Lay Perceptions of Algorithmic Discrimination in the Context of Systemic Injustice

Gabriel Lima
Max Planck Institute for Security and Privacy
Bochum, Germany
gabriel.lima@mpi-sp.org

Nina Grgić-Hlača
Max Planck Institute for Software Systems
Saarbrücken, Germany
Max Planck Institute for Research on Collective Goods
Bonn, Germany
nghlaca@mpi-sws.org

Markus Langer
University of Freiburg
Freiburg im Breisgau, Germany
markus.langer@psychologie.uni-freiburg.de

Yixin Zou
Max Planck Institute for Security and Privacy
Bochum, Germany
yixin.zou@mpi-sp.org

Abstract

Algorithmic fairness research often disregards concerns related to systemic injustice. We study how contextualizing algorithms within systemic injustice impacts lay perceptions of algorithmic discrimination. Using the hiring domain as a case-study, we conduct a 2x3 between-participants experiment (N=716), studying how people's views of algorithmic fairness are influenced by information about (i) *systemic injustice* in historical hiring decisions and (ii) algorithms' propensity to *perpetuate biases learned from past human decisions*. We find that shedding light on systemic injustice has heterogeneous effects: participants from historically advantaged groups became more negative about discriminatory algorithms, while those from disadvantaged groups reported more positive attitudes. Explaining that algorithms learn from past human decisions had null effects on people's views, adding nuances to calls for improving public understanding of algorithms. Our findings reveal that contextualizing algorithms in systemic injustice can have unintended consequences and show how different ways of framing existing inequalities influence perceptions of injustice.

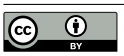
CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; • **Applied computing** → **Psychology**.

Keywords

Artificial Intelligence, Algorithms, Algorithmic Decision-Making, Injustice, Discrimination, Systemic Injustice

ACM Reference Format:
Gabriel Lima, Nina Grgić-Hlača, Markus Langer, and Yixin Zou. 2025. Lay Perceptions of Algorithmic Discrimination in the Context of Systemic Injustice. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/3706598.3713536



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713536

1 Introduction

The deployment of algorithms in high-risk decision-making scenarios has disproportionately impacted certain groups, such as gender and racial minorities [3, 33, 67]. Research on *algorithmic fairness* has sought to make these algorithms more computationally fair according to several statistical criteria [56, 63, 87, 152, 153]. However, many scholars argue that these computational approaches to algorithmic fairness—albeit well-intentioned—fail to achieve real justice because they overlook the underlying conditions that lead to the disproportionate outcomes they attempt to rectify [12, 45, 54, 69, 139, 156]. Instead, critics argue that one should conceptualize algorithmic discrimination through the lens of systemic injustice, proposing that algorithms should account for the unjust social structure (e.g., sexism and racism) that creates injustice [78, 102]. In other words, to develop algorithms that are truly fair and just, one must acknowledge that they are embedded in unjust social relations that cannot be solved solely by "fair" algorithms.

A similar gap regarding systemic injustice also exists in research capturing laypeople's perceptions of algorithmic fairness. Extensive literature has explored what individuals consider to be fair in the context of algorithmic decision-making [56, 57, 62, 65, 96, 133, 135, 136]; yet, these studies examine people's judgments of fairness without situating algorithms in systemic injustice. Given that this line of research aims to inform the development and regulation of algorithms [6, 91, 122], it is imperative that these studies also recognize how algorithms are socially situated to ensure that they do not end up reifying the same unjust conditions they attempt to rectify. This paper takes a first step towards exploring how contextualizing decision-making algorithms in systemic injustice may impact how laypeople perceive their harms.

We report a 2x3 between-participants vignette-based experiment (N=716), in which we manipulated whether participants are provided information about systemic injustice in hiring and examined whether doing so impacts how they perceive a racially biased hiring algorithm with respect to fairness and trust. Namely, we explored the following research question:

RQ1: How does shedding light on historical injustice in a particular domain impact how laypeople perceive algorithmic discrimination in the same domain?

Comprehending that algorithms can perpetuate systemic injustice relies on understanding that these systems learn from past human decisions and their biases. Nevertheless, prior research suggests that laypeople's understanding of algorithms is largely detached from reality [8, 25]. Unless individuals comprehend that decision-making algorithms learn from past human decisions, their perceptions of algorithmic fairness will not necessarily change, even if they acknowledge systemic injustice. Hence, our experiment also manipulated whether participants were told that algorithms (1) learn from past human decisions and (2) perpetuate their biases. Namely, we explored whether different ways of explaining how algorithms work moderate the effect of shedding light on injustice:

RQ2: How does explaining that algorithms learn from past human decisions—and thus perpetuate human biases—moderate the effect of contextualizing algorithms in historical injustice?

Individuals' perceptions of algorithmic fairness can also be tied to their own positionality regarding injustice. Research suggests that members of dominant groups (e.g., men and White individuals) employ defense mechanisms to maintain their advantages when exposed to information that portrays them as benefiting from an unjust state of affairs [17, 48, 82]. In contrast, there is also evidence that privileged individuals may show support towards attempts to dismantle systemic injustice when provided with relevant information [43, 82, 114]. Hence, we also studied whether participants' social position in relation to injustice impacts how they judge decision-making algorithms contextualized within unjust structures:

RQ3: How does one's positionality moderate the effect of contextualizing algorithms in historical injustice?

Across all experimental conditions, participants judged algorithmic hiring discrimination as largely unfair and reported that they would not trust the algorithm to make hiring decisions. Explaining that algorithms learn from past human biases and thus perpetuate these biases had little to no effect on participants' perceptions of algorithmic hiring discrimination. This null effect was independent of whether the algorithm was contextualized in systemic injustice.

Nevertheless, our results indicate that shedding light on systemic injustice has heterogeneous effects on participants' perceptions of algorithms depending on their racial identity. When provided information about systemic injustice, participants belonging to the group portrayed as advantaged within the study (i.e., White participants) reported *more negative* opinions regarding the algorithm. Surprisingly, participants belonging to racial minority groups exhibited the opposite trend, reporting higher trust and perceived fairness and becoming *more positive* about the algorithm, even though the algorithm was racially discriminatory against Black applicants.

An exploratory analysis of our experiment indicates that the extent to which participants believed in systemic injustice was strongly associated with their perceptions of the algorithm. Individuals with stronger beliefs in systemic injustice considered algorithmic discrimination more unfair and reported lower trust in the algorithm. Finally, we found that participants who judged the algorithm to be unfair and untrustworthy were more likely to indicate that the algorithm should be redesigned and even banned.

Our results reveal the difficulty of explaining how algorithms work to laypeople, particularly when contextualized in systemic injustice—a domain about which people might also have varying attitudes and understanding [83, 84]. We discuss our null effects of explaining how algorithms work on lay perceptions of algorithmic discrimination in the context of regulatory and design calls for improving lay understanding of algorithms for effective human oversight.

Our findings demonstrate how contextualizing algorithms in systemic injustice may have unintended consequences on people's judgments of algorithmic discrimination. Underscoring that humans have been historically biased may fuel the image of algorithms being more objective and less discriminatory, persuading members of historically disadvantaged groups towards algorithmic decision-making—even when clearly discriminatory against them. Our findings highlight how different ways of framing existing inequalities may impact individuals' perceptions of (in)justice—an important lesson for future efforts aiming to rectify injustices sustained by both algorithms and humans.

2 Background

2.1 Algorithmic Fairness and Its Critics

Algorithms are becoming increasingly prevalent in many high-stakes domains, assisting humans in making hiring decisions [33], granting bail [3], identifying fraud [67], allocating resources to refugees [142], among many other tasks. However, many algorithms have been found to be discriminatory, leading to disproportionate outcomes across groups. Studies have found, for instance, algorithms aimed at improving the provenance of child welfare services punishing poor households [42]; systems designed to detect fraud in social benefits disproportionately impacting ethnic minorities and foreigners [67]; and models deployed to assess health risk underdiagnosing conditions of Black patients [117].

To deal with these potential discriminatory outcomes, researchers have proposed computational methods to achieve various different notions of fairness in algorithmic outputs [56, 63, 87]. These approaches often rely on defining specific statistical criteria and proposing methods for training algorithms that satisfy them. For instance, algorithms may be trained to achieve equal error rates across different demographic groups [152] or assign benefits to different groups in equal proportions [153]. Research suggests that these definitions of fairness vary widely, several of them being computationally and socially incompatible with each other [26, 53, 80].

These computational approaches to algorithmic fairness have been criticized for overlooking the context within which algorithms are developed and deployed [53, 61, 149]. Critics argue that these systems fail to account for the conditions that lead to the disproportionate outcomes they surface, perpetuating and potentially exacerbating the social disparities computational approaches to fairness attempt to solve [78, 139, 155]. By disregarding how an unjust state of the world has come to exist, statistically fair algorithms end up providing misguided interventions to social problems and failing to identify who is accountable for righting specific injustices [44]. This computational approach to injustice can thus create a veil of objectivity and neutrality, by "letting the data speak for itself" [69], justifying and reifying an unjust reality through algorithmic interventions [139, 156].

For instance, although algorithms deployed in the criminal justice system may be fair with respect to some statistical criteria, they can still disproportionately impact defendants from historically disadvantaged groups. Because members of these groups are more likely to encounter the criminal justice system due to overpolicing [141], they are also more likely to be subjected to algorithmic decisions and therefore to their harms. Even if the algorithm's accuracy is the same for all racial groups, its harms will be disproportionately felt by those who have been historically harmed by compounding existing injustices [155].

Alternatively, scholars have built upon the humanities and social sciences and argued for conceptualizing algorithmic fairness as a systemic and structural problem rather than a computational one. For instance, Kasirzadeh [78] and Lin and Chen [102] have conceptualized algorithmic discrimination through the lens of structural injustice, which acknowledges how unjust social structures and power dynamics shape algorithmic outcomes. Green and Viljoen [54] have advocated for a shift from algorithmic formalism to algorithmic realism by centering contextual and political concerns. Mohamed et al. [111] and Mhlambi and Tiribelli [110] have proposed a decolonial approach to algorithmic fairness, focusing on how systems reify colonial-like relationships in ways that restrict the behavior of and pattern outcomes for those subjected to algorithms.

Scholars have also challenged the predominant notions of fairness fueling the algorithmic fairness literature, which often rely on achieving an equal distribution of goods and opportunities (i.e., distributive fairness). Fazelpour et al. [45] and Wong [150] have advocated for a more procedural approach to algorithmic fairness, looking at the process through which fairness is defined and achieved. Others have argued for affirmative algorithms [155] and algorithmic reparations [34, 134]. While still focusing on distributive notions of fairness, affirmative algorithms would undo the allocative harms posed by algorithms by considering how the past has shaped the present experience of algorithmic subjects.

Our research builds upon these efforts that recognize the importance of accounting for historical imbalances when discussing algorithmic fairness. Namely, we study how informing people about the existence of such historical biases influences their *perceptions* of algorithmic fairness.

2.2 Perceptions of Algorithmic Fairness

The prior work discussed above represents a normative approach to algorithmic fairness based on scholars' normative claims concerning how algorithms should work and how injustice should be rectified. Another line of research focuses on descriptive notions of algorithmic fairness. Namely, it involves studying what laypeople consider to be fair regarding decision-making algorithms [136].

By capturing people's opinions about algorithmic fairness in a wide range of circumstances, researchers aim to embed societal values in the development of decision-making algorithms [122] and inform the regulation of these systems so that policies do not conflict with lay expectations [6]. Studying lay perceptions of algorithmic harm is also relevant for legislation and ethical guidelines because they are likely to require laypeople to detect unfairness for effective human oversight [91]. Users play an increasing role in algorithm auditing to identify and surface potential errors and unfairness [89, 90, 112], as shown in the case of Twitter/X users collectively discovering that the platform's image-cropping algorithm was favoring White faces over Black faces [68].

Extensive literature has captured laypeople's perceptions of algorithmic fairness. For instance, Grgic-Hlaca et al. [56] identified that people rely on properties such as relevance, volitionality, and privacy when judging whether certain features are fair to be used in algorithms. Focusing on the fairness of the decision-making process itself, Lee [96] found that people consider human decisions more fair than algorithmic ones in tasks perceived as requiring "human skills." Prior work has examined perceived algorithmic fairness in specific domains, such as targeted advertising [120], work evaluation and hiring decisions [93, 96], and bail decision-making [57, 65, 135]. All in all, research suggests that perceptions of algorithmic fairness vary significantly between individuals and the contexts in which algorithms are deployed [62, 133, 136].

Our study examines how providing information about a particular context may impact beliefs and attitudes towards algorithmic discrimination in the same context—a gap unexplored in prior work. Prior work capturing people's perceptions of algorithmic fairness in particular domains rarely introduces how this domain is socially situated. For instance, although studies have captured people's judgments of fairness concerning hiring algorithms [93, 96], they have not examined how shedding light on how hiring decisions have been historically unfair towards certain groups [13, 60, 109, 121, 148] may change how people judge algorithmic hiring decisions. Similarly, studies capturing people's perceptions of algorithmic fairness in the criminal justice domain [7, 56, 57, 65, 135] have not probed participants about historical injustice behind the system, such as their views on mass incarceration of communities of color. Similar to the oversight in computational solutions to algorithmic fairness, using decontextualized lay perceptions of fairness to inform the design, deployment, and regulation of algorithms may reify the same unjust conditions research aims to rectify.

Building upon prior work arguing that a decontextualized analysis of algorithmic fairness may perpetuate and exacerbate existing injustices [12, 45, 54, 78, 102], we study how highlighting the context in which algorithms are developed and deployed may change the way that laypeople perceive algorithmic decisions. More specifically, we study how shedding light on historical injustice in contexts in which algorithms are embedded may impact laypeople's perceptions of algorithms (RQ1).

2.3 Do Laypeople Understand That Algorithms Can Perpetuate Injustices?

The argument that algorithms perpetuate past biases relies on the fact that they *learn from* past decision-making patterns that disproportionately harm specific groups. Unless something is done to address historical injustice, those who have been marginalized will continue to be so, but now through algorithmic means. However, prior work suggests that laypeople do not understand that algorithms learn from past human decisions and thus perpetuate their biases.

Prior work has captured laypeople's mental models of algorithmic systems [79, 116], finding that their understanding of how

algorithms function is often detached from reality [25]. These misperceptions are more common among those without technological expertise [107], who may overlook potential risks due to simplified mental models that do not account for certain threats [77]. All in all, research suggests that laypeople's understanding of AI is deficient, if not incorrect [8]. Looking at work focusing on algorithmic fairness, laypeople tend to consider decision-making algorithms more fair than humans because they believe that algorithms cannot hold prejudices [9] and thus can blindly apply rules without considering who is being judged [16]. These results seem to point out that laypeople may not fully understand that algorithms learn from human biases.

Another line of research examining people's understanding of algorithms investigates how individuals develop algorithmic folk theories to make sense of how such systems work. Although much of the prior work has focused on social media recommendation and moderation algorithms [35, 40, 41, 108], its findings also provide insights into how laypeople may react to algorithmic discrimination. Different people employ varying folk theories that impact how they react to potential algorithmic changes and harms [36]. While some laypeople view algorithms more positively as rational assistants [49], others hold beliefs that algorithmic systems are marginalizing [108], reductive, and exploitative [151], which in turn contribute to their negative views.

Aiming to increase lay understanding of algorithms, research on explainable AI (XAI) has proposed methods to help those subjected to algorithms understand how algorithms make decisions [5]. Prior work suggests that explanations affect people's mental models of algorithms [85] and their judgments of algorithmic decisions with respect to fairness, trust, and other factors [11, 38, 46, 73, 97, 100, 132]. XAI research, however, largely focuses on explaining why an algorithm made a particular decision without addressing how the system works as a whole [101].

Our study explores how explaining that algorithms learn from past human biases (as a way of explaining how algorithms work in general) may impact how laypeople judge an instance of algorithmic discrimination. We manipulate whether participants are told that algorithms *learn from* past human decisions and potentially *perpetuate* biases and study how doing so influences their judgments of algorithmic decision-making. Considering that the effect of shedding light on systemic injustice may hinge on laypeople's understanding that algorithms learn from past human biases, we study whether explaining this possibility moderates the effect of contextualizing algorithms in injustice (RQ2).

