BJM BRITISH JOURNAL of MANAGEMENT
BRITISH ACADEMY OF MANAGEMENT
British Journal of Management, Vol. 00, e70045 (2026)
DOI: 10.1111/1467-8551.70045

Like Human, Like Algorithm: Responses to Algorithmic Discrimination Among Individuals From Protected Classes

Gülen Sarial-Abi [1] and Verdiana Giannetti[2]
[1]Copenhagen Business School, Solbjerg Plads 3, Frederiksberg, 2000, Denmark [2]Marketing Department, Leeds University Business School, University of Leeds, Maurice Keyworth Building, Moorland Rd, Leeds, LS2 9JT, UK
Corresponding author e-mail: gsa.marktg@cbs.dk

Algorithms, commonly used in business practice, often discriminate against members of protected classes (e.g. racial minorities). Previous research findings suggest that individuals, including those from protected classes, under some circumstances, may not respond negatively to discriminatory algorithms. Other evidence suggests the opposite. Given the conflicting evidence, there is an opportunity to understand *how* and *when* protected class members respond to businesses that employ algorithms when these algorithms make predictions or decisions resulting in discrimination. Drawing on an empirical package comprising one secondary data study and four experiments, our research demonstrates that when algorithms are perceived to engage in human-like social categorization, they elicit more negative responses from members of protected classes. This effect is observed across various algorithm features, including nonrepresentative training data, proxy classification rules and non-statistical classification rules. The research's findings extend the literature on algorithmic discrimination and business ethics, providing suggestions to mitigate algorithmic discrimination and improve societal well-being.

Introduction

The increasing reliance on algorithms—programmes that combine statistical models, decision parameters and data inputs to generate outcomes via a set of if-then rules—in decision-making processes has raised significant ethical concerns (Kordzadeh and Ghasemaghaei, 2022; Srinivasan and Sarial-Abi, 2021). Of particular concern is the reinforcement of biases against members of protected classes, or groups protected by law from discrimination based on their characteristics, such as sex, religion, colour, national origin, race, age (individuals over 40) and physical or mental disabilities.

Algorithms target men more than women for ads related to science, technology, engineering and math (STEM) careers (Lambrecht and Tucker, 2019) and reject mortgage loans for qualified Black applicants at a rate 80% higher than for White applicants (Martinez and Kirchner, 2021). In a Pew survey of US adults (n = 4594) on algorithmic decision-making, the majority of respondents (67%) expressed concerns over fairness (Smith, 2018), though attitudes towards algorithmic fairness varied significantly by race; 25% of White respondents considered financial algorithms to be unfair versus 45% of Black respondents, and 49% of White respondents (61% of Black respondents) viewed criminal justice algorithms as unfair. It is now acknowledged that algorithms can, and sometimes do, make discriminatory decisions.

The prevalent use of algorithms in business is driving interest in theory-driven perspectives on artificial intelligence (AI) in business and management among management researchers (Brown *et al.*, 2024). Much of the work on business ethics, generally, and algorithm accountability, specifically, has focused on the ethical implications of using AI in business contexts, such as auditing (Buhmann, Paßmann and Fieseler, 2020; John-Mathews, Cardon and Balagué, 2022; Munoko, Brown-Liburd and Vasarhelyi, 2020). Previous research on business ethics has mainly focused on organizations that rely on algorithm-based HR decision-making to monitor their employees (Vaillant *et al.*, 2025) or has assessed ethicality inferences about the organization (Yan *et al.*, 2024) and fairness perceptions of recruitment processes (Lavanchy *et al.*, 2023; see Table WA1 in the Supporting Information for more literature).

Yet, significant gaps remain in our understanding of algorithmic discrimination. One major unaddressed puzzle—the focus of this research—is how protected class members respond to businesses that employ algorithms that make predictions or decisions resulting in discrimination.

We define algorithmic discrimination (algorithmic preference) as when there are systematically unfavourable (favourable) outcomes for members of a protected class, which are not justified by their non-protected characteristics (Zliobaite, 2017). Our research demonstrates that when algorithms are perceived to engage in *human-like social categorization*, they elicit more negative responses from protected class members. Consistent with prior research (Acerbi and Stubbersfield, 2023; Gao *et al.*, 2025), we define human-like social categorization as the process by which algorithms employ social categorization mechanisms analogous to those used by humans when forming predictions and making decisions. This effect is observed across various algorithm features, including nonrepresentative training data, proxy classification rules and non-statistical classification rules.

This research makes several novel contributions to the study of algorithmic discrimination and the broader management literature. Most notably, it addresses a critical yet underexplored question: How do members of protected classes respond to businesses that use algorithms when those algorithms produce discriminatory outcomes? Existing research has offered two contrasting perspectives. One stream suggests that members of low-status or marginalized groups may internalize a sense of inferiority within unequal social systems (Jost, 2020). In contrast, another stream, grounded in reactance theory (Brehm, 1966), argues that individuals who perceive repeated and unjust restrictions to their freedoms are likely to push back, especially when these restrictions threaten their identity or autonomy (Seemann *et al.*, 2004). Our research provides empirical evidence consistent with psychological reactance theory: members of protected classes exhibit stronger negative responses to businesses that employ discriminatory algorithms. In doing so, we move beyond viewing algorithmic discrimination as a purely technical issue and demonstrate its downstream effects on consumer perceptions and behaviour, particularly among those most directly affected.

Secondly, we empirically demonstrate how members of protected classes respond to algorithmic discrimination, depending on the source of that discrimination. While prior work has identified several sources of algorithmic discrimination—including nonrepresentative training data, proxy classification rules and statistical classification rules (Wang *et al.*, 2024)—most research has focused on detecting and categorizing these issues from technical, legal or system design perspectives. There is not yet a theoretical understanding of how the source of algorithmic discrimination shapes the subjective experiences and reactions of those most directly affected. Our research shows that members of protected classes are more likely to respond negatively to algorithmic discrimination when it arises from specific features of the algorithm used by a company, specifically the use of nonrepresentative training data and proxy classification rules, but not statistical classification rules. By examining how different sources of algorithmic discrimination are perceived and evaluated by protected class members, we demonstrate that individuals do not simply respond to the presence of discrimination in a binary way (i.e. biased vs. unbiased); rather, they interpret and evaluate *how* the discrimination is produced.

Our findings also reveal that protected class members perceive the use of nonrepresentative training data and proxy classification as forms of human-like social categorization, but do not interpret statistical classification in the same way. In the context of algorithmic discrimination, we show that when algorithms are perceived to use human-like social categorization, they elicit more negative responses from protected class members. This finding introduces a novel perspective on how the perceived nature of an algorithm's operation can influence reactions to discriminatory outcomes.

Thirdly, this research investigates the perspectives of protected class members, who have been largely overlooked in empirical investigations. A review of leading academic journals found that fewer than 1% of studies focused on Black individuals (Pittman Claytor, 2019), highlighting a broad neglect in management and organizational research. To help address this gap, we empirically examine how members of three protected classes—racial minorities, women and older adults—respond when their group is subjected to discrimination by a business using algorithmic decision-making. We do not simply ask whether algorithmic discrimination occurs but theorize how such discrimination is perceived and processed by those directly affected.

Our investigation of *how* and *when* protected class members respond to businesses that employ algorithms includes one secondary data study and four experiments, conducted across multiple contexts (e.g. financial services, product recommendations, hiring decisions etc.) and three protected classes (i.e. racial minorities, women, older adults). Our study fundamentally advances scholars' understanding of algorithmic discrimination, providing critical insights into how algorithmic design choices impact perceptions of discrimination, with significant importance for management, marketing, psychology and information systems research.

