-about-americas-stem-workforce-and-those-training-for-it/

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. https://doi.org/10.1177/2053951718756684

Manyika, J., Silberg, J., & Presten, B. (2019, October 25). *What do we do about the biases in AI?* Harvard Business Review. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Mullainathan, S. (2019, December 6). Biased algorithms are easier to fix than biased people. *The New York Times.* https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

Noble, S. M., Foster, L. L., & Craig, S. B. (2021). The procedural and interpersonal justice of automated application and resume screening. *International Journal of Selection and Assessment, 29*(2), 139-153. https://doi.org/10.1111/ijsa.12320

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

Kantayya, S. (Director). (2020). *Coded bias* [Documentary]. 7th Empire Media.

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review, 94,* 192-233. https://ssrn.com/abstract=3333423

Tippins, N., Oswald, F. L., & Morton McPhail, S. (2021). Scientific, legal, and ethical concerns about AI-based personnel selection tools: A call to action. *Personnel Assessments and Decisions, 7*(2), Article 1. https://doi.org/10.25035/pad.2021.02.001

Turner Lee, N., Resnick, P., & Barton, G. (2019). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms.* The Brookings Institution. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Appendix A – Consent Form

Title of Project*:* Understanding Reactions to Human versus Algorithmic Bias

Participation is voluntary. It is your decision to participate in this study. If you decide to participate, you may change your mind and leave the study at any point. Your data will not be collected in the case that you leave before the survey is complete. If you do submit your response, it cannot be withdrawn due to the anonymous nature of the survey. Refusal to participate or withdrawing your participation will involve no penalty or loss of benefits to which you are otherwise entitled in accordance with Prolific's user terms and conditions.

Introduction & Purpose: This study will examine how people react to various outcomes. Please note you must be 18 or older to participate in this study.

Study Procedures: During the study, you will be asked to read a scenario and then complete a short questionnaire about it. There are no known risks to participating in this study. The study will take up to 12 minutes to complete.

Are there any benefits from being in this research study? There are no foreseeable benefits for study participants.

Will I be compensated for participating in this research? You will be compensated 1.50 US dollars for participating in this study. Prolific will apply the Reward to your Prolific Account on the Website within seven working days of such approval. Payment to you will be made through Prolific's payment processing service provider PayPal, Inc. (to your PayPal account).

Confidentiality: Prolific's standard privacy policy applies to this study. Prolific has various technical and organizational security measures in place to protect your personal data and prevent the loss, misuse, or alteration of your personal data. Personal data will be stored on Prolific's secure servers. Data relating to your financial transactions that is sent from your web browser to

Prolific, or from Prolific to your web browser, will be protected using encryption technology. Your data will be stored indefinitely on Prolific's servers.

Contact: Please contact Jordan Axt via email at jordan.axt@mcgill.ca if you have any questions or desire further information with respect to this study.

If you have any ethical concerns or complaints about your participation in this study and want to speak with someone not on the research team, please contact the McGill Ethics Manager at 514-398-6831 or lynda.mcneil@mcgill.ca.

Consent: You have the right to choose to not answer some or any of the questions. By clicking the button below you are indicating that you have read the informed consent statements above and agree to participate. Please save or print a copy of this consent information for your records.

Appendix B – Questionnaire

1. How upset are you by this outcome?

    1) not upset at all
    2) upset
    3) somewhat upset
    4) very upset
    5) extremely upset

2. How harmful do you think this outcome is for the excluded applicants?

    1) not harmful at all
    2) slightly harmful
    3) moderately harmful
    4) very harmful
    5) extremely harmful

3. How responsible is the company for this outcome?

    1) not responsible at all
    2) slightly responsible,
    3) somewhat responsible
    4) mostly responsible
    5) fully responsible

4. How upset do you believe the excluded applicants would be if they found out about the bias in the selection process?

    1) not upset at all

2) slightly upset
3) somewhat upset
4) very upset
5) extremely upset

5. Considering what happened, how likely do you think it is that the company will switch to a different hiring method?
   1) extremely unlikely
   2) somewhat unlikely
   3) likely
   4) very likely
   5) extremely likely
6. How much do you trust this hiring method?
   1) I do not trust this hiring method at all
   2) I trust the hiring method a little bit
   3) I trust the hiring method a moderate amount
   4) I trust the hiring method a great deal
   5) I have extreme trust in the hiring method

Appendix C – Demographics

1. What is your current gender identity? (select all that apply)

(Male, Female, Trans male/ Trans man, Trans female/ Trans woman, Genderqueer/ Gender nonconforming, A different identity)

2. What is your age?

____

3. What is your race? (select all that apply)

Black (African, Afro-Caribbean, African Canadian descent)

East Asian (Chinese, Korean, Japanese, Taiwanese descent or Filipino)

Southeast Asian (Vietnamese, Cambodian, Thai, Indonesian, other Southeast Asian descent)

Indigenous (First Nations, Métis, Inuk/Inuit)

Latino (Latin American, Hispanic descent)

Middle Eastern (Arab, Persian, West Asian descent (e.g., Afghan, Egyptian, Iranian, Lebanese, Turkish, Kurdish)

South Asian (South Asian descent (e.g., East Indian, Pakistani, Bangladeshi, Sri Lankan, Indo-Caribbean)

White (European descent)

Another race category

Do not know

Prefer not to answer

4. What is your political orientation?

Strongly conservative, Moderately conservative, Slightly conservative, Neutral, Slightly liberal, Moderately liberal, Strongly liberal

5. How familiar are you with how algorithms work?

Not familiar at all, slightly familiar, moderately familiar, very familiar, extremely familiar

Appendix D – Debriefing Form

You have completed the study. Thank you for participating!

What was this study about?

We are investigating how individuals react to a computer algorithm versus a human making biased hiring decisions. We would also like to explore how people react to a historically disadvantaged group (women) versus a historically advantaged group (men) being discriminated against during the hiring process. Results can help researchers understand how people make sense of the growing use of algorithms for decision-making on tasks often thought of as requiring human skill and that may have a significant impact on their lives.

I still have questions about the study

If you have any questions or comments about the study, please email the lead investigator, Jordan Axt (jordan.axt@mcgill.ca).

If you have any ethical concerns or complaints about your participation in this study and want to speak with someone not on the research team, please contact the McGill Ethics Manager at 514-398-6831 or lynda.mcneil@mcgill.ca.