2.4 How Identity Shapes Perceptions of Injustice

Another factor that can moderate the effect of shedding light on systemic injustice is one's positionality in relation to injustice. Members of dominant groups, such as men and White individuals, often deny and distance themselves from injustices that grant them advantages as a way to maintain their social position and meritocratic ideals [82]. This trend is stronger for those with a strong identification with their groups [17] and emerges from threats to their group image [82]. Members of these groups also have a psychological motive to defend and justify the status quo for their own benefit according to system justification theory [76]. Considering these protective behaviors, those who benefit from discriminatory algorithms may be more likely to dismiss the historical, social, and political context in which algorithms are embedded, refusing to change their beliefs about algorithm unfairness.

Going beyond identity-based inequality, past social psychology research has also studied how people process information that is negative to them or that contradicts their beliefs. Some studies have found that people tend to resist information that does not align with their beliefs [2], meaning that if one does not believe that historical decisions were unjust, presenting them with evidence of biases in past decisions may induce cognitive dissonance, a discomfort stemming from conflicting information [47, 64]. To alleviate this discomfort, individuals may disregard information about historical injustices that contradict their prior beliefs, a phenomenon known as biased assimilation [104] or disconfirmation bias [39]. Additionally, research shows that people often react negatively to negative feedback, exhibiting defensiveness [103], denial [103], rejection of the feedback [72], and unwillingness to change their behavior [137].

In contrast, other studies suggest that members of dominant groups fail to acknowledge injustices not intentionally but due to their obliviousness about how an unjust state of affairs came to exist [43, 114]. By either educating people about injustice or making it more salient, members of the dominant group may become more supportive of initiatives that try to dismantle the structures fueling injustice [82]. Hence, giving information about historical and social injustice may have stronger effects on the perceptions of individuals from dominant demographic groups by educating them [17, 84]. For instance, Callaghan et al. [22] found that highlighting racial inequality leads to more accurate estimates of the Black-White wealth gap.

Our study explores how one's racial identity may influence how one reacts to information that describes racial injustice. We build upon prior social psychology research looking at human injustice and explore how its findings may translate to algorithmic injustice. More specifically, we study whether one's position in relation to racial injustice moderates the effect of shedding light on systemic injustice on laypeople's perceptions of algorithms (RQ3).

3 Methodology

We conducted a 2x3 between-participants vignette-based experiment. For RQ1, we manipulated whether participants were presented with information concerning historical injustice in a particular domain (context: historical injustice vs. none/control). For RQ2, we varied the way in which we explained how algorithms work (explanation: no explanation/control vs. algorithms *learn* from past human decisions vs. algorithms *perpetuate* human biases). For RQ3, we explored whether one's identity moderates the effect of contextualizing algorithmic discrimination in historical injustice. The study was approved by the authors' Ethical Review Board (ERB), and all of the data and scripts used for analysis are available online: https://tinyurl.com/AIFairPerceptions-Injustice.

3.1 Study Design

Figure 1 presents an overview of our methodology. All study materials are available in Appendix A. Participants were recruited

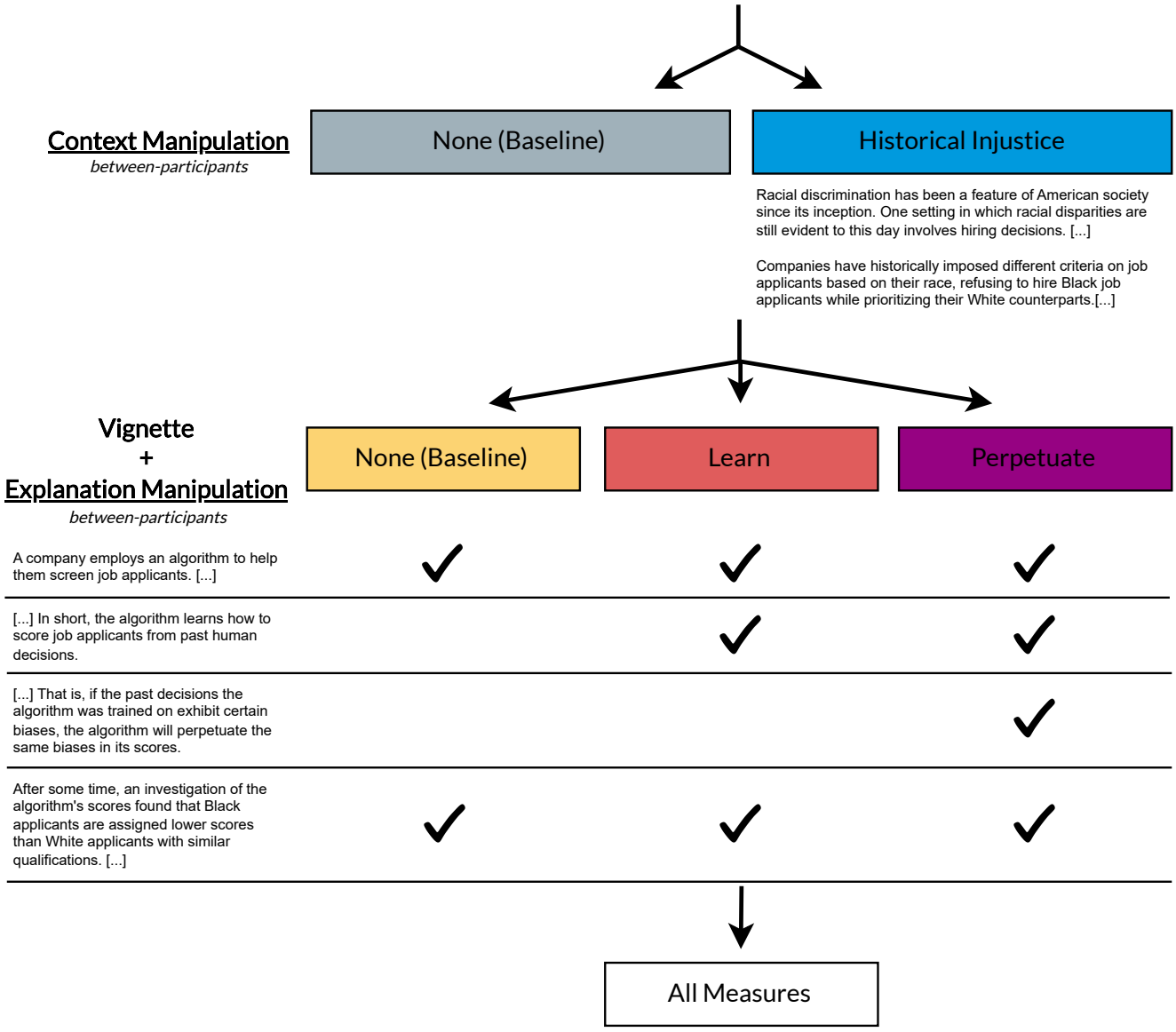


Figure 1: High-level overview of our study methodology. The full text of our vignette and manipulations are presented in Appendix A.

through Prolific [118] to complete a study titled "Survey Study on People's Perceptions of Hiring Decision-Making." After signing the consent form, participants were randomly assigned to one of two *context* treatment conditions. Those assigned to the condition in which we provided information about racial injustice (*context = Historical Injustice* below) read two paragraphs explaining how Black job applicants have been historically (and to this day) discriminated, while White job applicants have been advantaged in hiring decisions. In contrast, those assigned to the control condition (*context = None*) were not shown any information concerning injustice in hiring.

Participants were then shown the study vignette, in which a company screens job applicants with the assistance of an algorithm, which is later found to be discriminatory. Although all participants received the same baseline vignette, some of them were also provided additional information about how the algorithm learns to make decisions according to which *explanation* treatment condition they were randomly assigned.

Participants in the *explanation = Learn* condition were shown an additional paragraph explaining that the algorithm learns to score applicants from past human decisions. This paragraph was placed in between the two paragraphs of the baseline vignette. Another random set of participants was also *explicitly* told that the

algorithm has the potential to replicate past human biases in its scores. Those assigned to this *explanation = Perpetuate* condition read both the paragraph explaining that the algorithm learns from past human decisions and another paragraph explicitly stating that the algorithm can perpetuate human biases.

A few design choices are worth clarifying. We chose to focus on the hiring domain due to the increasing use of decision-making systems in filtering resumes and selecting candidates [14, 33, 71, 146]. We also build upon extensive literature capturing people's perceptions of algorithmic fairness in the hiring domain [92, 93, 96, 119]. Most importantly, there is widespread evidence of identity-based discrimination in the hiring domain across race [60, 109, 121], gender [13], sexual orientation [148], and other identity axes.

We also chose to focus on *racial* discrimination in the United States (US) for our vignette, building upon prior work on how people receive and react to information regarding racial injustice [17, 82]. Our vignette depicts discrimination between White and Black Americans without mentioning other racial groups that have been historically (and continue to be) discriminated against in hiring decisions. We did this for a cleaner experimental design and analysis and to situate our findings in prior social psychology research on the White-Black racial dichotomy in the US [83, 84]. Nonetheless, we call for future work to explore algorithmic injustice across different identity axes and contexts.

3.2 Measures

After reading the vignette, participants answered four groups of questions in the following order:

(1) **Exploratory Variables:** Five exploratory questions concerning the algorithm and its scores (e.g., their perceived objectivity, similarity with past human decisions), which might explain any potential experimental effects observed in our study.
(2) **Main Dependent Variables:** Participants' judgments of fairness concerning the vignette, as well as their reported trust in the algorithm.
(3) **Downstream Effects:** Participants' beliefs about the extent to which the algorithm should be used and redesigned, which were operationalized as downstream effects of their judgments of fairness and trust.
(4) **Background Variables:** Participants' beliefs about whether racial injustice exists, self-reported racial identity, and political leaning.

We viewed the exploratory variables as potential mediators that could help us interpret our experimental findings. As such, we measured them first in the study. In contrast, we collected participants' beliefs in racial injustice after measuring the main dependent variables to avoid their influence on the main dependent variables. Had we asked these questions before fairness and trust, participants in the baseline context condition would also have been prompted to think about racial injustice before answering questions concerning our dependent variables.

While the exploratory variables were intended to serve as potential explanations of any experimental effects from the context and explanation manipulations, we did not observe significant experimental effects on participants' judgments of fairness and trust unless we also accounted for their racial group. Because the exploratory variables were not intended to help explain the racial heterogeneity we report below, we omit the analysis of exploratory variables from the main text for conciseness and instead report it in Appendix C. Nonetheless, we note that our analysis of these exploratory variables is aligned with our findings concerning our main dependent variables (i.e., they also show differences between racial groups).

3.2.1 Main Dependent Variables. We captured participants' perceived fairness of and trust in the algorithm, operationalizing these variables as measures of participants' perceptions of the algorithm and its decisions. We studied participants' fairness perceptions across three dimensions: distributive, procedural, and interpersonal fairness [27]. Distributive fairness refers to whether the distribution of (algorithmic) outcomes is fair across different groups [128, 135]. Procedural fairness looks at whether the decision-making process is consistent, ethical, and unbiased [97, 145]. Interpersonal fairness [123, 129] refers to the respectful treatment of individuals concerning their dignity. Although not as well defined as fairness [143], trust has also been used to capture people's expectations of algorithms [46, 96]. To trust someone is to decide to remain subjected to them [81].

We asked participants the extent to which they agreed with the following statements addressing the perceived fairness of the algorithm, as well as their trust in the system (-3 = Strongly disagree, 3 = Strongly agree). The presentation order of the statement groups was randomized between participants.

- **Distributional Fairness**: 1) "The scores determined by the algorithm are fair" & 2) "The outcome of using the algorithm's scores seems fair" (items adapted from Newman et al. [115]).
- **Procedural Fairness**: 1) "The way that the algorithm calculates scores seems fair" & 2) "The algorithm's process for determining scores is fair" (items adapted from Newman et al. [115]).
- **Interpersonal Fairness**: 1) "The algorithm treats job applicants with respect" & 2) "The algorithm's scores take into consideration the dignity of job applicants."
- **Trust**: 1) "I would strongly rely on the algorithm to calculate scores" & 2) "I trust the algorithm to determine good-quality scores" (items adapted from Thielsch et al. [140]).

Participants were also asked to attribute blame to the algorithm itself, its developer, and its user. We included these questions based on prior work examining blame judgments for instances of algorithmic harm [30, 95, 99]. We note, however, that they do not directly refer to people's perceptions of the algorithm and its scores but instead capture their reactive attitudes resulting from algorithmic harm [100]. For brevity, we omit our analyses of these variables from the main text and report them in Appendix C.

3.2.2 Downstream Effects. Participants were next asked the extent to which they agreed with the two following statements in random order (-3 = Strongly disagree, 3 = Strongly agree):

- **Should It Be Banned?:** "The algorithm should not be determining scores for screening job applicants."

- **Should It Be Changed?:** "The process used by the algorithm to determine scores for screening job applicants should be changed."

These questions aimed to explore whether people's perceptions of fairness and trust had downstream effects on their opinions concerning the deployment of the algorithm. Although examined in a few studies related to algorithmic fairness [1, 105, 131] and trust [46], downstream effects have been largely overlooked by prior work [136], even though they are practically relevant and well-documented in psychology research on trust and justice related to human decision-makers [28, 29].

3.2.3 Background Variables. We also gathered participants' beliefs concerning racial injustice. The following questions were presented in random order:

(1) **Belief in Racial Injustice:** Participants indicated the extent to which they agreed with three statements (e.g., "Hiring decisions are marked by racial disparities to this day") affirming that there exists racial discrimination in hiring decisions (-3 = Strongly disagree, 3 = Strongly agree).
(2) **Inequality of Opportunity:** "Do you think that White Americans have more opportunities than they should, that Black Americans have more opportunities than they should, or that opportunities are about equal between racial groups?" (1 = Black Americans have too much, 4 = Things are about equal, and 7 = White Americans have too much; from Callaghan et al. [22]).

We also measured participants' understanding that algorithms learn from past human decisions to gauge whether our explanation manipulations significantly influenced participants' comprehension of how algorithms work. Because our explanation manipulations had only marginal effects on perceived fairness and trust, we omit the analysis of these variables from the main text and report them in Appendix C.

Participants also indicated whether they had any training or work experience in professions related to machine learning (ML) or artificial intelligence (AI). Finally, they reported their political leaning using a 5-point scale (-2 = Conservative, 2 = Liberal), as well as their racial identity out of the following options: 1) White, 2) Black or African American, 3) Asian, 4) American Indian or Alaska Native, 5) Native Hawaiian or Other Pacific Islander, and Other (please self-describe). These options were based on the US Census questions on race. Participants were also allowed to withhold their racial identity.

3.3 Data Analysis

We used linear regressions to analyze our data. In all models, the treatment conditions to which participants were assigned were treated as dummy independent variables with the control conditions as the baseline. In other words, we regressed participants' responses to dummy variables that are one if they were assigned to treatments other than the controls. As a robustness check, we replicated our results using ordinal regressions and found consistent results (see Appendix C); we report the results of linear regressions below for easier interpretation.

We also explored how our manipulations influenced participants' perceptions of algorithms differently depending on their racial identity. Hence, some regression models also include participants' self-reported race as an independent variable. We mapped responses such that those who self-identified as White were categorized as the "Advantaged" group, while all other participants were grouped into "Disadvantaged." We did this to align our analysis with our vignette, which depicts a scenario in which White applicants are advantaged compared to Black applicants.

Since we recruited participants from the US, this categorization is also aligned with the US context, where most research on race focuses on how White people are privileged in comparison to other racial groups, such as Black [83, 84], Asian [86], and Latinx [51] communities. We note, however, that our "Disadvantaged" group includes a diverse group of participants with a wide range of identities. In the main text, we present results using the Advantaged vs. Disadvantaged dichotomy for simplicity and better statistical power. We show in Appendix C that our main findings are consistent if we analyze our data using more granular categorizations of race.

First, we analyzed participants' judgments concerning fairness and trust irrespective of the experimental manipulations to which they were assigned. Second, we investigated how our manipulations impacted participants' perceived fairness of the algorithm and trust in the system (RQ1, RQ2). Third, we looked at the interaction between our treatment conditions and participants' self-reported racial identity (RQ3). Fourth, we explored how one's beliefs in racial injustice help explain the effect of the manipulations on our main dependent variables using (moderated) mediation models.[1] Finally, we examined the relationship between perceived fairness and trust and the downstream effects of these judgments.