Theoretical background

Algorithmic discrimination

While algorithms are designed for scalability and impartiality, their predictions do not always align with reality, resulting in algorithmic bias (Angwin *et al.*, 2016). Algorithmic discrimination (preference) occurs when algorithmic bias causes systematically unfavourable (favourable) outcomes for protected class members that are not justified by their non-protected characteristics (Zliobaite, 2017).

Algorithmic discrimination is multifaceted (Wang *et al.*, 2024). It can arise from discriminatory feature selection (Zliobaite, 2015), proxy discrimination (Prince and Schwarcz, 2020), disparate impact (Hellman, 2020), targeted advertising (Speicher *et al.*, 2018) and biased training data (Barocas and Selbst, 2016). Human choices and social and historical contexts—including a lack of diversity in development teams (West, Whittaker and Crawford, 2019), the opacity of algorithmic systems (Pasquale, 2015) and historical biases in data (Crawford and Schultz, 2014)—significantly influence algorithmic outcomes.

Previous research has explored various strategies for mitigating algorithmic discrimination, including incorporating fairness constraints into machine learning models (Zafar *et al.*, 2017), algorithmic auditing (Reisman *et al.*, 2018) and pre-processing training data (Kamiran and Calders, 2012). In discussing the policy and legal implications of algorithmic bias, the literature has emphasized the need for new regulations for algorithmic decision-making (Selbst and Powles, 2018); the challenges in enforcing accountability and transparency in algorithmic systems (Kroll *et al.*, 2017); and the applicability of existing anti-discrimination laws (Kim, 2017).

Despite extensive research on algorithmic bias (Dietvorst, Simmons and Massey, 2018; Lavanchy *et al.*, 2023; Longoni, Bonezzi and Morewedge, 2019; Srinivasan and Sarial-Abi, 2021; Yan *et al.*, 2024), empirical evidence of the real-world impact of algorithmic discrimination is scarce, and there is limited understanding of how specific algorithm features impact perceptions of discrimination. Recent studies (e.g. Pethig and Kroenung, 2023; Smith, 2018; Wang *et al.*, 2024) have highlighted the need for examining these features in detail.

The role of social categorization

Social categorization simplifies the world's social structure, allowing individuals to efficiently process information about others by grouping individuals based on salient characteristics such as race, gender or age (Fiske and Neuberg, 1990; Tajfel and Turner, 1986). Individuals use social categorization to infer others' roles, abilities and personality traits based on their membership in a given social category—which may include protected classes (Bodenhausen and Macrae, 1998)—resulting in stereotyping and bias. For instance, individuals perceive a face as having a darker complexion if non-facial cues suggest that the face is that of a Black (vs. Hispanic) person (MacLin and Malpass, 2001).

We propose that when an algorithm discriminates against members of a protected class, those individuals perceive the algorithm as a non-human entity engaging in human-like behaviour—specifically, social categorization. According to social categorization theory (Tajfel and Turner, 1979), this process is a fundamental aspect of human cognition that helps individuals simplify complex social information. However, social categorization is not merely a neutral act of classification; in shaping how people perceive others and how they are perceived in return, it carries social and moral implications, influencing stereotypes, intergroup dynamics and social judgements (Fiske and Neuberg, 1990).

In interpersonal contexts, social categorization typically involves contextual judgements and moral awareness, especially when applied to socially meaningful and historically sensitive categories. When algorithms are perceived as performing similar categorization—particularly when it leads to discriminatory outcomes—they may be seen as stepping outside of their appropriate role. Rather than acting as neutral, data-driven tools, such algorithms appear to imitate human social behaviour in ways that are ethically charged. This perceived blurring of boundaries between human and machine behaviour may violate normative expectations about the objectivity and impartiality of algorithmic systems. For members of protected classes, who are especially attuned to the harms of group-based discrimination, the idea that an algorithm is sorting people based on group membership may be particularly troubling. Building on these ideas, we propose that, following algorithmic discrimination, protected class members respond negatively to the algorithm and, by extension, to the organization that employs it. These negative reactions stem from the perception that the algorithm is engaging in human-like social categorization, an act seen as inappropriate and morally problematic for a non-human system.

With respect to algorithmic preference, we expect that individuals will respond positively or neutrally following an algorithmic preference for members of their protected class. Research on stereotyping and discrimination suggests that Black and Hispanic individuals in the United States generally (e.g. Harrison *et al.*, 2006), but not always (Aberson, 2003; Peterson, 1994), endorse affirmative action programmes aimed at reducing disparities arising from structural racism. In addition, because certain forms of preference for historically disadvantaged protected classes are allowed under the law (e.g.

the Civil Rights Act of 1964), protected class members may not perceive the algorithm as using human-like social categorization to develop preferential predictions. Hence, we propose:

H 1. *Algorithmic discrimination against (vs. preference for) members of a protected class leads to the perception that the algorithm uses human-like social categorization, resulting in more negative responses to the company following algorithmic discrimination against (vs. preference for) members of one's protected class.*

Utilizing the moderation-of-process design approach (Spencer, Zanna and Fong, 2005), we next introduce three algorithm features (i.e. nonrepresentative training data, proxy classification rules, statistical classification rules) that may change the way individuals perceive the use of human-like social categorization by algorithms (Wang *et al.*, 2024).

Nonrepresentative training data as a source of algorithmic discrimination

Training data is the foundation on which algorithms learn about targets for making predictions. One way algorithmic discrimination may occur is when algorithms use incomplete or nonrepresentative training data (Haim *et al.*, 2022; Wang *et al.*, 2024). Nonrepresentative training data refers to data that does not accurately reflect the target population the algorithm is intended to serve (Cawley and Talbot, 2010) and can lead to discrimination as algorithms reproduce, perpetuate and exacerbate existing societal biases (Barocas and Selbst, 2016). If women or minorities, for instance, were underrepresented or underpromoted in its training data, a hiring algorithm trained on past employment records may learn to discriminate against women or minorities (Ajunwa, 2019). Indeed, nonrepresentative training data is the primary cause of bias against ethnic minorities in facial recognition algorithms (Centre for Data Ethics and Innovation, 2020). Most training datasets for facial recognition algorithms are more than 75% male and 80% pale-skinned White, leading to misidentification of darker-skinned women (Hardesty, 2018). This underrepresentation of protected classes in training data can occur for many reasons, including lower representation in the real world (Calders and Zliobaite, 2013).

We suggest that individuals expect algorithms to learn autonomously (i.e. without biased human intervention) to improve their predictions (Mitchell, 1997; Youssef, Abramoff and Char, 2023). Such ongoing machine learning is crucial when historical data is not representative of a target population because of changing characteristics (Berger *et al.*, 2021). When an algorithm is perceived not to be continuously learning, as in the case of nonrepresentative training data, individuals may perceive the algorithm as using human-like social categorization of protected class members to develop its predictions, and may feel its decisions and predictions are based on biased agents' use of incomplete or nonrepresentative training data. We predict that this perception leads to more negative responses to the organization by protected class members following algorithmic discrimination against members of their protected class. When discrimination is not the result of nonrepresentative training data, individuals may perceive the algorithm to be learning autonomously and not influenced by biased agents' use of incomplete or unrepresentative data, making its decisions and predictions less biased. Hence, we propose:

H 2. *Individuals respond more negatively to a company, following algorithmic discrimination against members of their protected class, when the source of algorithmic discrimination is (vs. is not) nonrepresentative training data.*

Proxy classification as a source of algorithmic discrimination

Algorithmic discrimination may also occur when algorithms use proxy variables that correlate with protected characteristics, known as proxy classification (Citron and Pasquale, 2014; Fagan, 2025; Wang *et al.*, 2024). Regardless of its benign intent or underlying efficiency logic, proxy classification results in the denial of deserved outcomes to individuals in protected classes. ZIP codes, for instance, may be used as a proxy for race to determine credit risk in the United States, leading to discriminatory outcomes even when race is not explicitly considered in decisions and predictions (Citron and Pasquale, 2014; Lindholm *et al.*, 2024). In such situations, individuals from a given protected class have significantly worse outcomes based on characteristics associated with their membership in the protected class than those not associated with the class (Prince and Schwarcz, 2020). Policing algorithms also overemphasize the predictive role of ZIP codes, incorrectly associating low-income Black and Latino neighbourhoods with high criminality (Barocas and Selbst, 2016).