3.4 Participants

We recruited study participants through Prolific [118]. Prolific asks its workers their ethnicity out of the following options: White, Black, Asian, Mixed, and Other. To ensure that we obtained a balanced number of responses across White participants (categorized as Advantaged) and those from other racial groups (categorized as Disadvantaged), we recruited participants through two concurrent studies targeting workers who self-identified as White and other racial groups, respectively.

We conducted a power analysis to determine our sample size. A two-tailed t-test requires 105 respondents per treatment group to detect a medium effect size (Cohen's d=0.5) at the significance level of 0.05 with 0.95 power. Considering that we have six treatment conditions, we aimed to recruit at least 662 participants, assuming a 5% attention-check failure rate.

In total, we recruited 725 participants, out of which 362 had self-identified as White on Prolific. Participants were required to be US residents, fluent in English, and have completed at least 50 tasks on Prolific with an approval rate of over 95%. We sampled participants at different hours over several days to mitigate sampling biases that may occur due to time [24]. We discarded responses from nine participants who did not pass an instructed response

[1]For our mediation analysis, we use model 5 from the PROCESS macro by Hayes [66] with 5000 bootstrap iterations.

	(1)	(2)	(3)	(4)
(1) Distributional Fairness	1	0.877	0.686	0.831
(2) Procedural Fairness	0.877	1	0.656	0.826
(3) Interpersonal Fairness	0.686	0.656	1	0.685
(4) Trust	0.831	0.826	0.685	1

Table 1: Pearson's (r) correlation matrix of the dependent variables.

question or failed to identify that the vignette that they read focused on hiring decisions. Thus, our final sample comprised 716 participants. All participants were paid 1.60 GBP (approximately 2.00 USD) for their participation, with a median pay of 12.57 GBP per hour (approximately 16.00 USD per hour).

We did not ask participants for their demographic information other than their racial identity, political orientation, and prior training in AI-related fields. Nonetheless, Prolific also keeps track of their workers' self-reported gender and age information. Our sample comprised 361 (50.42%) women—and 355 (49.58%) men—with an average age of 36.90 years old (SD = 12.08). By design, nearly half of the participants were classified as "Advantaged" ($n_{\text{Adv.}}$=372, 51.96%) while the remaining were classified as "Disadvantaged" ($n_{\text{Dis.}}$=344, 48.04%). Within the latter group, 127 (36.92%) self-identified as Black/African American, 123 (35.75%) as Asian, 39 (11.34%) as Mixed, and 55 (15.99%) either chose other racial group or decided to withhold this information. Our sample leaned slightly liberal (M=0.63, SD=1.17), and most participants (80.72%) did not have any training in AI-related disciplines.[2] We conducted a follow-up study to gather participants' additional demographic information—as well as their racial identity and political orientation—and report this information in Appendix B.

4 Results

4.1 Perceived Fairness of and Trust in the Algorithm

Participants judged the vignette as largely unfair concerning its distributive (M=−1.41, SD=1.51, Cronbach's α=0.954), procedural (M=−1.28, SD=1.54, α=0.966), and interpersonal aspects (M=−1.33, SD=1.38, α= 0.844). Moreover, participants suggested that they do not trust the algorithm to make screening decisions (M=−1.40, SD=1.52, α=0.946). Table 1 presents the correlation matrix between these four measures. All measures are moderate to highly associated with each other, with interpersonal fairness exhibiting a slightly lower correlation with the other variables. These results are consistent with theory and empirical research showing that these dimensions of fairness are usually correlated—both in terms of people's perceptions about fairness more broadly [27, 28], as well as in the context of algorithmic fairness [11]. Similarly, these findings are in line with prior work showing that the perceived fairness of an algorithm is associated with trust in it [96, 136].

[2]We reran our analysis including participants' reported experience and training in AI-related fields (or lack thereof) as a covariate and found consistent results. Moreover, we conducted sub-group analyses and also observed results in line with what we report in the paper, barring some changes in the significance level of our results due to smaller sample sizes.

4.2 Experimental Effects on Perceived Fairness and Trust (RQ1, RQ2)

Table 2 presents the effect of our experimental conditions on perceived fairness and trust. Surprisingly, none of the four measures were influenced by the context manipulation, explanation manipulation, or their interaction (see Figure 2). In other words, highlighting that the context in which the algorithm is deployed is racially unjust did not impact how participants judged a particular instance of algorithmic racial discrimination. Moreover, explaining that algorithms learn from past human decisions and may perpetuate human biases also did not influence how people perceived biased algorithmic decisions.

4.3 Differences Between Racial Groups on Perceived Fairness and Trust (RQ3)

We also explored whether participants' self-reported race is associated with their perceptions of the algorithm and whether it interacts with our experimental manipulations. Table 3 shows that providing information about systemic injustice has a significant effect on the perceived fairness of the algorithm when we account for participants' racial group. We identified an interaction between these two factors across all four variables, suggesting that our context manipulation consistently produced heterogeneous effects on people's perceptions of algorithmic discrimination depending on their racial group.

Figure 3 presents the perceived fairness of the algorithm, as well as participants' trust in it depending on their self-reported race and the context condition to which they were assigned. Among participants that were not shown our context manipulation, we only observed differences between racial groups in judgments of procedural fairness (see *group = Advantaged* in Table 3), such that those categorized as Advantaged considered the algorithm more procedurally fair than those in the Disadvantaged group. Our results thus suggest that people from different racial groups largely agree on perceived (distributive and interpersonal) fairness and how much they trust the algorithm when its deployment is not contextualized in systemic injustice.

Focusing on those who read about systemic racial injustice, we found that those categorized as *Disadvantaged* judged the algorithm to be **more** fair across the distributive, procedural, and interpersonal dimensions (see *context = Historical Injustice* in Table 3). In contrast, participants labeled as *Advantaged* considered the algorithm **less** fair and trustworthy when exposed to the same information (see *(context = Historical Injustice):(group = Advantaged)* in Table 3). In other words, participants from minority groups became *more positive* about algorithmic racial discrimination when contextualized in systemic injustice, while those from the Advantaged group exhibited the opposite trend. Although participants from the Advantaged and Disadvantaged groups largely agreed on fairness and trust when not shown information about systemic injustice, their views started to diverge when the hiring algorithm was contextualized in historical discrimination.

In terms of interaction effects between the explanation manipulation and participants' race, Figure 3 shows that participants from the Advantaged group consider the algorithm less distributively fair

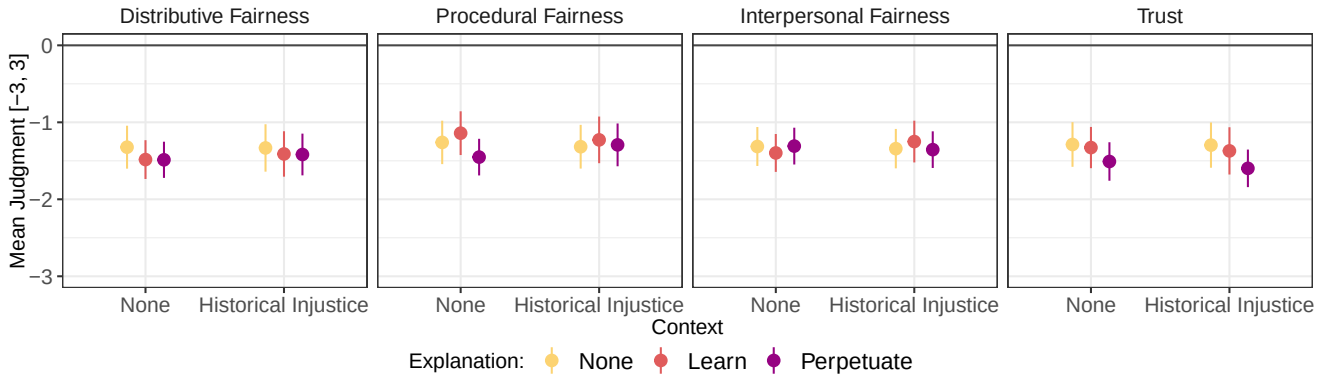


Figure 2: Perceived fairness of and trust in the algorithm depending on the treatment condition. Participants either did not receive any information about systemic injustice in the hiring domain (Context = *None*) or read two paragraphs explaining how Black job applicants have been (and continue to be) systematically disadvantaged in hiring decisions (Context = Historical Injustice). Our vignette either did not explain how the algorithm learns how to make decisions (Explanation = *None*), stated that the algorithm learns from past human decisions (Explanation = *Learn*), or explicitly mentioned that the algorithm has the potential to perpetuate past human biases (Explanation = *Perpetuate*). Error bars correspond to 95% confidence intervals.

	Dependent Variables			
	Distributional Fairness	Procedural Fairness	Interpersonal Fairness	Trust
	(1)	(2)	(3)	(4)
context = Historical Injustice	−0.010	−0.056	−0.028	−0.008
	(0.198)	(0.202)	(0.181)	(0.199)
explanation = Learn	−0.161	0.120	−0.084	−0.041
	(0.195)	(0.199)	(0.178)	(0.196)
explanation = Perpetuate	−0.164	−0.191	0.005	−0.221
	(0.201)	(0.204)	(0.183)	(0.201)
(context = Historical Injustice):(explanation = Learn)	0.084	−0.031	0.176	−0.034
	(0.279)	(0.284)	(0.255)	(0.280)
(context = Historical Injustice):(explanation = Perpetuate)	0.079	0.215	−0.019	−0.081
	(0.278)	(0.283)	(0.254)	(0.279)
Constant	−1.323***	−1.261***	−1.314***	−1.288***
	(0.142)	(0.145)	(0.130)	(0.143)
Observations	716	716	716	716

Table 2: Linear regressions of perceived fairness and trust. Dependent variables: perceived distributional fairness, procedural fairness, interpersonal fairness, and trust in the algorithm. Independent variables: dummy variables (context and explanation) indicating to which treatment condition participants' were assigned. Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

and trustworthy when explicitly told that algorithms can perpetuate human bias. This finding reflects a trend similar to that found for the context manipulation—people from the Advantaged group became less positive about the algorithm when receiving more information concerning injustice. We present the full regression table in Appendix C for brevity since most coefficients are non-significant. The only significant coefficients refer to the interaction between racial group and the Perpetuate explanation condition in relation to distributive fairness (b=−0.562, SE=0.277, p<.05) and trust (b=−0.692, SE=0.277, p<.05).

4.4 Exploratory Analysis: Race as a Proxy For Beliefs

Up to this point, we have operationalized participants' racial group (Advantaged vs. Disadvantaged) as a moderator in our analysis. With this approach, we view a participant's race as a factor that determines their perceptions of discriminatory algorithms and leads to heterogeneous experimental effects. Although this approach allows us to observe differences across racial groups, it assumes an essentialist conception of race [124], i.e., race is a fixed property

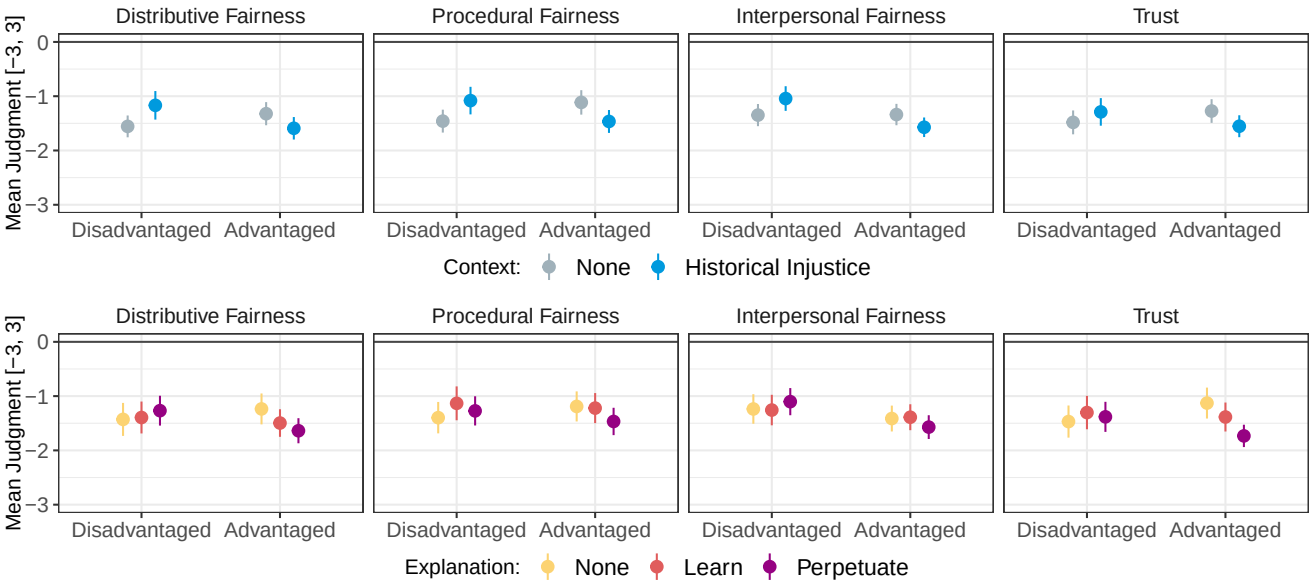


Figure 3: Perceived fairness of and trust in the algorithm depending on the treatment conditions and the participants' racial group (Advantaged = White, Disadvantaged = all others). Participants either did not receive any information about systemic injustice in the hiring domain (Context = *None*) or read two paragraphs explaining how Black job applicants have been (and continue to be) systematically disadvantaged in hiring decisions (Context = Historical Injustice). Our experimental vignette either did not explain how the algorithm learns how to make decisions (Explanation = *None*), stated that the algorithm learns from past human decisions (Explanation = *Learn*), or explicitly mentioned that the algorithm has the potential to perpetuate past human biases (Explanation = *Perpetuate*). Error bars correspond to 95% confidence intervals.

	Dependent Variables			
	Distributional Fairness	Procedural Fairness	Interpersonal Fairness	Trust
	(1)	(2)	(3)	(4)
context = Historical Injustice	0.388*	0.378*	0.305*	0.193
	(0.162)	(0.165)	(0.148)	(0.163)
group = Advantaged	0.234	0.346*	0.010	0.209
	(0.159)	(0.162)	(0.145)	(0.161)
(context = Historical Injustice):(group = Advantaged)	−0.657**	−0.730**	−0.539**	−0.474*
	(0.225)	(0.229)	(0.205)	(0.227)
Constant	−1.556***	−1.459***	−1.348***	−1.482***
	(0.115)	(0.117)	(0.105)	(0.116)
Observations	716	716	716	716

Table 3: Linear regressions of perceived fairness and trust. Dependent variables: perceived distributional fairness, procedural fairness, interpersonal fairness, and trust in the algorithm. Independent variables: dummy variables (context) indicating to which treatment condition participants' were assigned and participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

that determines a person's characteristics, beliefs, and opinions, including judgments regarding discriminatory algorithms.

In contrast, the constructivist view of race offers another possibility: one's racial identity does not directly determine their beliefs and judgments but instead works as a proxy for their social position in a system of privilege and oppression [124]. This social position shapes an individual's social outcomes, experiences, and beliefs, which in turn influence their views of specific circumstances [70]. Taking the constructivist view of race to our study, one's racial identity will not directly determine their judgments concerning algorithmic discrimination or reactions to our manipulation; instead, a significant effect of race suggests that there might be some

underlying factors captured by racial identity that are correlated with perceptions of algorithms.

We studied three factors for which race is a proxy. We explored participants' beliefs that racial injustice and inequality of opportunity exist—both of which are strongly associated with racial identity [17, 114, 154]—as well as their political orientation—which can also influence people's reaction to race-salient manipulations [19]. On average, participants agreed that racial injustice is a problem in hiring (M=1.91, SD=1.17, Cronbach's α=0.944) and affirmed that White individuals have more advantages than Black individuals (M=5.51, SD=1.18).

We first examined whether these three factors are associated with people's racial identity and our context manipulation (see Table 4). In comparison to those in the Advantaged group, participants in the Disadvantaged group agreed to a larger extent that racial injustice exists and were more likely to affirm that White individuals have more opportunities than Black individuals. These results are aligned with prior work [22, 84]. Our context manipulation had no statistically significant effect on participants' belief in racial injustice and unequal opportunities.

Concerning political orientation, participants in the Advantaged group reported being more conservative than those in the Disadvantaged group. Surprisingly, we also observed an interaction between racial group and our context manipulation. Participants in the Advantaged group who were shown additional information about racial injustice in hiring decisions reported being more liberal than those in the same racial group but assigned to the control context condition. Additional analyses indicate that this interaction effect may have been caused by an imperfect random assignment of participants instead of being an effect of our context manipulation (see Appendix D for a more detailed discussion). Given that we account for participants' race in our analysis—and the fact that the indirect effect of political learning on judgments of fairness and trust is non-significant, as shown below—we do not expect this bias to influence our findings.

Finally, we examined whether these three factors help unravel the underlying reasons behind the observed heterogeneous effects across racial groups using the mediation model in Figure 4. We had initially decided to use a moderated mediation model to account for the interaction between racial group and our context manipulation (see the gray dashed lines in Figure 4). However, the context manipulation did not moderate any indirect effects. Instead, it only moderated the direct effect of the racial group on the dependent variables. This moderation of the direct effect suggests that other unobserved factors interact with our context manipulation to determine people's perceptions of algorithms. We decided to retain this moderation of the direct effect in our final model and not to include any other moderation effects. Consequently, we report the results of a parallel mediation model, in which differences across racial groups can be explained by differences in how much participants believe in racial injustice, in inequality of opportunity, and by participants' political orientation (see solid lines in Figure 4).

We present the estimated direct and indirect effects of the mediation model in Table 13 in Appendix C for conciseness. The extent to which participants believe in racial injustice is a significant mediator of racial differences for judgments of distributive fairness, procedural fairness, and trustworthiness. Perceptions of inequality of opportunities between White and Black individuals also help explain differences across racial groups in for all dependent variables. In contrast, participants' political orientation does not mediate any racial differences.

4.5 Exploratory Analysis: Downstream Effects of Perceived Fairness and Trust

Finally, we conducted an exploratory analysis of the potential downstream effects of people's judgments of fairness and trust. Participants somewhat agreed that the algorithm depicted in the vignette should *not* be used for screening job applicants (M=1.34, SD=1.57). Furthermore, we found that participants believe that the algorithm should be changed (M=1.95, SD=1.30).

Similar to our analysis of perceived fairness and trust, we did not identify any significant effects of our context and explanation manipulations (and their interaction) on participants' belief that the algorithm should be changed or used. However, our results here are consistent with the heterogeneous effects of our context manipulation depending on participants' racial group. Those categorized as Advantaged became more likely to support that the algorithm should be redesigned (b=0.425, SE=0.194, p<.05) or not used at all (b=0.660, SE=0.235, p<.01) when it was contextualized in systemic injustice. In contrast, participants in the Disadvantaged group who were shown the context manipulation reported more positive attitudes towards the algorithm, indicating that it should be banned to a lesser extent (b=−0.352, SE=0.169, p<.05). We present all of these results in detail in Appendix C.

To model the relationship between fairness and trust and people's opinions concerning whether the algorithm should be changed or used, we regressed these potential downstream effects to participants' judgments of fairness and trust. Table 5 shows that, when we account for perceived fairness and trust, the context manipulation and its interaction with participants' racial group become non-significant. We also find that trust is strongly associated with people's agreement that the algorithm should not be used. Similarly, participants' judgments of distributional fairness, procedural fairness, and trust are correlated with participants' belief that the algorithm should be redesigned.

5 Discussion

Below, we contextualize participants' negative attitudes towards algorithmic discrimination in prior work and discuss their downstream effects on people's opinions concerning the deployment of algorithms (§5.1). We examine how our null effects of explaining how algorithms work on perceived fairness and trust demonstrate the complexity involved in explaining how algorithms work to laypeople, adding nuances to calls for the development of explainable algorithms (§5.2). We then discuss how providing information about systemic injustice may have unintended consequences to lay perceptions of algorithms, leading to disagreements between different racial groups by making members of historically disadvantaged groups more positive towards algorithmic discrimination (5.3). Finally, we reflect on our work's limitations (§5.4) and share some concluding remarks (§6).

	Dependent Variables		
	Belief in Racial Injustice	Inequality of Opportunity	Political Orientation
	(1)	(2)	(3)
context = Historical Injustice	0.013	−0.003	−0.142
	(0.125)	(0.124)	(0.125)
group = Advantaged	−0.316*	−0.622***	−0.350**
	(0.123)	(0.122)	(0.123)
(context = Historical Injustice):(group = Advantaged)	0.222	0.162	0.442*
	(0.174)	(0.172)	(0.174)
Constant	2.006***	5.795***	0.766***
	(0.089)	(0.088)	(0.089)
Observations	716	716	716

Table 4: Linear regressions of participants' belief in racial injustice, belief in inequality of opportunity between White and Black people, and political orientation. Dependent variables: participants' agreement that racial injustice exists in the hiring domain; participants' views concerning who has more opportunities between Black and White Americans; political orientation (-2 = Conservative, 2 = Liberal). Independent variables: dummy variables (context) indicating to which treatment condition participants were assigned and participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

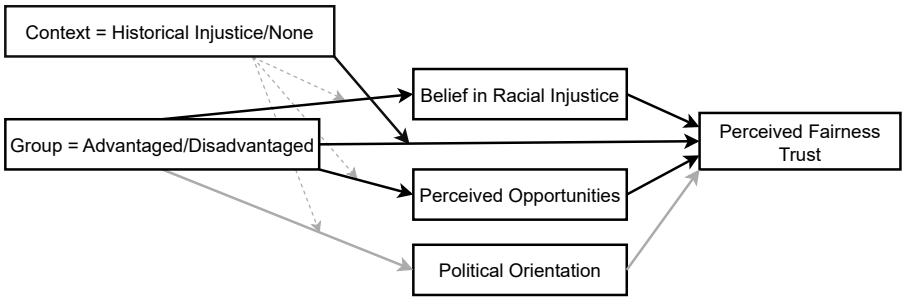


Figure 4: Mediation model exploring the underlying reasons behind the observed varying perceptions of discriminatory algorithms across racial groups. Gray lines are not statistically significant, and dashed lines are omitted from the final model.

5.1 Negative Views Towards Algorithmic Discrimination

Across all experimental conditions, participants judged the algorithm as largely unfair concerning its distributive, procedural, and interpersonal aspects. Similarly, participants indicated that they would not trust the algorithm to make hiring decisions. Our findings are aligned with prior work showing that people are averse to algorithms making consequential decisions [21, 37], including in the hiring domain [46, 96]. We note that our vignette described an algorithm that was explicitly discriminatory against a racial minority; hence, such negative perceptions are somewhat expected. Nonetheless, our findings underscore the importance of algorithmic fairness on people's perceptions of algorithms: people will not trust algorithms if they are discriminatory.

We did not identify significant differences between racial groups in most baseline judgments concerning the algorithm. Participants in the Advantaged and Disadvantaged groups *who did not read about systemic injustice in hiring* judged the algorithm as similarly (un)fair concerning its distributive and interpersonal components and reported (not) trusting it to a similar extent. This result is surprising given prior work suggesting that people from different racial groups have distinct intuitions about what counts as discrimination and the severity of racial injustice [23]. In the context of algorithmic decision-making, however, this finding is in line with research showing that race has no significant effect on people's perceptions of algorithmic fairness [55]. All in all, our findings indicate that people—no matter their racial identity—denounce algorithmic discrimination to a similar extent when not contextualized within systemic injustice.

We also explored whether fairness and trust have downstream effects on people's opinions regarding the deployment of algorithms—an aspect that has received considerably less attention in prior work [136]. Consistent with low perceived fairness and trust, participants somewhat agreed that the algorithm in the vignette should be changed or not used at all, in line with the conceptualization of trust as one's willingness to remain subjected to a decision-maker [81]. Because people did not trust the system, they believed

	Dependent Variables	
	Should It Be Banned?	Should It Be Changed?
	(1)	(2)
Distributional Fairness	−0.093	−0.293***
	(0.073)	(0.051)
Procedural Fairness	−0.092	−0.149**
	(0.069)	(0.049)
Interpersonal Fairness	−0.052	0.004
	(0.050)	(0.035)
Trust	−0.417***	−0.213***
	(0.062)	(0.044)
context = Historical Injustice	−0.185	0.065
	(0.138)	(0.098)
group = Advantaged	−0.111	−0.020
	(0.135)	(0.096)
(context = Historical Injustice):(group = Advantaged)	0.306	0.025
	(0.191)	(0.136)
Constant	0.505***	1.024***
	(0.110)	(0.078)
Observations	716	716

Table 5: Linear regressions of participants' belief that the algorithm should be banned and redesigned. Dependent variables: participants agreement with a statement indicating that the system should not be used and another statement affirming that the algorithm should be changed. Independent variables: perceived fairness; reported trust; a dummy variable (context) indicating to which treatment condition participants' were assigned and participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

that it should not be used to evaluate job applicants (and potentially them). In contrast, perceived fairness was a significant predictor only for participants' agreement that the algorithm should be changed rather than banned. Nonetheless, we caution that all of our variables around perceived fairness and trust are strongly correlated, meaning that it is impossible to disentangle which of these factors is the main determining factor of participants' belief that the algorithm should be changed or banned.

Our results are aligned with findings from Acikgoz et al. [1] on different downstream effects of perceived algorithmic fairness: in their study, higher perceived fairness decreases the likelihood that one will pursue litigation against a company using a hiring algorithm and increases one's intention to accept a job at the same company [1]. Our observed downstream effects are also in line with the findings by Marcinkowski et al. [105], i.e., lower perceived fairness is associated with intentions to protest against an algorithm and withdraw from algorithmic decisions. All in all, lower perceived fairness leads to more negative attitudes towards algorithms, ranging from beliefs that the system should not be used at all (as found in our study) to reported intentions to protest against algorithmic decision-making.

Given these negative perceptions, it is possible that the adoption of algorithms, particularly in the hiring domain, will be challenging when specific instances of harm are made public. Algorithms are becoming increasingly common in hiring processes, with 99% of Fortune 500 companies employing similar systems in some part of their selection processes [71]. These systems' limitations are widely known, with biases found in several steps of the hiring process [14]. Job applicants have varying levels of knowledge about hiring algorithms, with some comprehending that algorithms can perpetuate inequalities [4]. The numerous examples of how algorithms can be biased, in hiring and other domains [3, 33, 42, 67, 117], may thus make it difficult to convince those subjected to algorithms to accept these systems in the first place. As suggested by our work, these harms may even lead people to indicate that these systems should not be used at all, potentially even if they are already deployed [74].

5.2 Explaining How Algorithms Work Has Little to No Effect on Fairness and Trust

Critical algorithmic studies argue that algorithms that do not account for the context in which they are deployed can exacerbate injustice [12, 45, 54, 78, 102]. However, this argument relies on the understanding that algorithms learn from—and thus perpetuate—past human biases. Hence, our research also explored whether highlighting how algorithms can replicate biased human decision-making has any impact on people's perceptions of discriminatory algorithms. We found little to no evidence that explicitly mentioning this fact influences lay perceptions.

A potential explanation of our mostly null effects is that laypeople already understand how algorithms can perpetuate biases. However, this hypothesis is not completely aligned with some additional analysis in Appendix C (see Table 8), which shows that exposure to the *Learn* and *Perpetuate* manipulations increased participants' belief that algorithms perpetuate biases compared to the control condition. In other words, our explanation conditions increased

participants' understanding that algorithms learn past human biases, but this increased understanding had no effect on judgments of fairness and trust. Another possibility is that our vignette makes it implicitly clear that algorithms perpetuate biases by portraying an algorithm aligned with an unjust status quo. Even without the vignette telling people that algorithms perpetuate past biases, participants may have made this connection between their prior knowledge about injustice and our vignette portraying a discriminatory algorithm. Future work is needed to explore whether laypeople comprehend how algorithms work and how this knowledge may impact their judgments of specific instances of algorithmic harm.

Our results demonstrate the difficulty in explaining how algorithms work to laypeople. Prior work suggests that laypeople have false mental models about algorithms and AI models [8, 25]. These misconceptions seem to be difficult to correct given that explaining how algorithms work had null effects on people's judgments of fairness and trust, as shown by our findings. In our study's particular context, explaining that algorithms can perpetuate injustice can introduce even more nuances given widespread misperceptions about racial and economic inequality [83, 84]. Teaching laypeople about algorithms is hard; explaining that algorithms can perpetuate existing injustices is an even bigger challenge.

Our findings add more nuances to calls for the development of algorithms that are explainable and interpretable [5]. Explainable AI has been put forward by academics [94], industry leaders [113], and policymakers [32] as a potential solution to issues posed by decision-making algorithms, such as algorithmic discrimination [20]. However, we found that explaining how an algorithm works had no effect on laypeople's perceptions of the algorithm. Our findings are also contrary to prior work showing that explanations of specific algorithmic decisions impact lay perceptions of algorithms in hiring [11, 132] and other domains [38, 97, 100], calling into question the importance of interpretability when assessing the fairness and trustworthiness of discriminatory algorithms.

However, we call attention to a distinction between our approach and how prior work explored explainability. Past research has largely examined the effect of explaining why the algorithm made a particular decision [101]; in contrast, our manipulation was designed to educate participants about how the algorithm as a whole works, i.e., by learning from past human decisions. It is possible that educating people about the limitations of algorithmic decision-making requires interventions that are more complex than the textual manipulation we explored in this study. Future work could examine different ways of educating laypeople about potential algorithmic biases. For instance, studies could contrast the effect of textual manipulations with longitudinal interactions with discriminatory algorithms [46], which could indirectly teach people that algorithms are biased without saying so explicitly.

Our findings also have implications for policymaking. Algorithmic transparency is a common proposition across regulations around the world. For instance, the European Union (EU) AI Act demands systems that are deemed high-risk to have clear instructions regarding how they work and should be used. Similarly, the United States (US) AI Bill of Rights posits that algorithms should be accompanied by "accessible plain language documentation including clear descriptions of the overall system functioning." Regulatory calls for human oversight that would require users to identify and report unfairness, as proposed by the EU AI Act [91], might not be successful if educating people about how algorithms work has little to no impact on their perceptions of fairness.

5.3 Highlighting Systemic Injustice Makes Advantaged Groups More Negative—and Disadvantaged Groups More Positive—About Algorithmic Discrimination

Our findings show that participants from different racial groups reacted differently to information about systemic injustice: participants belonging to the racial group portrayed as advantaged in our manipulation became more negative about algorithmic injustice, whereas participants from disadvantaged groups became more positive, even though the algorithm was discriminatory against one particular disadvantaged group. Although people initially agreed on algorithmic fairness and trustworthiness when decontextualized from systemic injustice, the context manipulation sparked disagreements between individuals depending on their positionality with respect to injustice.

Our study also found disagreements in people's opinions concerning the algorithm's deployment depending on their positionality. However, this heterogeneity disappeared when we accounted for judgments of fairness and trust. This result suggests that the varying effects of shedding light on systemic injustice on people's attitudes towards the deployment of algorithms can be explained by the heterogeneous effects of the manipulation on perceived fairness and trust. In other words, contextualizing algorithms in systemic injustice leads to disagreements concerning fairness and trust between racial groups, which in turn lead to their contrasting opinions about algorithms being redesigned and banned.

Although prior work has also shown that individuals from different racial groups perceive racial injustice differently [23, 84], the directions of change found by our study are unexpected and worth exploring further. Considering that racially advantaged groups may reject the notion that they are privileged [82] and try to justify injustice was an attempt to maintain it [75], we expected participants from the Advantaged group to either not change their views concerning the algorithm or potentially become more positive towards it since it perpetuates a status quo that favors them. Below, we provide a few possible explanations: some are partially supported or contradicted by our results; others leave room for further exploration in future work.

Explanation 1: Advantaged Participants Were Taught or Reminded of Racial Injustice: Participants in the Advantaged group could have been initially unaware of racial injustice in hiring, with our manipulation teaching them about its existence [15, 114, 154]. Although this explanation seems plausible at first, it is not entirely aligned with our results, which shows that our context manipulation did not increase the extent to which people believe in racial injustice (see Table 4). It may also be that our manipulation did not "teach" participants about racial injustice but instead served as a reminder and made the concept more salient. This increased attention could have made people more supportive of behaviors that attempt to dismantle racial injustice [82] by denouncing algorithmic discrimination.

This interpretation could also help explain our heterogeneous effects. Because members of the Disadvantaged group are already more likely to experience discrimination, they might naturally think of racial injustice when they read about a particular instance of discrimination; in contrast, those in the Advantaged group may have to be reminded that systemic injustice exists.