We propose that when the source of algorithmic discrimination is a proxy classification rule, members of protected classes may perceive the algorithm as using human-like social categorization of protected class members to develop its predictions because people often use proxies to categorize others outside the algorithmic context. Non-native accents, for example, which are correlated with the protected characteristic of immigrant status, are used as proxies for lower intelligence and competence, including in workplace settings (Gluszek and Dovidio, 2010). When the discrimination is not a result of proxy classification, individuals may perceive

the algorithm as not influenced by biased agents' use of proxies, making its decisions and predictions less biased. Hence, we propose:

H 3. *Individuals respond more negatively to a company, following algorithmic discrimination against members of their protected class, when the source of algorithmic discrimination is (vs. is not) a proxy classification rule.*

Statistical classification as a source of algorithmic discrimination

A third way algorithmic discrimination occurs is when algorithms use factual, readily accessible, historical information on features of protected class members to develop their predictions, such as that women take more leave than men because of elder and child care responsibilities (Coate and Loury, 1993; Lang and Kahn-Lang Spitzer, 2020). This statistical classification can give rise to informational issues that influence rational decision-making, leading to unfair outcomes even in the absence of discriminatory intent (Patty and Penn, 2023). Similar to proxy classification, statistical classification is widely used in practice and has been extensively examined in economics (Becker, 1971).

We propose that when the source of algorithmic discrimination is a statistical classification rule, protected class individuals may be less likely to perceive the algorithm as using human-like social categorization of protected class members to develop its predictions. This is because statistical discrimination outside the algorithmic context is typically not attributed to prejudice. Profit-maximizing employers, for example, seek to cost-effectively hire the highest-ability workers (Phelps, 1972). When organizations face a signal extraction problem and cannot accurately estimate workers' skills, they rely on other individual-level indicators of productivity (Prince and Schwarcz, 2020), thereby rationalizing any resultant discriminatory actions (Tilcsik, 2020). When algorithms use statistical classification to develop predictions, they may be similarly perceived as using 'objective' information about the characteristics of protected class members. This use of objective information is consistent with observers' views that algorithms promise neutrality in decision-making by avoiding human-like social categorization. When the discrimination is not the result of a statistical classification rule, individuals may perceive that the algorithm is not objective, making its decisions and predictions more biased. Hence, we propose:

H 4. *Individuals respond less negatively to a company, following algorithmic discrimination against members of their protected class, when the source of algorithmic discrimination is (vs. is not) a statistical classification rule.*

Studies

We conducted one secondary data study and four experimental studies that test the hypotheses. The Institutional Review Board of the authors' home institutions reviewed and approved the experimental designs before commencement of the studies. For all experimental studies, the target sample was based on a priori power analyses (power of 0.90, small-to-medium effect sizes [$d = 0.50$], alpha level of 0.05; Faul *et al.*, 2007). We provide the study stimuli in Web Appendix B in the Supporting Information. For Studies 2–4, we tested whether the manipulations change the extent to which participants perceive the algorithm as using human-like social categorization for members of protected classes to develop its predictions in separate studies (see Web Appendix C in the Supporting Information). The five studies were designed to build cumulatively, with each addressing limitations and open questions from the preceding one (see Web Appendix D in the Supporting Information for an overview of the studies).

Study 1A used secondary data to provide initial correlational evidence for H1, linking algorithmic discrimination against a protected class (i.e. older adults) to perceptions of human-like social categorization and negative company responses. Study 1B established causality through an experiment in the gender context.

Study 2 tested H2 by examining nonrepresentative training data as a source of discrimination. Study 3 tested H3, focusing on proxy classification rules to assess whether perceptions of human-like social categorization generalize across different sources of discrimination. Study 4 tested H4 by investigating statistical classification rules, providing further evidence on how varying sources of algorithmic discrimination influence perceptions of algorithms as engaging in human-like social categorization and, subsequently, company responses.

Together, the studies offer conceptual replication across contexts (age, gender, race) and methodological triangulation (secondary data and experiments), progressively strengthening both the internal validity of our causal claims and the external validity of our theoretical insights. This multi-study design moves beyond documenting the existence of algorithmic discrimination to explain when, why and how such discrimination triggers company-directed backlash through perceptions of human-like social categorization.

Study 1A

Study 1A analyses complaints from the Consumer Financial Protection Bureau's (CFPB) Consumer Complaint Database filed between 2018 and 2021 that included a narrative (i.e. text-based complaint; $N =$

Table 1. Results of Study 1A

Dependent variable:	Negative response		Social categorization
Variable	Column 1	Column 2	Column 3
Algorithmic discrimination	0.20*** (0.05)	0.16*** (0.06)	0.51*** (0.09)
Social categorization		0.23*** (0.05)	
Year indicators	Yes	Yes	Yes
Issue indicators	Yes	Yes	No
Observations	2244	2244	2244
Wald χ^2 (4)			38.15
R-squared	5.48%	6.46%	

Note: All models include a constant. Unstandardized parameter estimates and robust standard errors in parentheses.
*** indicates $p < 0.01$.

621,219) to examine whether perceived algorithmic discrimination against older adults is associated with the perception that the algorithm uses human-like social categorization.

To identify relevant complaints against algorithms in the database (e.g. 'The credit companies won't change my score because the algorithms say we are in trouble […]. No human looks at the data and corrects it […]'), we searched for narratives containing algorithm-related keywords (i.e. algorithm*, software*, computer*, or automat*; $n = 24{,}468$). The CFPB flags complaints from 'Older Americans' (i.e. older consumers), who are often discriminated against in the financial services sector (Silberg and Manyika, 2019), enabling us to focus on complaints filed by older individuals that included algorithm-related keywords ($n = 2244$). Complaints were then classified as involving *Algorithmic Discrimination* (binary variable = 1, 0 otherwise) if they included discrimination-related keywords (i.e. discriminat*, bias*, prejudic*, *fair*, *justic*; $n = 264$, 11.76%).

To operationalize human-like social categorization, we developed a custom dictionary encompassing six stereotypes of older adults with corresponding keywords (Hummert *et al.*, 1994; Remedios, Chasteen and Packer, 2010): 'impaired' (e.g. senile, fragile); 'despondent/vulnerable' (e.g. alienated, dejected); 'recluse' (e.g. sedentary, isolated); 'golden' (e.g. cottage, vacation); 'grandparent' (e.g. grandchild, grandparent) and 'conservative' (e.g. nostalgic, religious). Complaints containing at least one stereotype-related keyword were classified as referring to the social categorization of older adults ($n = 444$, 19.79%). The full list of keywords is available upon request.