Explanation 2: Racial Injustice Renders Algorithms as Less Biased Than Humans: When looking at participants from the Disadvantaged group, we identified that they became more positive about algorithmic discrimination, even when it was against members of their own group. This unexpected result might have emerged from the fact that *human* decision-makers were underscored in our context manipulation. By highlighting that *humans* have been and continue to be biased against racial minorities, algorithms started to seem relatively more fair and trustworthy, even if they tended to replicate past biases. This hypothesis is aligned with prior work in the medical domain: emphasizing that human doctors can be biased increases people's support for algorithmic systems in medicine [10]. Similarly, past research on hiring has shown that women are relatively more supportive of algorithms making hiring decisions when the alternative is to be evaluated by men [119].

The first two potential explanations can also be formulated through psychological theories on how people make moral judgments. Gray and Pratt [52] argue that all moral judgments—including those about fairness violations—are based on comparisons with a template of an agent causing harm to a patient. The theory acknowledges that moral judgments have no ground truth and posits that different judgments emerge from disagreements about who is a vulnerable patient and/or an intentional agent. Our context manipulation could have made the vulnerability of Black job applicants more salient to participants from the Advantaged group, decreasing perceived fairness. In contrast, given past first-hand experiences of discrimination by members of the Disadvantaged group, our manipulation could have instead highlighted the role of human agents in past discrimination, portraying algorithms as less harmful agents and thus increasing perceived fairness. Future work can investigate the effect of varying the agents and patients involved in algorithmic discrimination to uncover the underlying explanation of our heterogeneous effects.

Explanation 3: Different Folk Theories of Algorithms: Analyzing our results through the lens of algorithmic folk theories [36, 49], our manipulations could have fueled varying folk theories depending on participants' positionality. While participants from the Disadvantaged group could have pictured algorithms as "rational" decision-makers [49] (in comparison to humans), highlighting that algorithms are embedded in unjust social structures could have prompted those from the Advantaged group to theorize algorithms as "exploitative" [151]. Future work can scrutinize the folk theories people hold about algorithms—especially considering how they may vary based on one's positionality—and link them to our observed heterogeneous effects.

It is also imperative that future work examines the mechanism through which these folk theories are created. For instance, scholars could scrutinize the discourse that advertises algorithms as solutions to human-created problems [88, 106], which could shape people's initial expectations of algorithms as objective decision-makers. Furthermore, studies could explore whether explicitly describing that existing algorithms—instead of humans—are systematically biased has a different effect. For instance, providing concrete examples of algorithms perpetuating systemic injustice could be more effective in turning laypeople more critical of algorithmic decision-making.

Explanation 4: Varying Beliefs in Racial Injustice Between Racial Groups: We also explored whether participants' beliefs in racial injustice and inequality of opportunity in hiring can help explain the observed heterogeneous effects. Although these beliefs were strongly associated with judgments of fairness and trust—as suggested by prior work on human discrimination [125, 126, 154]—they were not influenced by our manipulation describing historical injustice in hiring. In other words, we found that these beliefs are not easily mutable. Although beliefs in racial injustice and inequality of opportunity helped explain differences between racial groups, it did not provide a full picture of how our manipulation impacted those categorized as Advantaged and Disadvantaged differently.

Looking Forward: A promising approach to explore potential explanations of our heterogeneous effects relies on qualitative methods. Instead of restricting people's reactions to algorithmic discrimination to a set of measures, such as fairness and trust, qualitative methods have the potential to examine which specific concepts individuals bring up when interpreting instances of algorithmic injustice. Future work could, for instance, interview participants with different positionalities to examine how they interpret experimental manipulations that contextualize algorithms in systemic injustice. Such interviews could then identify, for instance, whether people use different folk theories when judging algorithmic discrimination or how individuals judge the vulnerability of victims of fairness violations.

Our finding that some people become less critical of algorithmic discrimination when contextualized in systemic injustice calls for a reflection of the approaches advocated by critics of computational algorithmic fairness. Many scholars from this stream of research argue that a more socially situated analysis of algorithmic discrimination should make individuals more critical of computational solutions to injustice [12, 45, 54, 78, 102]. Yet, our findings point to the opposite direction for some individuals. We thus raise the question of whether highlighting systemic injustice in the context of algorithmic decision-making could inadvertently make some people more supportive of algorithmic systems that perpetuate injustice behind a veil of computational objectivity.

5.4 Limitations

All participants, regardless of their racial group, were largely negative towards the algorithm depicted in the study vignette. Thus, the small effect sizes that we found for our experimental manipulations may result from a floor effect. Future work could try replicating our study using more neutral vignettes, in which the algorithm is not explicitly described as being discriminatory or using more subtle manipulations such as showing people algorithmic outputs that are more ambiguous regarding whether they are discriminator as done by Langer et al. [92].

The small effect sizes may also originate from the simplicity of our context manipulation. Future work could compare different ways and formats of doing the manipulation, such as basing the intervention on inequality data [22]; letting participants visualize racial inequality through interactive tools [145]; simulating (dis)advantages experimentally within the context of the study [144]; and using large language models to influence participants' belief about racial injustice. Inspirations for this latter study can be drawn from prior work using a similar approach on political opinions and belief in conspiracy theories [31, 59, 127].

It is also possible that some of our findings are the result of experimental demands: because our manipulation and vignette were clearly negative towards racial discrimination, participants could have answered our questions to fit the framing. It is, however, noteworthy that if there really would have been strong experimental demands in our data, we would have expected this trend for all participants, not only those in the Advantaged group. Another potential limitation of our study comes from social desirability biases [58], which are a common limitation of survey studies that rely on self-reported measures. It is possible that these biases might have influenced participants to report more negative opinions about algorithms to be perceived more favorably in the context of the study. Future studies could try to mitigate this potential bias by relying on more implicit manipulations and examining behavioral responses rather than self-reported measures.

Our vignette is also limited to hiring and racial inequality in the US. Our context manipulation is also restricted to two racial groups, and it is unclear whether our result would replicate had it introduced injustice between other groups. Future studies could replicate our study with different vignettes, in different countries, and using different identity-based injustices.

6 Concluding Remarks

Our study shows that although people from different racial groups may initially agree on the perceived fairness and trustworthiness of an algorithm, framing algorithmic discrimination as part of a systemic problem can spark disagreements between them. These disagreements are not restricted to people's perceptions of algorithms but also relevant to their opinions concerning whether algorithms should even be deployed. Our findings mirror discussions surrounding injustice and how to rectify it; while some argue that enough has been done to remedy past harms, others believe more is needed to ensure that individuals are treated equally [18].

Future debates on how to address inequality—caused through algorithmic means or otherwise—should consider how different ways of framing injustice impact how people perceive it. This framing effect is particularly relevant for those proposing solutions to injustice. In the same way that different ways of framing algorithmic fairness interventions may change the arguments one can use to defend their ideas [147], discussing algorithmic discrimination as either a one-time bug or as a symptom of a systemic problem will change how these systems are perceived and hence which solutions are put forward.

In this paper, we explored how a more socially situated analysis of algorithmic discrimination could impact how laypeople perceive injustice perpetuated by algorithms. Our results underscore the importance of prior beliefs concerning systemic injustice in how people judge particular instances of discrimination, be they caused by algorithms or humans. We found that people respond differently to a more contextualized framing of algorithmic discrimination depending on their identity and suggest that this approach could lead to disagreements between groups portrayed as privileged and marginalized. Our analysis demonstrates that these disagreements might not only emerge at the level of people's perceptions of algorithms but also impact whether they think algorithms should be redesigned or deployed at all. We call for future work to bridge these gaps in perceptions of injustice to ensure that algorithms are held to a standard that does not perpetuate but instead rectify the injustices they surface.

Acknowledgments

This research was funded by the Max Planck Society, the VolkswagenStiftung through the project "Explainable Intelligent System" (AZ 98513), and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the Germany's Excellence Strategy—EXC 2092 CASA—390781972 and the project 389792660 as part of the Transregional Collaborative Research Center TRR 248 "Foundations of Perspicuous Software Systems."

References

[1] Yalcin Acikgoz, Kristl H Davison, Maira Compagnone, and Matt Laske. 2020. Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment* 28, 4 (2020), 399–416.
[2] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. 1954. *The nature of prejudice.* Addison-wesley Reading, MA.
[3] Julia Angwin, Madeleine Varner, and Ariana Tobin. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. ProPublica. https://tinyurl.com/5t3apr69.
[4] Lena Armstrong, Jayne Everson, and Amy J Ko. 2023. Navigating a black box: Students' experiences and perceptions of automated hiring. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1.* 148–158.
[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.
[6] Edmond Awad, Sohan Dsouza, Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. Crowdsourcing moral machines. *Commun. ACM* 63, 3 (2020), 48–55.
[7] Ruha Benjamin. 2023. Race after technology. In *Social Theory Re-Wired.* Routledge, 405–415.
[8] Arne Bewersdorff, Xiaoming Zhai, Jessica Roberts, and Claudia Nerdel. 2023. Myths, mis-and preconceptions of artificial intelligence: A review of the literature. *Computers and Education: Artificial Intelligence* 4 (2023), 100143.
[9] Yochanan E Bigman, Desman Wilson, Mads N Arnestad, Adam Waytz, and Kurt Gray. 2023. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General* 152, 1 (2023), 4.
[10] Yochanan E Bigman, Kai Chi Yam, Déborah Marciano, Scott J Reynolds, and Kurt Gray. 2021. Threat of racial and economic inequality increases preference for algorithm decision-making. *Computers in Human Behavior* 122 (2021), 106859.
[11] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems.* 1–14.
[12] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021).
[13] Gunn Elisabeth Birkelund, Bram Lancee, Edvard Nergård Larsen, Javier G Polavieja, Jonas Radl, and Ruta Yemane. 2022. Gender discrimination in hiring: Evidence from a cross-national harmonized field experiment. *European Sociological Review* 38, 3 (2022), 337–354.
[14] Miranda Bogen. 2019. All the Ways Hiring Algorithms Can Introduce Bias. Harvard Business Review. https://tinyurl.com/4rew6v6d.

[15] Courtney M Bonam, Vinoadharen Nair Das, Brett R Coleman, and Phia Salter. 2019. Ignoring history, denying racism: Mounting evidence for the Marley hypothesis and epistemologies of ignorance. *Social Psychological and Personality Science* 10, 2 (2019), 257–265.

[16] Andrea Bonezzi and Massimiliano Ostinelli. 2021. Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied* 27, 2 (2021), 447.

[17] Nyla R Branscombe, Michael T Schmitt, and Kristin Schiffhauer. 2007. Racial attitudes in response to thoughts of White privilege. *European Journal of Social Psychology* 37, 2 (2007), 203–215.

[18] Roy L Brooks. 1999. *When sorry isn't enough: The controversy over apologies and reparations for human injustice*. Vol. 10. nyu Press.

[19] Xanni Brown, Julian M Rucker, and Jennifer A Richeson. 2022. Political ideology moderates White Americans' reactions to racial demographic change. *Group Processes & Intergroup Relations* 25, 3 (2022), 642–660.

[20] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society* 3, 1 (2016), 2053951715622512.

[21] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making* 33, 2 (2020), 220–239.

[22] Bennett Callaghan, Leilah Harouni, Cydney H Dupree, Michael W Kraus, and Jennifer A Richeson. 2021. Testing the efficacy of three informational interventions for reducing misperceptions of the Black–White wealth gap. *Proceedings of the National Academy of Sciences* 118, 38 (2021), e2108875118.

[23] Evelyn R Carter and Mary C Murphy. 2015. Group-based differences in perceptions of racism: What counts, to whom, and why? *Social and personality psychology compass* 9, 6 (2015), 269–280.

[24] Logan S Casey, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z Strolovitch. 2017. Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *Sage Open* 7, 2 (2017), 2158244017712774.

[25] Stephen Cave and Kanta Dihal. 2019. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* 1, 2 (2019), 74–78.

[26] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. PMLR, 134–148.

[27] Jason A Colquitt. 2001. On the dimensionality of organizational justice: a construct validation of a measure. *Journal of applied psychology* 86, 3 (2001), 386.

[28] Jason A Colquitt, Donald E Conlon, Michael J Wesson, Christopher OLH Porter, and K Yee Ng. 2001. Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. *Journal of applied psychology* 86, 3 (2001), 425.

[29] Jason A Colquitt, Brent A Scott, and Jeffery A LePine. 2007. Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. *Journal of applied psychology* 92, 4 (2007), 909.

[30] Mihaela Constantinescu, Constantin Vică, Radu Uszkai, and Cristina Voinea. 2022. Blame it on the AI? on the moral responsibility of artificial moral advisors. *Philosophy & Technology* 35, 2 (2022), 35.

[31] Thomas H Costello, Gordon Pennycook, and David Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI.

[32] ACM US Public Policy Council. 2017. Statement on algorithmic transparency and accountability. *Commun. ACM* (2017).

[33] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://tinyurl.com/y64598bh.

[34] Jenny L Davis, Apryl Williams, and Michael W Yang. 2021. Algorithmic reparation. *Big Data & Society* 8, 2 (2021), 20539517211044808.

[35] Michael Ann DeVito. 2022. How transfeminine TikTok creators navigate the algorithmic trap of visibility via folk theorization. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31.

[36] Michael A DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. " Algorithms ruin everything" # RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3163–3174.

[37] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[38] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.

[39] Kari Edwards and Edward E Smith. 1996. A disconfirmation bias in the evaluation of arguments. *Journal of personality and social psychology* 71, 1 (1996), 5.

[40] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *proc. of the ACM Conference on Human Factors in Computing Systems*. 2371–2382.

[41] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *proc. of the ACM Conference on Human Factors in Computing Systems*. 153–162.

[42] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[43] Albert H Fang and Steven White. 2022. Historical information and beliefs about racial inequality. *Politics, Groups, and Identities* (2022), 1–22.

[44] Sina Fazelpour and Zachary C Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 57–63.

[45] Sina Fazelpour, Zachary C Lipton, and David Danks. 2022. Algorithmic fairness and the situated dynamics of justice. *Canadian Journal of Philosophy* 52, 1 (2022), 44–60.

[46] Tina Feldkamp, Markus Langer, Leo Wies, and Cornelius J König. 2023. Justice, trust, and moral judgements when personnel selection is supported by algorithms. *European Journal of Work and Organizational Psychology* (2023), 1–16.

[47] Leon Festinger. 1962. *A theory of cognitive dissonance*. Vol. 2. Stanford university press.

[48] Brett Q Ford, Dorainne J Green, and James J Gross. 2022. White fragility: An emotion regulation perspective. *American psychologist* 77, 4 (2022), 510.

[49] Megan French and Jeff Hancock. 2017. What's the folk theory? Reasoning about cyber-social systems. *Reasoning About Cyber-Social Systems (February 2, 2017)* (2017).

[50] Caleb Furlough, Thomas Stokes, and Douglas J Gillan. 2021. Attributing blame to robots: I. The influence of robot autonomy. *Human factors* 63, 4 (2021), 592–602.

[51] Pat Antonio Goldsmith. 2004. Schools' racial mix, students' optimism, and the Black-White and Latino-White achievement gaps. *Sociology of education* 77, 2 (2004), 121–147.

[52] Kurt Gray and Samuel Pratt. 2025. Morality in Our Mind and Across Cultures and Politics. *Annual Review of Psychology* 76 (2025).

[53] Ben Green. 2022. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology* 35, 4 (2022), 90.

[54] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 19–31.

[55] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M Redmiles. 2022. Dimensions of diversity in human perceptions of algorithmic fairness. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–12.

[56] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *proc. of the Web conference*. 903–912.

[57] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *AAAI*.

[58] Pamela Grimm. 2010. Social desirability bias. *Wiley international encyclopedia of marketing* (2010).

[59] Kobi Hackenburg and Ben M Tappin. 2024. Scaling laws for political persuasion with large language models. (2024).

[60] Dominik Hangartner, Daniel Kopp, and Michael Siegenthaler. 2021. Monitoring hiring discrimination through online recruitment platforms. *Nature* 589, 7843 (2021), 572–576.

[61] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 501–512.

[62] Jacqueline Hannan, Huei-Yen Winnie Chen, and Kenneth Joseph. 2021. Who gets what, according to whom? An analysis of fairness perceptions in service allocation. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 555–565.

[63] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS*.

[64] Eddie Harmon-Jones and Judson Mills. 2019. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In *Cognitive Dissonance, Second Edition: Reexamining a Pivotal Theory in Psychology*. American Psychological Association.

[65] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.

[66] Andrew F Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

[67] Melissa Heikkilä. 2022. Dutch scandal serves as a warning for Europe over risks of using algorithms. Politico. https://tinyurl.com/yaye24zc.

[68] Alex Hern. 2020. Twitter apologises for 'racist' image-cropping algorithm. The Guardian. https://tinyurl.com/yh75yrym.

[69] Sun-ha Hong. 2023. Prediction as extraction of discretion. *Big Data & Society* 10, 1 (2023), 20539517231171053.

[70] Lily Hu. 2023. What is "Race" in Algorithmic Discrimination on the Basis of Race? *Journal of Moral Philosophy* 1, aop (2023), 1–26.

[71] IFOW. 2022. All the Ways Hiring Algorithms Can Introduce Bias. Institute for the Future of Work. https://tinyurl.com/4xwn6p79.

[72] Daniel R Ilgen, Cynthia D Fisher, and M Susan Taylor. 1979. Consequences of individual feedback on behavior in organizations. *Journal of applied psychology* 64, 4 (1979), 349.

[73] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[74] Nari Johnson, Sanika Moharana, Christina Harrington, Nazanin Andalibi, Hoda Heidari, and Motahhare Eslami. 2024. The Fall of an Algorithm: Characterizing the Dynamics Toward Abandonment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 337–358.

[75] John T Jost, Mahzarin R Banaji, and Brian A Nosek. 2004. A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political psychology* 25, 6 (2004), 881–919.

[76] John T Jost and Orsolya Hunyady. 2005. Antecedents and consequences of system-justifying ideologies. *Current directions in psychological science* 14, 5 (2005), 260–265.

[77] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. 2015. {"My} data just goes {Everywhere:"} user mental models of the internet and implications for privacy and security. In *Eleventh symposium on usable privacy and security (SOUPS 2015)*. 39–52.

[78] Atoosa Kasirzadeh. 2022. Algorithmic fairness and structural injustice: Insights from feminist political philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 349–356.

[79] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining interpretability and explainability using sensemaking theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 702–714.

[80] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[81] Bran Knowles, Jasmine Fledderjohann, John T Richards, and Kush R Varshney. 2023. Trustworthy AI and the Logics of Intersectional Resistance. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 172–182.

[82] Eric D Knowles, Brian S Lowery, Rosalind M Chow, and Miguel M Unzueta. 2014. Deny, distance, or dismantle? How white Americans manage a privileged identity. *Perspectives on Psychological Science* 9, 6 (2014), 594–609.

[83] Michael W Kraus, Sa-kiera TJ Hudson, and Jennifer A Richeson. 2022. Framing, context, and the misperception of Black–White wealth inequality. *Social Psychological and Personality Science* 13, 1 (2022), 4–13.

[84] Michael W Kraus, Julian M Rucker, and Jennifer A Richeson. 2017. Americans misperceive racial economic equality. *Proceedings of the National Academy of Sciences* 114, 39 (2017), 10324–10331.

[85] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[86] Entung Enya Kuo, Michael W Kraus, and Jennifer A Richeson. 2020. High-status exemplars and the misperception of the Asian-White wealth gap. *Social Psychological and Personality Science* 11, 3 (2020), 397–405.

[87] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[88] Kevin LaGrandeur. 2023. The consequences of AI hype. *AI and Ethics* (2023), 1–4.

[89] Michelle S Lam, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. 2022. End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. *proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–34.

[90] Michelle S Lam, Ayush Pandit, Colin H Kalicki, Rachit Gupta, Poonam Sahoo, and Danaë Metaxa. 2023. Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–37.

[91] Markus Langer, Kevin Baum, and Nadine Schlicker. 2023. A signal detection perspective on error and unfairness detection as a critical aspect of human oversight of AI-based systems. (2023).

[92] Markus Langer, Cornelius J König, Caroline Back, and Victoria Hemsing. 2023. Trust in Artificial Intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology* 38, 3 (2023), 493–508.

[93] Markus Langer, Cornelius J König, and Maria Papathanasiou. 2019. Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment* 27, 3 (2019), 217–234.

[94] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.

[95] Minha Lee, Peter Ruijten, Lily Frank, Yvonne de Kort, and Wijnand IJsselsteijn. 2021. People May Punish, But Not Blame Robots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[96] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[97] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[98] Gabriel Lima, Meeyoung Cha, Chihyung Jeon, and Kyung Sin Park. 2021. The Conflict Between People's Urge to Punish AI and Legal Systems. *Frontiers in Robotics and AI* 8 (2021), 339. https://doi.org/10.3389/frobt.2021.756242

[99] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In *proc. of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[100] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2023. Blaming humans and machines: What shapes people's reactions to algorithmic harm. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–26.

[101] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *proc. of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

[102] Ting-An Lin and Po-Hsuan Cameron Chen. 2022. Artificial Intelligence in a Structurally Unjust Society. *Feminist Philosophy Quarterly* 8, 3/4 (2022).

[103] Manuel London. 2003. *Job feedback: Giving, seeking, and using feedback for performance improvement*. Psychology Press.

[104] Charles G Lord, Lee Ross, and Mark R Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology* 37, 11 (1979), 2098.

[105] Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. Implications of AI (un-) fairness in higher education admissions: the effects of perceived AI (un-) fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 122–130.

[106] Alva Markelius, Connor Wright, Joahna Kuiper, Natalie Delille, and Yu-Ting Kuo. 2024. The mechanisms of AI hype and its planetary and social costs. *AI and Ethics* (2024), 1–16.

[107] Marijn Martens, Ralf De Wolf, Bettina Berendt, and Lieven De Marez. 2023. Decoding algorithms: Exploring end-users' mental models of the inner workings of algorithmic news recommenders. *Digital Journalism* 11, 1 (2023), 203–225.

[108] Samuel Mayworm, Michael Ann DeVito, Daniel Delmonaco, Hibby Thach, and Oliver L Haimson. 2024. Content moderation folk theories and perceptions of platform spirit among marginalized social media users. *ACM Transactions on Social Computing* 7, 1 (2024), 1–27.

[109] John B McConahay. 1983. Modern racism and modern discrimination: The effects of race, racial attitudes, and context on simulated hiring decisions. *Personality and Social Psychology Bulletin* 9, 4 (1983), 551–558.

[110] Sabelo Mhlambi and Simona Tiribelli. 2023. Decolonizing AI ethics: Relational autonomy as a means to counter AI Harms. *Topoi* (2023), 1–14.

[111] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33 (2020), 659–684.

[112] Luis Morales-Navarro, Yasmin Kafai, Vedya Konda, and Danaë Metaxa. 2024. Youth as Peer Auditors: Engaging Teenagers with Algorithm Auditing of Machine Learning Applications. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. 560–573.

[113] Satya Nadella. 2016. The Partnership of the Future. Slate. https://tinyurl.com/yhm2ah6c.

[114] Jessica C Nelson, Glenn Adams, and Phia S Salter. 2013. The Marley hypothesis: Denial of racism reflects ignorance of history. *Psychological science* 24, 2 (2013), 213–218.

[115] David T Newman, Nathanael J Fast, and Derek J Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160 (2020), 149–167.

[116] Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. 2020. Exploring mental models for transparent and controllable recommender systems: a qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 183–191.

[117] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations.

Science 366, 6464 (2019), 447–453.
[118] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.
[119] Florian Pethig and Julia Kroenung. 2023. Biased humans, (un)biased algorithms? *Journal of Business Ethics* 183, 3 (2023), 637–652.
[120] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. 2017. Exploring user perceptions of discrimination in online targeted advertising. In *proc. of the USENIX Security Symposium*. 935–951.
[121] Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H Midtbøen. 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences* 114, 41 (2017), 10870–10875.
[122] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and information technology* 20, 1 (2018), 5–14.
[123] BIES RJ. 1986. Interactional justice: Communication criteria of fairness. *Research on negotiation in organizations* 1 (1986), 43–55.
[124] Wendy D Roth, Elena G van Stee, and Alejandra Regla-Vargas. 2023. Conceptualizations of race: Essentialism and constructivism. *Annual Review of Sociology* 49 (2023), 39–58.
[125] Julian Rucker, Ajua Duker, and Jennifer Richeson. [n. d.]. Structurally unjust: How lay beliefs about racism relate to perceptions of and responses to racial inequality in criminal justice. ([n. d.]).
[126] Julian M Rucker and Jennifer A Richeson. 2021. Beliefs about the interpersonal vs. structural nature of racism and responses to racial inequality. In *The Routledge international handbook of discrimination, prejudice and stereotyping*. Routledge, 13–25.
[127] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. On the conversational persuasiveness of large language models: A randomized controlled trial. *arXiv preprint arXiv:2403.14380* (2024).
[128] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.
[129] Nadine Schlicker, Markus Langer, Sonja K Ötting, Kevin Baum, Cornelius J König, and Dieter Wallach. 2021. What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior* 122 (2021), 106837.
[130] Daniel B Shank and Alyssa DeSanti. 2018. Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in human behavior* 86 (2018), 401–411.
[131] Donghee Shin and Yong Jin Park. 2019. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98 (2019), 277–284.
[132] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology* 24, 1 (2022), 2.
[133] Jessie Smith, Nasim Sonboli, Casey Fiesler, and Robin Burke. 2020. Exploring user opinions of fairness in recommender systems. *arXiv preprint arXiv:2003.06461* (2020).
[134] Wonyoung So, Pranay Lohia, Rakesh Pimplikar, AE Hosoi, and Catherine D'Ignazio. 2022. Beyond Fairness: Reparative Algorithms to Address Historical Injustices of Housing Discrimination in the US. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 988–1004.
[135] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2459–2468.
[136] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (2022), 20539517221115189.
[137] Lisa A Steelman and Kelly A Rutkowski. 2004. Moderators of employee reactions to negative feedback. *Journal of Managerial Psychology* 19, 1 (2004), 6–18.
[138] Michael T Stuart and Markus Kneer. 2021. Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.
[139] Anna-Lena Theus. 2023. Striving for Affirmative Algorithmic Futures: How the Social Sciences can Promote more Equitable and Just Algorithmic System Design. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 558–568.
[140] Meinald T Thielsch, Sarah M Meeßen, and Guido Hertel. 2018. Trust and distrust in information systems at the workplace. *PeerJ* 6 (2018), e5483.
[141] Michael Tonry. 1998. A comparative perspective on minority groups, crime, and criminal justice. *Eur. J. Crime Crim. L. & Crim. Just.* 6 (1998), 60.
[142] Ana Valdivia, Júlia Corbera Serrajòrdia, and Aneta Swianiewicz. 2022. There is an elephant in the room: Towards a critique on the use of fairness in biometrics. *AI and Ethics* (2022), 1–16.
[143] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.
[144] Heidi A Vuletich, Kurt Gray, and B Keith Payne. 2023. People's Preferences for Inequality Respond Instantly to Changes in Status: A Simulated Society Experiment of Conflict Between the Rich and the Poor. *Cognitive science* 47, 6 (2023), e13306.
[145] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
[146] Julie Weed. 2021. Résumé-Writing Tips to Help You Get Past the A.I. Gatekeepers. New York Times. https://tinyurl.com/yc2hz9tp.
[147] Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. 2024. The Neutrality Fallacy: When Algorithmic Fairness Interventions are (Not) Positive Action. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2060–2070.
[148] Doris Weichselbaumer. 2003. Sexual orientation discrimination in hiring. *Labour economics* 10, 6 (2003), 629–642.
[149] Lindsay Weinberg. 2022. Rethinking fairness: an interdisciplinary survey of critiques of hegemonic ML fairness approaches. *Journal of Artificial Intelligence Research* 74 (2022), 75–109.
[150] Pak-Hang Wong. 2020. Democratizing algorithmic fairness. *Philosophy & Technology* 33 (2020), 225–244.
[151] Brita Ytre-Arne and Hallvard Moe. 2021. Folk theories of algorithms: Understanding digital irritation. *Media, Culture & Society* 43, 5 (2021), 807–824.
[152] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International World Wide Web Conference*. 1171–1180.
[153] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. 962–970.
[154] Ethan Zell and Tara L Lesick. 2022. Ignorance of history and political differences in perception of racism in the United States. *Social Psychological and Personality Science* 13, 6 (2022), 1022–1031.
[155] Marilyn Zhang. 2022. Affirmative Algorithms: Relational Equality as Algorithmic Fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 495–507.
[156] Annette Zimmermann and Chad Lee-Stronach. 2022. Proceed with caution. *Canadian Journal of Philosophy* 52, 1 (2022), 6–25.

A Supplementary Methods

A.1 Vignette

Participants assigned to the condition in which we provided information about racial injustice (*context = Historical Injustice*) read the following two paragraphs:

> *Racial discrimination has been a feature of American society since its inception. One setting in which racial disparities are still evident to this day involves hiring decisions. Historically, there has been a close connection between how companies make hiring decisions and race.*
>
> *Companies have historically imposed different criteria on job applicants based on their race, refusing to hire Black job applicants while prioritizing their White counterparts. These long-standing discriminatory practices have had profound and lasting effects on society. Although some efforts to confront and rectify these disparities exist, Black people still face difficulties while White people have advantages being hired to this day.*

All participants read the following baseline vignette, in which a company screens job applicants with the assistance of a discriminatory algorithm:

> *A company employs an algorithm to help them screen job applicants. This algorithm determines a score for each job applicant based on their application materials. These scores correspond to an evaluation of the applicant in relation to the job position, with higher scores indicating that the applicant is a better candidate for that particular position. The company then uses these scores to screen job applicants and decide who should be offered an interview.*
>
> *After some time, an investigation of the algorithm's scores found that Black applicants are assigned lower scores than White applicants with similar qualifications. That is, the algorithm is suggesting that Black applicants are worse candidates and thus should be interviewed less frequently than similar White applicants.*

The baseline vignette was modified according to the *explanation* treatment condition to which participants were assigned. Participants in the *explanation = Learn* condition were shown a paragraph explaining that the algorithm learns to score applicants from past human decisions, which was placed in between the two paragraphs of the baseline vignette:

> *The algorithm works by analyzing data from past human decisions. By analyzing historical data from the company's past screening decisions, the algorithm is able to learn the company's past screening criteria to score new applicants. In short, the algorithm learns how to score job applicants from past human decisions.*

Participants assigned to the *explanation = Perpetuate* also read a paragraph stating that algorithms can perpetuate human biases:

> *The algorithm's scores are not guaranteed to be accurate or fair. Since the algorithm is trained on past screening decisions, its scores will be largely consistent with these past decisions. That is, if the past decisions the algorithm was trained on exhibit certain biases, the algorithm will perpetuate the same biases in its scores.*

A.2 Measures

In addition to the description of our methods in Section 3, we now present all questions included in our study. In total, participants answered four sets of questions:

(1) **Exploratory Variables:** These questions refer to several factors that could help us explain our experimental effects.
(2) **Fairness, Trust, and Blame:** We collected participants' judgments of fairness and reported trust in the algorithm. We also asked participants to indicate how much blame some entities deserved for the harm described in the vignette.
(3) **Downstream Effects:** We asked participants the extent to which they believed the algorithm should be banned or redesigned.
(4) **Background Questions:** We also captured participants' beliefs in racial injustice and whether they think algorithms can reduce human biases. The study also asked participants to self-report their racial identity and political leaning, as well as answer some open-ended questions.

A.2.1 Exploratory Variables. After reading the vignette, participants were asked several exploratory questions. These questions referred to some factors that may come into play in determining people's judgments of fairness and reported trust in the algorithm. The study employed the following questions:

- **Perceived Objectivity:** Participants indicated the extent to which they agreed with four statements affirming that the algorithm's scores were objective (adapted from Pethig and Kroenung [119]; -3 = Strongly disagree, 3 = Strongly agree).
  (1) I believe scores determined by the algorithm are reasonable and logical.
  (2) I believe scores determined by the algorithm objectively consider all of the facts.
  (3) I believe scores determined by the algorithm are based on logical analysis.
  (4) I believe scores determined by the algorithm are rational and objective.
- **Potential To Reduce Biases:** "To what extent do you believe the algorithm can reduce human biases in screening decisions?" (1 = Not at all, 7 = Definitely).
- **Intentionality:** "To what extent do you believe the algorithm determines scores intentionally?" (1 = Not intentionally at all, 7 = Extremely intentionally).
- **Autonomy:** "To what extent do you believe the algorithm determines scores without human intervention?" (1 = Not at all, 7 = Definitely).
- **Similarity:** "To what extent are the algorithm's scores similar to the scores that a typical human resources (HR) manager would determine?" (1 = Not similar at all, 7 = Definitely similar).