We measured negative responses by the complaints' negative sentiment (i.e. tone_neg; M = 1.04, SD = 0.90, min = 0, max = 7.34) using LIWC-22 (Pennebaker *et al.*, 2015). We examined the effect of perceptions of algorithmic discrimination on negative responses, both without and with the proposed mediator, human-like social categorization, using the following model:

$$NegativeResponse_i = \alpha_0 + \alpha_1 AlgorithmicDiscrimination_i + \alpha_2 SocialCategorization_i + \gamma_i + \mu_i + \varepsilon_{1i}, \quad (1)$$

where $AlgorithmicDiscrimination_i$ is a binary variable (1 if complaint i includes at least one discrimination-related keyword, 0 otherwise), $SocialCategorization_i$ is a binary variable (1 if complaint i includes at least one stereotype-related keyword, 0 otherwise) and ε_{1i} is the error term. The model includes indicators for the year of the complaint (γ_i) and the issue type (μ_i; e.g. closing account, false statements etc.).

The results in Table 1, Column 1, show that perceptions of algorithmic discrimination are associated with more negative responses ($b = 0.20$, $p < 0.01$). The results in Column 2, which includes the mediator, indicate that both perceptions of algorithmic discrimination ($b = 0.16$, $p < 0.01$) and perceptions of human-like social categorization ($b = 0.23$, $p < 0.01$) contribute to heightened negative responses.

Next, we analysed the effect of perceptions of algorithmic discrimination on perceptions of human-like social categorization, a binary dependent variable, using a probit specification:

$$P\,(SocialCategorization_i = 1) = \Phi\,(\beta_0 + \beta_1 AlgorithmicDiscrimination_i + \pi_i) \quad (2)$$

where Φ is the cumulative distribution function of the standard normal distribution. While the model includes indicators for the year (π_i), we do not include indicators for the issue type as older consumers perceive no social categorization for some issues (always a value of 0), resulting in observations being dropped.

The results in Table 1, Column 3, indicate that perceptions of algorithmic discrimination are associated with a higher likelihood of perceptions of human-like social categorization ($b = 0.51$, $p < 0.01$), supporting

the proposed mediation mechanism: (1) the mediator (i.e. social categorization) significantly increases the dependent variable (i.e. negative responses; Table 1, Column 2); and (2) the independent variable (i.e. algorithmic discrimination) significantly increases the mediator (i.e. social categorization; Table 1, Column 3).

Study 1A provides initial, correlational evidence supporting our proposed mechanism: older adults respond negatively to companies following algorithmic discrimination because they perceive algorithms as using human-like social categorization. We report additional analyses and robustness analyses for this study in Web Appendix E in the Supporting Information.

Study 1B

Study 1B uses an experimental design to examine whether the perception that an algorithm employs human-like social categorization results in more negative responses to the company following algorithmic discrimination against (vs. preference for) members of one's protected class. In this study, we used gender as the context for the protected class.

Participants and procedure. One hundred ninety-one women (M_{age} = 40.71, SD = 13.60) participated in the study through the Prolific online platform in exchange for monetary compensation. All participants read a scenario in which a high-tech company, Tech Specialist, faces a crisis because the algorithm that assisted Tech Specialist in recruiting was systematically biased against female candidates. The context was chosen based on Amazon's AI recruiting tool, which was found to be discriminating against female candidates (Dastin, 2018). We then randomly assigned participants to the algorithmic discrimination or preference condition and measured participants' attitude towards Tech Specialist using a 5-item, 7-point semantic differential scale adapted from Ahluwalia, Burnkrant and Unnava (2000): bad/good, low quality/high quality, undesirable/desirable, harmful/beneficial, unfavourable/favourable (M = 3.05, SD = 1.34; $\alpha = 0.96$).

Our stimuli intentionally did not specify the algorithm's internal features. Instead, consistent with our theoretical framework, we examined whether discriminatory outcomes lead participants to perceive the algorithm as relying on gender-typed social categorization. We operationalized this perception using an adapted version of the Personal Attributes Questionnaire (Spence, Helmreich and Stapp, 2013) anchored to the algorithm's evaluation of women applicants (M = 3.43, SD = 1.48; $\alpha = 0.95$). Specifically, participants indicated the extent to which they agreed that, in rating female applicants, the automated system used attributes that people generally associate with women, on a 7-point scale. Sample attributes included 'women are emotional', 'women give up very easily' and 'women go to pieces under pressure'. This approach enabled a direct test of the proposed mechanism through mediation analysis (Hayes, 2013).

As a manipulation check, we asked participants to indicate the extent to which Tech Specialist caused harm to female applicants on a 7-point Likert scale (1 = not at all to 7 = very much; M = 4.87, SD = 2.11). The results of a one-way ANOVA on participants' perceptions of harm were significant, $F(1,189) = 119.06$, $p < 0.001$. Perceptions of harm were higher for the algorithmic discrimination (vs. preference) condition, $M_{DISCRIMINATION} = 6.22$, SD = 1.26 vs. $M_{PREFERENCE} = 3.60$, SD = 1.96. Finally, participants provided their basic demographic information.

Results. The results of a one-way ANOVA on participants' attitudes towards Tech Specialist are significant, $F(1,189) = 31.70$, $p < 0.001$. Participants' responses to the company are more negative for the algorithmic discrimination (vs. preference) condition, $M_{DISCRIMINATION} = 2.53$, SD = 1.31 vs. $M_{PREFERENCE} = 3.55$, SD = 1.18. The results hold, $F(1,188) = 33.16$, $p < 0.001$, when we control for age, $p = 0.022$.

We next used PROCESS model 4 (Hayes, 2013) to test for the mediation prediction (H1). We regressed participants' (1) attitudes towards Tech Specialist on the algorithmic discrimination (vs. preference) condition ($b = -1.02$, $p < 0.001$); (2) perceptions of the algorithm's use of human-like social categorization of protected class members on the algorithmic discrimination (vs. preference) condition ($b = 1.01$, $p < 0.001$); and (3) attitudes towards Tech Specialist on both the algorithmic discrimination (vs. preference; $b = -0.81$, $p < 0.001$) condition and perceptions of the algorithm's use of human-like social categorization ($b = -0.201$, $p = 0.002$). The indirect effect of the algorithmic discrimination (vs. preference) condition on female participants' attitude towards Tech Specialist is significant ($b = -0.2029$; SE = 0.0848; 95% CI = −0.3884, −0.0569; 10,000 bootstrap samples).

The results of Study 1B show that women respond more negatively to a company following algorithmic discrimination against (vs. preference for) other women, as they perceive the algorithm as using human-like social categorization, supporting H1.

Study 2

Study 2 investigates whether African American individuals respond more negatively to a company following algorithmic discrimination against members of their protected class when the algorithm is trained on nonrepresentative data.

Participants and procedure. A total of 366 African American individuals (175 female; M_{age} = 32.36, SD = 9.65) participated in the study through the Prolific

online platform in exchange for monetary compensation. We used a 2 × 2 (algorithmic bias: algorithmic discrimination, algorithmic preference × source: nonrepresentative training data, not) between-subjects design.

We informed participants that a leading US hospital and healthcare chain, HealthPoint Hospitals, experienced negative backlash in the press and on social media because a facial recognition algorithm they used made a mistake in providing automated access for new residency doctor fellows to its facilities. The context was chosen based on reports that facial recognition algorithms used in the US government's mobile app for asylum application failed to recognize Black individuals, preventing them from submitting applications (Del Bosque, 2023). We first randomly assigned participants to the algorithmic discrimination (vs. preference) condition and then to the nonrepresentative training data (vs. not) condition.