All questions within this group were presented on the same page and in random order. As mentioned in Section 3, we asked these questions before our main dependent variables because we

operationalized them as mediators that could help us explain any potential experimental effects.

We hypothesized that our context manipulation would decrease the perceived objectivity of the algorithm, particularly when the vignette explained that the algorithm perpetuates past human biases. Our expectation was that this decrease in objectivity would decrease the perceived fairness of and trust in the algorithm [119]. As suggested by prior work showing that people believe algorithms can be less biased than humans [9, 10], we also expected that telling participants that algorithms learn from past human decisions would decrease the extent to which they believe algorithms can reduce human biases, thereby decreasing perceived fairness and trust. As for our question on perceived similarity, we hypothesized that our explanation manipulations—which explicitly mention that algorithms learn from past human decisions—would increase the perceived similarity between scores determined by humans and algorithms. The questions related to intentionality and autonomy were added to complement our blame measures given prior work showing the correlation between blame and these two measures [50, 98, 130, 138]. We present an analysis of the measures relevant to fairness and trust in Table 7 and Appendix C below.

A.2.2 Fairness, Trust, and Blame. Participants then answered questions addressing fairness, trust, and blame. In addition to the questions we report in the main text, participants attributed blame to "the algorithm," "the developers of the algorithm," and "the company employing the algorithm:" "How much blame do the following entities deserve for the scores that the algorithm determines?" (1 = No blame at all, 7 = Extremely blame).

As explained in the main text, we included these questions in the study due to prior work examining blame judgments resulting from algorithmic harm. [30, 95, 99]. Because blame judgments refer to people's reactive attitudes following instances of harm—and not necessarily to their perceptions of the algorithm and its scores, as fairness and trust—we omit our analysis from the main text and report them below. Nonetheless, all of our data is available for further analysis at https://tinyurl.com/AIFairPerceptions-Injustice.

Questions concerning fairness and trust were shown in random order and on the same page. Questions addressing blame were shown on a different page, with the entities presented in random order. The presentation order of these two pages was randomized.

A.2.3 Downstream Effects. We next asked participants their opinions concerning the deployment of the algorithm and whether it should be changed, as reported in Section 3.

A.2.4 Background Questions. In addition to the questions capturing participants' beliefs in racial injustice presented in Section 3, we also asked participants questions directly related to our explanation manipulation:

(1) **Belief in Racial Injustice:** Participants indicated the extent to which they agreed with three statements affirming that there exists racial discrimination in hiring decisions (-3 = Strongly disagree, 3 = Strongly agree).
   (i) Racial disparities in hiring decisions are a long-standing problem in American society.
   (ii) Hiring decisions are marked by racial disparities to this day.
   (iii) Discrimination across racial lines in hiring decisions has had lasting effects on society.
(2) **Inequality of Opportunity:** "Do you think that White Americans have more opportunities than they should, that Black Americans have more opportunities than they should, or that opportunities are about equal between racial groups?" (1 = Black Americans have too much, 4 = Things are about equal, and 7 = White Americans have too much; from Callaghan et al. [22]).
(3) **Algorithms Perpetuate Biases?:** Participants agreed—or disagreed—with two statements asserting that algorithms can perpetuate past human biases (-3 = Strongly disagree, 3 = Strongly agree).
   (i) If people's past screening decisions are biased, algorithms will also make biased screening decisions.
   (ii) Algorithms' screening decisions replicate people's past screening decisions.
(4) **Who is More Biased?:** "Do you think algorithms or humans are more biased when making hiring screening decisions?" (1 = Algorithms are definitely more biased, 4 = Algorithms and humans are equally biased, 7 = Humans are definitely more biased).

We motivate our measures of belief in racial injustice and inequality of opportunity in Section 3. We hypothesized that our explanation manipulations would increase the extent to which participants agreed that algorithms perpetuate biases. Similarly, we expected our manipulations to modify people's judgments concerning who is relatively more biased between humans and algorithms. We decided to omit the results of these questions from the main text because our explanation manipulation had only marginal effects on people's perceptions of algorithmic discrimination. Nonetheless, we report an analysis of these questions in Table 8 below.

Finally, participants reported whether they had "any training or work experience in professions related to machine learning (ML) or artificial intelligence (AI)." We also asked them two open-ended questions for exploratory analysis: 1) "Explain in your own words how you think algorithms that screen job applicants work. Please, enter more than 20 characters" and 2) "If an algorithm leads to discriminatory hiring decisions, what do you think is the reason for that? Please, enter more than 20 characters." The study concluded with the political leaning and race measures presented in the main text.

B Supplementary Demographic Information

To gather more demographic information from our participants, we re-invited them to complete a short study in which they were asked some demographic questions. This study was conducted a month after the completion of the main study. Out of the 716 participants who provided valid responses to the main study, 569 (79.47%) answered this follow-up study. Participants were paid 0.15 GBP (approximately 0.19 USD), resulting in a median pay of 13.17 GBP per hour (approximately 16.96 USD per hour). Table 6 presents information about participants' income and education level.

C Supplementary Analysis

We present a series of supplementary analyses:

Income Level	**N (%)**
Less than $25,000	71 (12.48%)
$25,000 - $50,000	114 (20.04%)
$50,000 - $75,000	116 (20.39%)
$75,000 - $100,000	106 (18.63%)
$100,000 - $150,000	87 (15.29%)
More than $150,000	54 (9.49%)
Prefer not to say	21 (3.69%)
Education Level	**N (%)**
High-school	73 (12.83%)
Some college but no degree	129 (22.67%)
Associate degree	46 (8.08%)
Bachelor's degree (e.g., BS, BA)	215 (37.79%)
Professional degree (e.g., MD, JD)	11 (1.93%)
Master's degree (e.g., MS, MA, MBA)	78 (13.71%)
Doctorate degree (e.g., PhD, EdD)	11 (1.93%)
Prefer not to say	6 (1.05%)

Table 6: Additional demographic information from participants who completed the follow-up demographic survey.

- Table 7 reports an analysis of our exploratory measures depending on experimental conditions and participants' racial group. For each measure, we first look for the hypothesized effects of our experimental manipulations and then explore any heterogeneous effects across racial groups. Please refer to Appendix A for a more detailed rationale for the inclusion of each measure.
  - Perceived Objectivity: We hypothesized that contextualizing algorithms in systemic injustice would decrease the perceived objectivity of the algorithm, particularly when the vignette was explicit in how algorithms perpetuate biases (i.e., there would be a significant interaction between our manipulations). However, Model (1) in Table 7 shows no effect of our context and explanation manipulations. When we account for participants' racial group, we observed a similar moderation effect to what we found for our dependent variables: participants from the Advantaged group judged the algorithm to be less objective when contextualized in systemic injustice (Model (2) in Table 7). All in all, our analysis is aligned with our main findings.
  - Potential to Reduce Biases: We expected our Perpetuate manipulation to decrease people's perceptions that algorithms can reduce human biases given that it explicitly states that the algorithm can perpetuate human biases. Model (3) in Table 7 shows that explicitly telling people that algorithms perpetuate human biases decreases the perceived potential of algorithms to reduce them. We also found a moderation effect of participants' racial group (see Model (4) Table 7), such that those from the Advantaged group became less convinced that algorithms can reduce biases. This moderation effect is also aligned with our main findings.
  - Similarity: We hypothesized that our explanation manipulation would increase the perceived similarity between human and algorithmic decisions. Model (5) in Table 7 shows that our manipulations increased the perceived similarity between humans and algorithms. In contrast, we did not find any moderation with participants' racial group (see Model (6) in Table 7).
- Table 8 presents how our explanation manipulations impacted people's belief that algorithms perpetuate human biases.
- Table 9 shows the effect of our explanation manipulations and its interaction with participants' racial group on perceived fairness and reported trust.
- Tables 10, 11, 12 replicates our main findings using ordinal regressions as robustness checks.
- Table 13 presents the direct and indirect effects of our moderated mediation model.
- Tables 14 and 15 present how our experimental manipulations and participants' racial group impact people's belief that the algorithm should be changed or not used at all.
- Table 16 presents the results of regressions of participants' blame judgments to our experimental manipulations. We observe a shift of blame from developers to the user when participants were told that algorithms learn from past human biases.

We also replicate our main findings using different ways of grouping participants concerning their racial identity. Our results are consistent with those reported in the main text. More specifically, the coefficients referring to the heterogeneous effects of our context manipulation remain significant. Furthermore, we observe visually that participants who self-identified as Asian and Black become more positive about the vignette while White participants become more negative.

In this analysis, we use participants who self-identified as White (i.e., the main text's Advantaged group) as the baseline for comparison with our results in the main text. By doing this, we show that the heterogeneous effects reported in the main text are replicated even if we further divide the Disadvantaged group into more fine-grained groups.

- Table 17 replicates our main findings using a more fine-grained categorization of participants' self-reported race (see Figure 5). We grouped participants into four groups: White (N = 372), Black (N = 127), Asian (N = 123), Mixed/Other (N = 94). We group multi-racial participants with other minority groups because of their small sample size.
- Table 18 replicates our main findings using Prolific's data on ethnicity (see Figure 6). Prolific workers reported their ethnicity out of the following five options: Asian (N = 122), Black (N = 130), Mixed (N = 63), Other (N = 42), and White (N = 359).

D Self-Reported vs. Real Political Orientation

An unexpected result from our research is that participants from the Advantaged group who read additional information about racial injustice reported being more liberal than those who did not. We did not expect participants' political orientation to change depending on the experimental manipulation as it refers to a core set of beliefs held by individuals. We acknowledge two potential explanations: 1) it is possible that participants from the Advantaged group report

	Dependent variables					
	Perceived Objectivity		Potential to Reduce Biases		Similarity	
	(1)	(2)	(3)	(4)	(5)	(6)
context = Historical Injustice	0.111	0.173				
	(0.209)	(0.171)				
explanation = Learn	0.140		−0.161	0.078	0.451**	0.396*
	(0.206)		(0.147)	(0.214)	(0.137)	(0.201)
explanation = Perpetuate	0.180		−0.578***	−0.201	0.500***	0.484*
	(0.211)		(0.147)	(0.208)	(0.137)	(0.195)
(context = Historical Injustice):(explanation = Learn)	−0.332					
	(0.294)					
(context = Historical Injustice):(explanation = Perpetuate)	−0.273					
	(0.293)					
group = Advantaged		0.211		0.171		−0.177
		(0.169)		(0.209)		(0.195)
(context = Historical Injustice):(group = Advantaged)		−0.507*				
		(0.238)				
(explanation = Learn):(group = Advantaged)				−0.446		0.111
				(0.294)		(0.275)
(explanation = Perpetuate):(group = Advantaged)				−0.755*		0.023
				(0.292)		(0.273)
Constant	−0.827***	−0.829***	3.240***	3.152***	3.953***	4.045***
	(0.150)	(0.122)	(0.105)	(0.151)	(0.098)	(0.141)
Observations	716	716	716	716	716	716

Table 7: Linear regressions of our exploratory variables. Dependent variables: perceived objectivity; the algorithm's potential to reduce human biases; perceived similarity between human and algorithmic scores. Independent variables: dummy variables (context and explanation) indicating to which treatment condition participants' were assigned and participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001

	Dependent Variables			
	Algorithms Perpetuate Biases?		Who is More Biased?	
	(1)	(2)	(3)	(4)
explanation = Learn	0.595***	0.527**	0.225	0.237
	(0.111)	(0.162)	(0.130)	(0.190)
explanation = Perpetuate	0.764***	0.635***	−0.193	0.017
	(0.111)	(0.158)	(0.129)	(0.184)
group = Advantaged		−0.118		0.054
		(0.158)		(0.185)
(explanation = Learn):(group = Advantaged)		0.130		−0.025
		(0.223)		(0.260)
(explanation = Perpetuate):(group = Advantaged)		0.254		−0.422
		(0.221)		(0.258)
Constant	1.247***	1.308***	5.296***	5.268***
	(0.079)	(0.114)	(0.092)	(0.133)
Observations	716	716	716	716

Table 8: Linear regressions of measures directly related to our explanation manipulation. Dependent variables: participants' agreement that algorithms perpetuate past human biases and their views concerning who is more biased between algorithms and humans. Independent variables: dummy variables (explanation) indicating to which treatment condition participants were assigned. Some models also account for participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

	Dependent Variables			
	Distributional Fairness	Procedural Fairness	Interpersonal Fairness	Trust
	(1)	(2)	(3)	(4)
explanation = Learn	0.034	0.264	−0.020	0.166
	(0.203)	(0.207)	(0.185)	(0.203)
explanation = Perpetuate	0.160	0.125	0.135	0.087
	(0.197)	(0.201)	(0.180)	(0.197)
group = Advantaged	0.193	0.207	−0.177	0.341
	(0.198)	(0.202)	(0.180)	(0.198)
(explanation = Learn):(group = Advantaged)	−0.295	−0.296	0.044	−0.423
	(0.278)	(0.284)	(0.253)	(0.279)
(explanation = Perpetuate):(group = Advantaged)	−0.562*	−0.402	−0.293	−0.692*
	(0.277)	(0.282)	(0.252)	(0.277)
Constant	−1.429***	−1.397***	−1.237***	−1.469***
	(0.143)	(0.145)	(0.130)	(0.143)
Observations	716	716	716	716

Table 9: Linear regressions of perceived fairness and trust. Dependent variables: perceived distributional fairness, procedural fairness, interpersonal fairness, and trust in the algorithm. Independent variables: dummy variables (explanation) indicating to which treatment condition participants' were assigned and participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

	Dependent variable:			
	Distributional Fairness	Procedural Fairness	Interpersonal Fairness	Trust
	(1)	(2)	(3)	(4)
context = Historical Injustice	−0.146	−0.099	−0.073	−0.060
	(0.235)	(0.232)	(0.229)	(0.233)
explanation = Learn	−0.187	0.098	−0.134	−0.050
	(0.229)	(0.229)	(0.229)	(0.228)
explanation = Perpetuate	−0.105	−0.133	0.025	−0.187
	(0.230)	(0.229)	(0.229)	(0.231)
(context = Historical Injustice):(explanation = Learn)	0.148	−0.009	0.238	−0.059
	(0.328)	(0.327)	(0.324)	(0.327)
(context = Historical Injustice):(explanation = Perpetuate)	0.082	0.153	−0.016	−0.092
	(0.324)	(0.322)	(0.319)	(0.322)
Observations	716	716	716	716

Table 10: Ordinal regressions of perceived fairness and trust. Dependent variables: perceived distributional fairness, procedural fairness, interpersonal fairness, and trust in the algorithm. Independent variables: dummy variables (context and explanation) indicating to which treatment condition participants' were assigned. Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

being more liberal to portray themselves more favorably in the context of the study, or 2) it is caused by random assignment biases.