We measured participants' intentions to recommend HealthPoint Hospitals (i.e. 'I would tell other people about HealthPoint Hospitals' and 'I would recommend HealthPoint Hospitals to friends and family'; 1 = strongly disagree to 7 = strongly agree; adapted from Homburg, Schwemmle and Kuehnl (2015); M = 3.75, SD = 1.69; $\alpha = 0.80$). We also measured participants' attitude towards HealthPoint Hospitals and found supporting evidence for our prediction. Further details can be provided upon reasonable request. Participants also provided their perceptions of the extent to which the news was from a credible source (M = 4.62, SD = 1.68) and the extent to which the news was believable (M = 4.77, SD = 1.73) using a 7-point scale (1 = not at all to 7 = very much). We found no interaction effect of the algorithmic discrimination (vs. preference) and nonrepresentative training data (vs. not) conditions on information source credibility, $F(1,362) = 0.87$, $p = 0.351$, and news believability, $F(1,362) = 0.72$, $p = 0.396$. Finally, participants provided their basic demographic information.

Results and discussion. Consistent with H2, an ANOVA analysis on participants' intentions to recommend HealthPoint Hospitals confirmed the predicted interaction effect of the algorithmic discrimination (vs. preference) and nonrepresentative training data (vs. not) conditions, $F(1,362) = 7.64$, $p = 0.006$. There is no main effect of the algorithmic discrimination (vs. preference) condition ($p = 0.249$) but a marginally significant effect of the nonrepresentative training data (vs. not) condition ($p = 0.071$; see Figure 1).

Supporting H2, participants' recommendation intentions after algorithmic discrimination are lower when the source is nonrepresentative training data (vs. not), $M_{NONREPRESENTATIVE} = 3.24$, SD = 1.53 vs. $M_{NOT} = 4.04$, SD = 1.82, $F(1,362) = 10.74$, $p = 0.001$. Participants' recommendation intentions following algorithmic preference do not significantly differ with the source, $M_{NONREPRESENTATIVE} = 3.93$, SD = 1.78 vs. $M_{NOT} = 3.76$, SD = 1.50, $F(1,362) = 0.44$, $p = 0.506$.

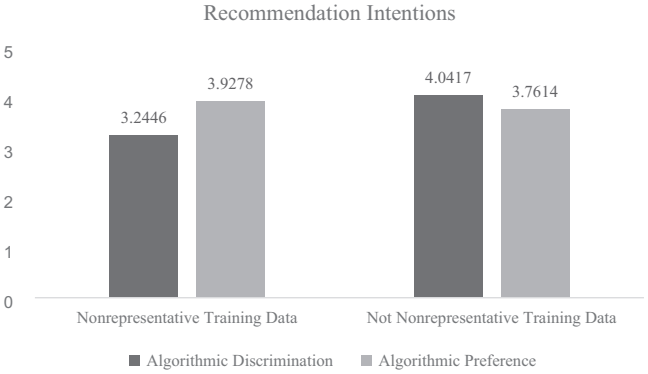

Figure 1. Effect of nonrepresentative training data on protected class responses to a hospital and healthcare chain following algorithmic discrimination against (vs. preference for) members of their protected class

Additional analyses indicated that when the source is nonrepresentative training data, participants' recommendation intentions following algorithmic discrimination are lower than following algorithmic preference, $M_{DISCRIMINATION} = 3.24$, SD = 1.53 vs. $M_{PREFERENCE} = 3.93$, SD = 1.78, $F(1,362) = 7.64$, $p = 0.006$. There is no difference between participants' recommendation intentions following algorithmic discrimination versus preference when the source of bias is unknown, $M_{DISCRIMINATION} = 4.04$, SD = 1.82 vs. $M_{PREFERENCE} = 3.76$, SD = 1.50, $F(1,362) = 1.30$, $p = 0.255$.

The results of Study 2 show that, following algorithmic discrimination against other African American individuals, African American participants are less likely to recommend a health institution when the source of algorithmic discrimination is nonrepresentative training data (vs. not), supporting H2.

Study 3

Study 3 investigates whether African American individuals respond more negatively to a company following algorithmic discrimination against members of their protected class when the algorithm uses a proxy classification rule.

Participants and procedure. A total of 380 African American individuals (240 female; $M_{age} = 35.32$, SD = 12.11) participated in the study through the Prolific online platform in exchange for monetary compensation. We used a 2 × 2 (algorithmic bias: algorithmic discrimination, algorithmic preference × source: proxy classification rule, not) between-subjects design.

We informed participants that an insurance provider company, Safe&Sound, was experiencing a crisis because of a mistake made by an algorithm in the assessment of car insurance premiums for its customers. The

context was chosen because Black drivers pay 46% more than White drivers on auto insurance premiums (Heller and DeLong, 2024). We randomly assigned participants to the algorithmic discrimination (vs. preference) condition and then to the proxy classification rule (vs. not) condition. We measured participants' attitude towards Safe&Sound, using the same 5-item scale used in Study 1B ($M = 3.16$, $SD = 1.70$; $\alpha = 0.97$).

As a manipulation check for the algorithmic discrimination (vs. preference) condition, participants indicated the extent to which they thought that the algorithm used by Safe&Sound had negative consequences for its Black customers (1 = not at all to 7 = very much; $M = 4.68$; $SD = 2.31$). The one-way ANOVA results for participants' perceptions of negative consequences were significant, $F(1,378) = 264.39$, $p < 0.001$. Perceived negative consequences were higher for the algorithmic discrimination (vs. preference) condition, $M_{DISCRIMINATION} = 6.19$, $SD = 1.42$ vs. $M_{PREFERENCE} = 3.24$, $SD = 2.05$. As a manipulation check for the proxy classification rule (vs. not) condition, participants indicated the extent to which they thought that the algorithm used proxies to determine the insurance premium for its customers (1 = not at all to 7 = very much; $M = 4.53$; $SD = 1.52$). The results of the one-way ANOVA on participants' perceptions of the use of proxies were significant, $F(1,378) = 55.47$, $p < 0.001$. Perceived use of proxies was higher for the proxy classification rule (vs. not) condition, $M_{PROXY} = 5.07$, $SD = 1.48$ vs. $M_{NOT} = 3.98$, $SD = 1.37$.

Participants also provided perceptions of the extent to which the news was believable (1 = not at all to 7 = very much; $M = 4.37$; $SD = 1.90$). We found no interaction effect of the algorithmic discrimination (vs. preference) and proxy classification rule (vs. not) conditions on believability, $F(1,376) = 1.49$, $p = 0.223$. Finally, participants provided their basic demographic information.

Results and discussion. Consistent with H3, an ANOVA analysis of participants' attitudes confirmed the predicted interaction effect of the algorithmic discrimination (vs. preference) and proxy classification rule (vs. not) conditions, $F(1,376) = 4.24$, $p = 0.040$. There is a main effect of the algorithmic discrimination (vs. preference) condition ($p < 0.001$) and a marginally significant main effect of the proxy classification rule (vs. not) condition ($p = 0.057$; see Figure 2).

Further supporting H3, participants' responses following algorithmic discrimination were more negative when the source was a proxy classification rule (vs. not), $M_{PROXY} = 1.81$, $SD = 0.74$ vs. $M_{NOT} = 2.34$, $SD = 1.36$, $F(1,376) = 7.67$, $p = 0.006$. There was no difference in participants' responses following algorithmic preference regardless of source, $M_{PROXY} = 4.21$, $SD = 1.57$ vs. $M_{NOT} = 4.19$, $SD = 1.43$, $F(1,376) = 0.012$, $p = 0.913$.

Additional analyses indicated that participants' responses following algorithmic discrimination (vs. preference) were more negative both when the source was a proxy classification rule, $M_{DISCRIMINATION} = 1.81$, $SD = 0.74$ vs. $M_{PREFERENCE} = 4.21$, $SD = 1.57$, $F(1,376) = 159.93$, $p < 0.001$, and when the source was unknown, $M_{DISCRIMINATION} = 2.34$, $SD = 1.36$ vs. $M_{PREFERENCE} = 4.19$, $SD = 1.43$, $F(1,376) = 93.23$, $p < 0.001$.

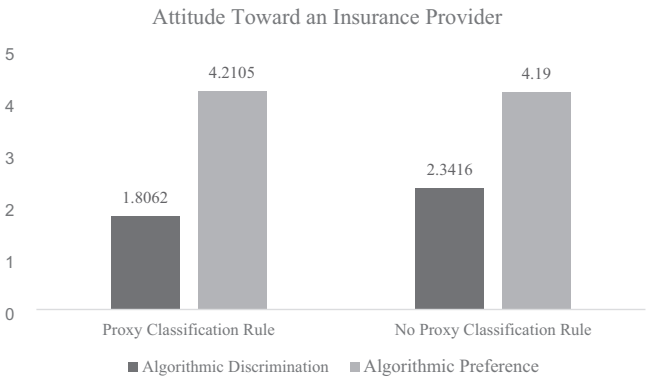

Figure 2. Effect of proxy classification rule on protected class responses to an insurance provider company following algorithmic discrimination against (vs. preference for) members of their protected class

The results of Study 3 show that African American participants are more likely to respond negatively to an insurance provider following algorithmic discrimination against other African American individuals when the source of algorithmic discrimination is a proxy classification rule (vs. not), supporting H3.

Study 4

Study 4 investigates whether women respond less negatively to a company following algorithmic discrimination against members of their protected class when the algorithm uses a statistical classification rule.

Participants and procedure. A total of 422 women ($M_{age} = 39.07$, $SD = 13.63$) participated in the study through the Prolific online platform in exchange for monetary compensation. We used a 2 × 2 (algorithmic bias: algorithmic discrimination, algorithmic preference × source: statistical classification rule, not) between-subjects design.

We informed participants that a leading university in the United States was experiencing a crisis because of a mistake in online course offerings to female students. The context was chosen based on a study of gender discrimination in STEM job ad delivery, in which Lambrecht and Tucker (2019) found that fewer women saw the ad than men as the algorithm used historical information to distribute the ad. We randomly assigned participants to the algorithmic discrimination (vs. preference) condition and then to the statistical classification rule (vs. not) condition.

We measured participants' attitude towards the university undergraduate programmes office, using the same 5-item scale used in Study 1B ($M = 4.37$, $SD =$

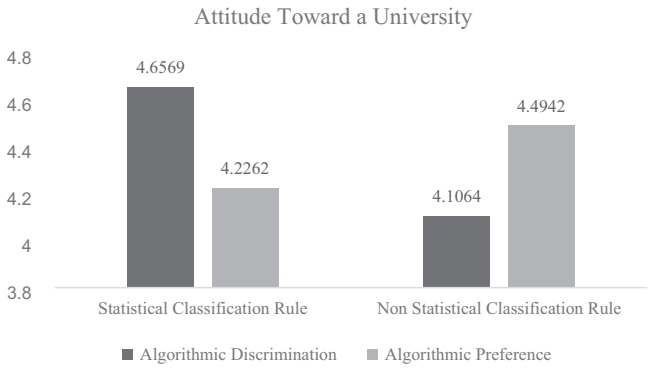

Figure 3. Effect of statistical classification rule on protected class responses to a university following algorithmic discrimination against (vs. preference for) members of their protected class

1.68; $\alpha = 0.97$). Participants also provided perceptions of the extent to which the news was believable (1 = not at all to 7 = very much; M = 4.97, SD = 1.57). We found no interaction effect of the algorithmic discrimination (vs. preference) and statistical classification rule (vs. not) conditions on news believability, $F(1,418) = 0.84$, $p = 0.359$. Finally, participants provided their basic demographic information.

Results and discussion. Consistent with H4, an ANOVA analysis on participants' attitudes towards the university confirmed the predicted interaction effect of the algorithmic discrimination (vs. preference) and statistical classification rule (vs. not) conditions, $F(1,418) = 6.33$, $p = 0.012$. There was no main effect of the algorithmic discrimination (vs. preference; $p = 0.895$) or statistical classification rule (vs. not; $p = 0.386$) conditions (see Figure 3).

Supporting H4, participants' responses following algorithmic discrimination were less negative when the source is a statistical classification rule (vs. not), $M_{STATISTICAL} = 4.66$, SD = 1.44 vs. $M_{NOT} = 4.11$, SD = 1.81, $F(1,418) = 5.72$, $p = 0.017$. Participants' responses following algorithmic preference did not differ when the source was a statistical classification rule (vs. not), $M_{STATISTICAL} = 4.23$, SD = 1.84 vs. $M_{NOT} = 4.49$, SD = 1.54, $F(1,418) = 1.36$, $p = 0.244$.

Additional analyses indicate that participants' responses following algorithmic discrimination (vs. preference) were marginally less negative when the source is a statistical classification rule, $M_{DISCRIMINATION} = 4.66$, SD = 1.44 vs. $M_{PREFERENCE} = 4.23$, SD = 1.84, $F(1,418) = 3.47$, $p = 0.063$, but marginally more negative when the source is unknown, $M_{DISCRIMINATION} = 4.11$, SD = 1.81 vs. $M_{PREFERENCE} = 4.49$, SD = 1.54, $F(1,418) = 2.87$, $p = 0.091$.

The results show that female participants are less likely to respond negatively to algorithmic discrimination against other women when the source is a statistical classification rule (vs. not), supporting H4.

Discussion

Evidence from a secondary data study and four experiments supports our hypotheses: protected class members respond more negatively to businesses following algorithmic discrimination against members of their protected class, and the features of algorithms play a role in their responses.

Theoretical contributions

Our findings advance several streams of research (see Web Appendix F in the Supporting Information for an extended discussion of their theoretical and managerial significance).

Firstly, we extend the business ethics and management literature on algorithmic bias by shifting the focus from merely documenting whether algorithms discriminate to explaining how those most affected respond to the companies that deploy them (Brown *et al.*, 2024; Chowdhury *et al.*, 2022; Chowdhury, Budhwar and Wood, 2024; Figueroa-Armijos, Clark and da Motta Veiga, 2023; Khalil, 1993; Kordzadeh and Ghasemaghaei, 2022; Yan *et al.*, 2024). Prior work has shown that algorithms can produce discriminatory outcomes for protected classes (Lambrecht and Tucker, 2019; Sweeney, 2013), yet there is limited theory on the downstream company-directed reactions of affected individuals. By integrating two contrasting perspectives—that protected class members may either internalize inferiority (Jost, 2020) or resist stigmatizing categorizations (Seemann *et al.*, 2004)—we demonstrate that when individuals perceive an algorithm as engaging in human-like social categorization, they are more likely to respond negatively towards the businesses using it. This pattern is consistent with psychological reactance theory (Brehm, 1966), whereby perceived constraints on identity and autonomy trigger efforts to restore freedom (Seemann *et al.*, 2004). Our contribution is non-trivial: discriminatory outcomes alone do not logically entail negative responses to the *company*; we show that the perceived mechanism—specifically, human-like social categorization—activates a theoretically grounded pathway from bias to backlash.