Prior work discusses some defensive behaviors that racially privileged people may display when provided with information that may threaten their own image or social position [2, 17, 47, 48, 64, 76, 82]. Considering that our context manipulation portrays members of the Advantaged group as having unearned benefits to the detriment of others, participants from this group could have self-reported their political orientation in a way that they can be perceived more favorably in this particular context. This interpretation is aligned with participants' tendency to become more negative towards algorithmic discrimination. Hence, it is possible that participants from the Advantaged group report being more liberal and thus more critical of algorithmic racial discrimination to present themselves more favorably in the context of the experiment.[3]

[3]It is also possible that participants hid their own racial identity due to the way that our study portrayed specific racial groups. We replicated our analysis using participants' self-reported ethnicity provided by Prolific and found consistent results (see Tables 17-18). Considering that participants answered these questions outside the scope of this study, we consider that participants' self-reported racial identity is not likely to follow the same trend as their self-reported political orientation. Furthermore, we also find

	Dependent variable:			
	Distributional Fairness	Procedural Fairness	Interpersonal Fairness	Trust
	(1)	(2)	(3)	(4)
context = Historical Injustice	0.279	0.342	0.337	0.137
	(0.193)	(0.193)	(0.191)	(0.194)
group = Advantaged	0.248	0.361	−0.002	0.273
	(0.185)	(0.186)	(0.187)	(0.186)
(context = Historical Injustice):(group = Advantaged)	−0.645*	−0.744**	−0.624*	−0.470
	(0.265)	(0.265)	(0.263)	(0.264)
Observations	716	716	716	716

Table 11: Ordinal regressions of perceived fairness and trust. Dependent variables: perceived distributional fairness, procedural fairness, interpersonal fairness, and trust in the algorithm. Independent variables: dummy variables (context) indicating to which treatment condition participants' were assigned and participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

	Dependent variable:			
	Distributional Fairness	Procedural Fairness	Interpersonal Fairness	Trust
	(1)	(2)	(3)	(4)
Explanation = Learn	0.069	0.288	−0.031	0.207
	(0.244)	(0.242)	(0.240)	(0.242)
explanation = Perpetuate	0.257	0.179	0.201	0.144
	(0.232)	(0.230)	(0.229)	(0.235)
group = Advantaged	0.271	0.263	−0.197	0.461*
	(0.235)	(0.232)	(0.229)	(0.234)
(Explanation = Learn):(group = Advantaged)	−0.341	−0.356	0.037	−0.526
	(0.330)	(0.328)	(0.326)	(0.329)
(explanation = Perpetuate):(group = Advantaged)	−0.631	−0.457	−0.374	−0.715*
	(0.325)	(0.322)	(0.320)	(0.323)
Observations	716	716	716	716

Table 12: Ordinal regressions of perceived fairness and trust. Dependent variables: perceived distributional fairness, procedural fairness, interpersonal fairness, and trust in the algorithm. Independent variables: dummy variables (explanation) indicating to which treatment condition participants' were assigned and participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

Because we conducted an additional demographic study (see Appendix B) a month later, we can analyze whether participants report different political orientations in these two time points. If participants report being more conservative in the demographic study—which did not contain any mention of racial injustice—it could indicate that participants' reported political leaning from the main study was influenced by our experimental manipulations.

Out of the participants that completed the additional demographic study, only 91 (16.31%) of participants did not answer the two questions in exactly the same way. Among these, we find that participants reported being marginally more conservative (M=−0.253, SD=1.16 on a 5-point scale) in main study. If we zoom in on the 28 participants from the Advantaged group that were shown the context manipulation *and* did not report the exact same political learning, we find that they also report being more conservative in the main study (M=−0.286, SD=0.976). All in all, our results do not suggest that participants reported being more liberal in the main study due to our context manipulation.

Finally, we do not expect our findings to be tainted by random assignment biases across racial groups since we also control for participants' self-reported racial identity in our analysis. Nonetheless, we call for future work to replicate our results by setting quotas across not only race but also political leaning.

that participants reported their racial identity consistently across the main study and the additional demographic study.

Path	Direct/Indirect Effects
Racial Group → Distributive Fairness (for context = No)	$c = -0.065$, $SE = 0.144$
Racial Group → Distributive Fairness (for context = Historical Injustice)	$c = -0.575$, $SE = 0.142^{\dagger}$
Racial Group → Belief in Racial Injustice → Distributive Fairness	$ab = 0.080$, $SE = 0.038^{\dagger}$
Racial Group → Inequality of Opportunity → Distributive Fairness	$ab = 0.140$, $SE = 0.046^{\dagger}$
Racial Group → Political Orientation → Distributive Fairness	$ab = 0.006$, $SE = 0.009$
Racial Group → Procedural Fairness (for context = No)	$c = 0.059$, $SE = 0.148$
Racial Group → Procedural Fairness (for context = Historical Injustice)	$c = -0.534$, $SE = 0.147^{\dagger}$
Racial Group → Belief in Racial Injustice → Procedural Fairness	$ab = 0.078$, $SE = 0.037^{\dagger}$
Racial Group → Inequality of Opportunity → Procedural Fairness	$ab = 0.136$, $SE = 0.046^{\dagger}$
Racial Group → Political Orientation → Procedural Fairness	$ab = 0.003$, $SE = 0.008$
Racial Group → Interpersonal Fairness (for context = No)	$c = -0.151$, $SE = 0.139$
Racial Group → Interpersonal Fairness (for context = Historical Injustice)	$c = -0.585$, $SE = 0.138^{\dagger}$
Racial Group → Belief in Racial Injustice → Interpersonal Fairness	$ab = 0.068$, $SE = 0.033^{\dagger}$
Racial Group → Inequality of Opportunity → Interpersonal Fairness	$ab = 0.034$, $SE = 0.037$
Racial Group → Political Orientation → Interpersonal Fairness	$ab = 0.006$, $SE = 0.009$
Racial Group → Trust (for context = No)	$c = -0.076$, $SE = 0.149$
Racial Group → Trust (for context = Historical Injustice)	$c = -0.416$, $SE = 0.148^{\dagger}$
Racial Group → Belief in Racial Injustice → Trust	$ab = 0.062$, $SE = 0.032^{\dagger}$
Racial Group → Inequality of Opportunity → Trust	$ab = 0.150$, $SE = 0.049^{\dagger}$
Racial Group → Political Orientation → Trust	$ab = 0.006$, $SE = 0.010$

Table 13: Direct and indirect effects of the mediation model depicted in Figure 4. The model explores how participants' belief in racial injustice, perceptions of inequality of opportunity across racial lines, and political orientation help explain differences in perceived fairness of and trust in the algorithm. † indicates that the bootstrapped 95% confidence interval does not include zero.

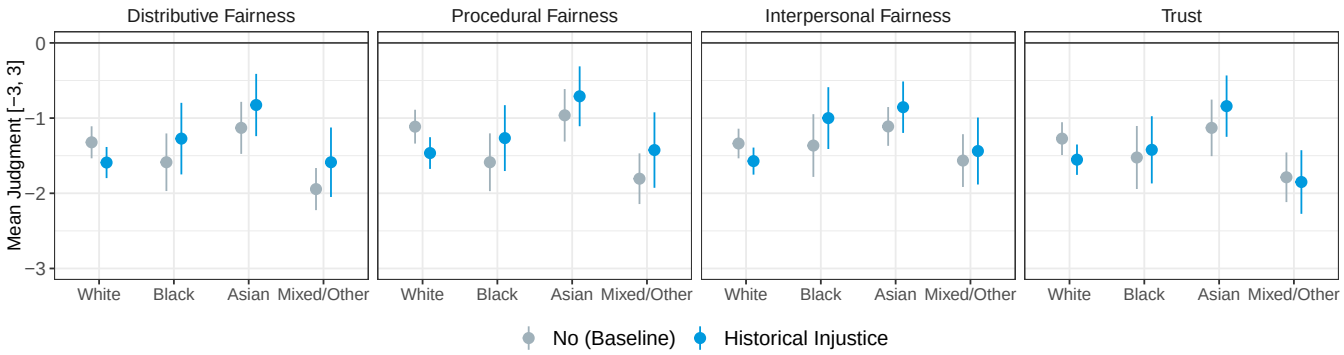


Figure 5: Perceived fairness of and trust in the algorithm depending on the treatment condition and the participants' self-reported racial group. Participants either did not receive any information about systemic injustice in the hiring domain (Context = *No (Baseline)*) or read two paragraphs explaining how Black job applicants have been (and continue to be) systematically disadvantaged in hiring decisions (Context = Historical Injustice).

	Dependent Variable		
	Should It Be Changed?		
	(1)	(2)	(3)
context = Historical Injustice	0.043	−0.145	
	(0.170)	(0.140)	
explanation = Learn	0.129		0.017
	(0.168)		(0.175)
explanation = Perpetuate	0.099		−0.052
	(0.172)		(0.169)
(context = Historical Injustice):(explanation = Learn)	0.041		
	(0.239)		
(context = Historical Injustice):(explanation = Perpetuate)	0.065		
	(0.239)		
(group = Advantaged):(explanation = Learn)			0.245
			(0.239)
(group = Advantaged):(explanation = Perpetuate)			0.367
			(0.238)
group = Advantaged		−0.184	−0.178
		(0.137)	(0.170)
(context = Historical Injustice):(group = Advantaged)		0.425*	
		(0.194)	
Constant	1.832***	2.006***	1.946***
	(0.122)	(0.099)	(0.123)
Observations	716	716	716

Table 14: Linear regressions of participants' belief that the algorithm should redesigned. Dependent variables: participants agreement with a statement affirming that the algorithm should be changed. Independent variables: dummy variables (context and explanation) indicating to which treatment condition participants' were assigned and participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

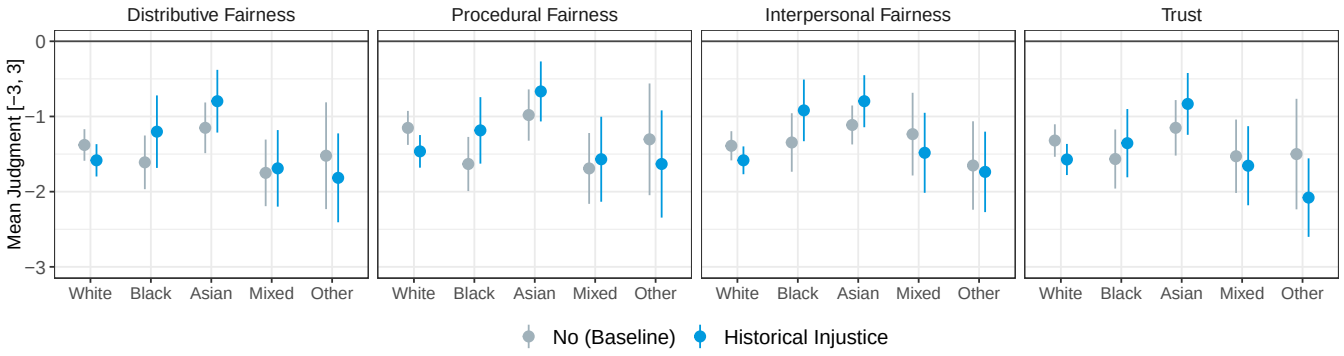


Figure 6: Perceived fairness of and trust in the algorithm depending on the treatment condition and the participants' ethnicity as reported by Prolific. Participants either did not receive any information about systemic injustice in the hiring domain (Context = *No (Baseline)*) or read two paragraphs explaining how Black job applicants have been (and continue to be) systematically disadvantaged in hiring decisions (Context = Historical Injustice).

	Dependent Variable		
	Should It Be Banned?		
	(1)	(2)	(3)
context = Historical Injustice	0.005	−0.352*	
	(0.206)	(0.169)	
explanation = Learn	−0.055		0.070
	(0.203)		(0.211)
explanation = Perpetuate	0.492*		0.193
	(0.208)		(0.205)
(context = Historical Injustice):(explanation = Learn)	0.195		
	(0.289)		
(context = Historical Injustice):(explanation = Perpetuate)	−0.263		
	(0.288)		
(group = Advantaged):(explanation = Learn)			−0.063
			(0.289)
(group = Advantaged):(explanation = Perpetuate)			0.325
			(0.288)
group = Advantaged		−0.252	0.001
		(0.166)	(0.206)
(context = Historical Injustice):(group = Advantaged)		0.660**	
		(0.235)	
Constant	1.204***	1.474***	1.205***
	(0.148)	(0.120)	(0.148)
Observations	716	716	716

Table 15: Linear regressions of participants' belief that the algorithm should be banned. Dependent variables: participants agreement with a statement indicating that the system should not be used. Independent variables: dummy variables indicating to which treatment condition participants' were assigned (context and explanation) and participants' racial identity (group). Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

	Dependent Variables		
	Blame - Algorithm	Blame - Company	Blame - Developer
	(1)	(2)	(3)
context = Historical Injustice	−0.328	0.434*	0.234
	(0.259)	(0.195)	(0.217)
explanation = Learn	−0.465	0.479*	−0.699**
	(0.255)	(0.192)	(0.214)
explanation = Perpetuate	−0.329	0.498*	−0.551*
	(0.261)	(0.197)	(0.220)
(context = Historical Injustice):(explanation = Learn)	0.328	−0.244	−0.261
	(0.363)	(0.274)	(0.305)
(context = Historical Injustice):(explanation = Perpetuate)	0.285	−0.063	−0.265
	(0.362)	(0.273)	(0.304)
Constant	3.903***	4.850***	5.699***
	(0.186)	(0.140)	(0.156)
Observations	716	716	716

Table 16: Linear regressions of blame judgments. Dependent variables: blame judgments of the algorithm itself, the company employing the algorithm, and the developers of the algorithm. Independent variables: dummy variables (context and explanation) indicating to which treatment condition participants' were assigned. Standard errors are shown inside parentheses. *p<0.05; **p<0.01; ***p<0.001.

	Dependent Variables			
	Distributional Fairness	Procedural Fairness	Interpersonal Fairness	Trust
	(1)	(2)	(3)	(4)
context = Historical Injustice	$-0.269^{\dagger}$	-0.352^{*}	$-0.234^{\dagger}$	$-0.281^{\dagger}$
	(0.155)	(0.157)	(0.142)	(0.156)
race = Black	-0.266	-0.474^{*}	-0.027	-0.251
	(0.218)	(0.221)	(0.199)	(0.219)
race = Asian	0.192	0.151	0.227	0.143
	(0.231)	(0.235)	(0.211)	(0.232)
race = Mixed/Other	-0.623^{**}	-0.692^{**}	-0.227	-0.514^{*}
	(0.231)	(0.235)	(0.211)	(0.232)
(context = Historical Injustice):(race = Black)	$0.583^{\dagger}$	0.673^{*}	0.599^{*}	0.382
	(0.307)	(0.312)	(0.281)	(0.309)
(context = Historical Injustice):(race = Asian)	$0.573^{\dagger}$	$0.605^{\dagger}$	$0.490^{\dagger}$	$0.570^{\dagger}$
	(0.312)	(0.318)	(0.285)	(0.314)
(context = Historical Injustice):(race = Mixed/Other)	$0.626^{\dagger}$	0.732^{*}	0.362	0.218
	(0.347)	(0.354)	(0.318)	(0.350)
Constant	-1.322^{***}	-1.114^{***}	-1.338^{***}	-1.273^{***}
	(0.110)	(0.112)	(0.100)	(0.110)
Observations	716	716	716	716

Table 17: Replication of our main findings using different racial groups. Dependent variables: perceived distributional fairness, procedural fairness, interpersonal fairness, and trust in the algorithm. Independent variables: dummy variables (context) indicating to which treatment condition participants' were assigned and participants' self-reported racial identity (race). We use participants who self-reported to be White as the baseline category. Standard errors are shown inside parentheses. †p<0.1; *p<0.05; **p<0.01; ***p<0.001.

	Dependent Variables			
	Distributional Fairness	Procedural Fairness	Interpersonal Fairness	Trust
	(1)	(2)	(3)	(4)
context = Historical Injustice	−0.204	−0.312†	−0.192	−0.252
	(0.158)	(0.161)	(0.144)	(0.159)
ethnicity = Black	−0.231	−0.481*	0.045	−0.246
	(0.213)	(0.217)	(0.194)	(0.215)
ethnicity = Asian	0.228	0.171	0.277	0.169
	(0.234)	(0.238)	(0.213)	(0.236)
ethnicity = Mixed	−0.371	−0.539†	0.155	−0.209
	(0.280)	(0.285)	(0.255)	(0.282)
ethnicity = Other	−0.143	−0.153	−0.262	−0.180
	(0.331)	(0.337)	(0.301)	(0.334)
(context = Historical Injustice):(ethnicity = Black)	0.612*	0.759*	0.619*	0.463
	(0.306)	(0.312)	(0.279)	(0.309)
(context = Historical Injustice):(ethnicity = Asian)	0.558†	0.627†	0.509†	0.569†
	(0.316)	(0.321)	(0.287)	(0.318)
(context = Historical Injustice):(ethnicity = Mixed)	0.264	0.435	−0.055	0.126
	(0.410)	(0.417)	(0.373)	(0.413)
(context = Historical Injustice):(ethnicity = Other)	−0.090	−0.015	0.108	−0.327
	(0.490)	(0.498)	(0.446)	(0.493)
Constant	−1.379***	−1.152***	−1.390***	−1.320***
	(0.112)	(0.114)	(0.102)	(0.113)
Observations	716	716	716	716

Table 18: Replication of our main findings using different racial groups. Dependent variables: perceived distributional fairness, procedural fairness, interpersonal fairness, and trust in the algorithm. Independent variables: dummy variables (context) indicating to which treatment condition participants' were assigned and participants' self-reported ethnicity provided by Prolific (ethnicity). We use White participants as the baseline category. Standard errors are shown inside parentheses. †p<0.1; *p<0.05; **p<0.01; ***p<0.001.