Secondly, we extend the algorithm accountability and legitimacy literature by showing that the source of discrimination matters to those subjectively experiencing it. Research on legitimacy has found that the perceived arbitrariness or morally dubiousness of algorithms reduce legitimacy (Martin and Waldman, 2023). As a complement to existing typologies of discriminatory mechanisms—nonrepresentative training data (Cawley and Talbot, 2010), proxy classification rules (Citron and Pasquale, 2014; Wang *et al.*, 2024) and statistical classification rules (Prince and Schwarcz, 2020)—we

provide empirical evidence that protected class members respond more negatively to discrimination when it stems from nonrepresentative training data or a proxy classification rule, and less negatively when it arises from a statistical classification rule. This finding advances accountability research by linking mechanistic features of algorithms to perceived legitimacy among those directly affected. Though not self-evident, this result is theoretically significant: even though all three mechanisms produce discriminatory outcomes, individuals do not respond uniformly. Instead, their reactions vary systematically by source, uncovering previously untheorized heterogeneity in perceived fairness and blame attribution.

Thirdly, we contribute to research on algorithm aversion and anthropomorphism by identifying perceived human-likeness in social categorization as a psychologically meaningful cue that amplifies negative reactions (Srinivasan and Sarial-Abi, 2021). Specifically, we show that protected class members tend to perceive nonrepresentative training data and proxy classification rules as human-like social categorization, whereas they do not perceive statistical classification rules in the same way. This clarifies *why* reactions diverge by source and extends the literature beyond a general 'aversion to algorithms' towards a more mechanism-based account of when and why such aversion intensifies. We identify a mediating perception that companies and policymakers can target (e.g. by auditing for human-like social categorization in feature design and data selection).

Finally, we extend scholarship on structural disadvantage in business systems (McPhail *et al.*, 2024), adding to the sparse evidence on protected class members' responses to algorithmic discrimination (e.g. Pethig and Kroenung, 2023). We show how members of three distinct protected classes—racial minorities, women and older adults—react when peers within their protected class are discriminated against by a company using an algorithm in its decision-making. This broadens the scope of prior research by demonstrating cross-context generalizability in company-directed responses and by identifying the conditions under which protected class members are especially likely to push back. Contributing to both theory and practice, this finding links systemic disadvantage to behavioural responses towards companies, connecting macro-level inequities to micro-level responses.

In short, we advance the literature by (1) specifying when discriminatory algorithmic decisions translate into company-directed backlash; (2) showing how distinct sources of discrimination map onto different perceptions and responses; and (3) explaining why these effects emerge. It is clear that not all algorithmic discrimination is experienced equally: the perceived human-likeness of the categorization process is a pivotal factor connecting technical features to social judgement and company legitimacy.

Practical implications

The first key practical implication of our research is that managers should be prepared for some negative backlash from protected class members. This backlash happens because individuals perceive algorithmic discrimination as occurring due to algorithms' use of human-like social categorization in generating predictions. Communicating clearly that the algorithm does not develop predictions based on human-like social categorization of protected class members may help mitigate individuals' negative responses following algorithmic discrimination.

The absence of negative responses when individuals do not perceive the algorithm as using human-like social categorization, such as in cases involving statistical classification, does not imply that businesses should be exempt from accountability for algorithmic discrimination. As Martin (2019) suggested, businesses must avoid accountability dissonance by understanding how their algorithms make decisions and by taking responsibility for eliminating bias and discriminatory decisions. Organizational policies and practices for ethical system design should be evaluated based on whether they enable and motivate data engineers, machine learning developers and user experience designers to proactively identify and properly mitigate algorithmic bias. For example, a financial institution using an algorithm to approve loan applications may find that its algorithm disproportionately denies loans to applicants from certain ethnic backgrounds. Following an audit, the company could take steps to mitigate this bias, such as adjusting the algorithm's decision-making criteria to ensure fairer treatment of all applicants. Communicating these efforts to the public can signal the company's commitment to fairness and help strengthen trust among its customers. Understanding the responses of protected class members to algorithmic discrimination enables organizations to employ appropriate mitigation strategies and better manage potential negative consequences.

Secondly, our findings show that individuals respond more negatively to a company, following algorithmic discrimination against members of their protected class, when the source of algorithmic discrimination is nonrepresentative training data (vs. not). This finding has implications for a wide variety of institutions. For instance, if a healthcare provider uses an algorithm to predict patient outcomes or recommend treatments, and the training data lacks sufficient representation of African American patients, the algorithm may yield biased results. This could lead to suboptimal care for African American patients, who may then perceive the

organization as discriminatory. Healthcare providers should prioritize the use of representative training data in their algorithms to prevent discriminatory outcomes and subsequent negative responses from protected class members.

Thirdly, our findings show that individuals respond more negatively to a company, following algorithmic discrimination against members of their protected class, when the source of algorithmic discrimination is a proxy classification rule (vs. not). This finding has implications for businesses that use proxies to make their predictions or decisions, such as insurance providers. For example, if an insurance company uses an algorithm to determine premiums based on factors such as ZIP codes, it may unintentionally engage in discriminatory pricing. If certain ZIP codes associated with higher premiums predominantly correspond to Latino communities, those communities could face disproportionate impacts. By identifying and removing such proxy variables, organizations can ensure that premiums are based on relevant, non-discriminatory factors, leading to fairer pricing for all customers.

Finally, our findings show that individuals respond less negatively to a company, following algorithmic discrimination against members of their protected class, when the source is a statistical classification rule (vs. not). This muted response is concerning, as it may inadvertently legitimize discrimination by businesses and, more importantly, cause broader societal harm. Public policymakers should foster algorithmic literacy in society through educational programmes so that individuals from protected classes understand how algorithmic decisions are made and do not legitimize algorithmic discrimination, regardless of its source.

Limitations and further research

In this study, we are agnostic about whether algorithmic discrimination occurs because of a false negative/positive mismatch between the algorithm's predictions and the underlying reality, or whether it arises because of differential treatment (e.g. algorithm designers' prejudice) or disparate impact. False-positive mismatches in algorithmic decision-making—which have serious implications in judicial and healthcare contexts—are useful areas for further work.

Additionally, we only focused on the responses of protected class members to discrimination against members of their protected class. Further studies focusing on the responses of non-protected class consumers to algorithmic discrimination against protected class members (e.g. women, Black individuals, older adults) will be a useful extension of this work. We also focused on a limited number of algorithm features that can lead to algorithmic discrimination. Future research should explore additional algorithm features and more diverse demographic groups to enhance the generalizability of our findings.

Finally, as representation and allocation harms are often confounded in practice, we focus on allocation harm only. A promising extension to this research would be to examine how protected class members respond to algorithmic discrimination incorporating allocation harm and representation harm.

Conclusion

Our findings contribute to the literature by showing that perceptions of algorithmic human-like social categorization significantly impact protected class members' negative responses to algorithmic discrimination. From a practical standpoint, businesses should be aware of these perceptions to ensure transparency and fairness in algorithmic decision-making. We hope that this research stimulates further work on algorithmic discrimination, a problem that will become increasingly important over time.

Conflict of interest

The authors have no conflicts of interest to declare that are relevant to this article.

References

Aberson, C. L. (2003). 'Support for race-based affirmative action: self-interest and procedural justice', *Journal of Applied Social Psychology*, **33**, pp. 1212–1225.

Acerbi, A. and J. M. Stubbersfield (2023). 'Large language models show human-like content biases in transmission chain experiments', *PNAS*, **120**, art. e2313790120.

Ahluwalia, R., R. Burnkrant and H. R. Unnava (2000). 'Consumer response to negative publicity: the moderating role of commitment', *Journal of Marketing Research*, **37**, pp. 203–214.

Ajunwa, I. (2021). 'An auditing imperative for automated hiring systems', *Harvard Journal of Law & Technology*, **34**, pp. 621–685.

Angwin, J., J. Larson, S. Mattu and L. Kirchner (2016). 'Machine bias: there's software used across the country to predict future criminals and it's biased against Blacks', *Propublica*, May 23, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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

Becker, G. S. (1971). *The Economics of Discrimination*. Chicago, IL: University of Chicago Press.

Berger, B., M. Adam, A. Ruhr and A. Benlian (2021). 'Watch me improve—algorithm aversion and demonstrating the ability to learn', *Business & Information Systems Engineering*, **63**, pp. 55–68.

Bodenhausen, G. V. and C. N. Macrae (1998). 'Stereotype activation and inhibition'. In R. S., Jr. Wyer, (eds), *Advances in Social Cognition*, pp. 1–52. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Brehm, J. W. (1966). *A Theory of Psychological Reactance*. New York, NY: Academic Press.

Brown, O., R. M. Davison, S. Decker, D. A. Ellis, J. Faulconbridge, J. Gore, et al. (2024). 'Theory-driven perspectives on generative ar-

tificial intelligence in business and management', *British Journal of Management*, **35**, pp. 3–23.
Buhmann, A., J. Paßmann and C. Fieseler (2020). 'Managing algorithmic accountability: balancing reputational concerns, engagement strategies, and the potential of rational discourse', *Journal of Business Ethics*, **163**, pp. 265–280.
Calders, T. and I. Zliobaite (2013). 'Why unbiased computational processes can lead to discriminative decision procedures'. In B. Custers, T. Calders, B. Schermer and T. Zarsky (eds), *Discrimination and Privacy in the Information Society*, pp. 43–57. Society. Berlin: Springer.
Cawley, G. C. and N. L. C. Talbot (2010). 'On over-fitting in model selection and subsequent selection bias in performance evaluation', *Journal of Machine Learning Research*, **11**, pp. 2079–2107.
Centre for Data Ethics and Innovation. (2020). 'Review into bias in algorithmic decision-making', https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making
Chowdhury, S., P. Budhwar, P. K. Dey, S. Joel-Edgar and A. Abadie (2022). 'AI–employee collaboration and business performance: integrating knowledge-based view, sociotechnical systems and organisational socialisation framework', *Journal of Business Research*, **144**, pp. 31–49.
Chowdhury, S., P. Budhwar and G. Wood. (2024). 'Generative artificial intelligence in business: towards a strategic human resource management framework', *British Journal of Management*, **35**, pp. 1680–1691.
Citron, D. K. and F. Pasquale (2014). 'The scored society: due process for automated predictions', *Washington Law Review*, **89**, art. 1.
Coate, S. and G. C. Loury (1993). 'Why affirmative-action policies eliminate negative stereotypes?', *American Economic Review*, **83**, pp. 1220–1240.
Crawford, K. and J. Schultz (2014). 'Big data and due process: toward a framework to redress predictive privacy harms', *Boston College Law Review*, **55**, pp. 93–128.
Del Bosque, M. (2023). 'Facial recognition bias frustrates Black asylum applicants to US, advocates say', *The Guardian*, https://www.theguardian.com/us-news/2023/feb/08/us-immigration-cbp-one-app-facial-recognition-bias
Dietvorst, B. J., J. P. Simmons and C. Massey (2018). 'Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them', *Management Science*, **64**, pp. 1155–1170.
Fagan, F. (2025). 'Proxy discrimination after students for fair admissions', https://arxiv.org/abs/2501.03946
Faul, F., E. Erdfelder, A. G. Lang and A. Buchner (2007). 'GPower 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences', *Behavior Research Methods*, **39**, pp 175–191.
Figueroa-Armijos, M., B. B. Clark and S. P. da Motta Veiga. (2023). 'Ethical perceptions of AI in hiring and organizational trust: the role of performance expectancy and social influence', *Journal of Business Ethics*, **186**, pp. 179–197.
Fiske, S. T. and S. L. Neuberg (1990). 'A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation', *Advances in Experimental Social Psychology*, **23**, pp. 1–74.
Gao, Y., D. Lee, G. Burtch and S. Fazelpour (2025). 'Take caution in using LLMs as human surrogates', *PNAS*, **122**, art. e2501660122.
Gluszek, A. and J. F. Dovidio (2010). 'The way they speak: a social psychological perspective on the stigma of nonnative accents in communication', *Personality and Social Psychology Review*, **14**, pp. 214–237.
Haim, N., G. Vardi, G. Yehudai, O. Shamir and M. Irani (2022). 'Reconstructing training data from trained neural networks', https://arxiv.org/abs/2206.07758
Hardesty, L. (2018). 'Study finds gender and skin-type bias in commercial artificial-intelligence systems', *MIT News*, February 11, https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212
Harrison, D. A., D. A. Kravitz, D. M. Mayer, L. M. Leslie and D. Lev-Arey (2006). 'Understanding attitudes toward affirmative action programs in employment: summary and meta-analysis of 35 years of research', *Journal of Applied Psychology*, **91**, pp. 1013–1036.
Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York: The Guilford Press.
Heller, D. and M. DeLong (2024). 'Landmark Washington report on unintentional bias finds racial premium gap in auto insurance pricing in Washington D.C.', *Consumer Federation of America*, November 19, https://consumerfed.org/press_release/landmark-washington-report-on-unintentional-bias-finds-racial-premium-gap-in-auto-insurance-pricing-in-washington-d-c/
Hellman, D. (2020). 'Measuring algorithmic fairness', *Virginia Law Review*, **106**, pp. 811–866.
Homburg, C., M. Schwemmle and C. Kuehnl (2015). 'New product design: concept, measurement, and consequences', *Journal of Marketing*, **79**, pp. 41–56.
Hummert, M. L., T. A. Gartska, J. L. Shaner and S. Strahm (1994). 'Stereotypes of the elderly held by young, middle-aged, and elderly adults', *Journal of Gerontology*, **49**, pp. 240–249.
Dastin, R. (2018). 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters*, 10 October. Available at: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG
John-Mathews, J., D. Cardon and C. Balagué (2022). 'From reality to world. A critical perspective on AI fairness', *Journal of Business Ethics*, **178**, pp. 945–959.
Jost, J. T. (2020). *A Theory of System Justification*. Cambridge, MA: Harvard University Press.
Kamiran, F. and T. Calders (2012). 'Data preprocessing techniques for classification without discrimination', *Knowledge Information Systems*, **33**, pp. 1–33.
Khalil, O. E. M. (1993). 'Artificial decision-making and artificial ethics: a management concern', *Journal of Business Ethics*, **12**, pp. 313–321.
Kim, P. T. (2017). 'Data-driven discrimination at work', *William & Mary Law Review*, **58**, art. 875.
Kordzadeh, N. and M. Ghasemaghaei (2022). 'Algorithmic bias: review, synthesis, and future research directions', *European Journal of Information Systems*, **31**, pp. 388–409.
Kroll, J. A., J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, et al. (2017). 'Accountable algorithms', *University of Pennsylvania Law Review*, **165**, pp. 633–705.
Lambrecht, A. and C. Tucker (2019). 'Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads', *Management Science*, **65**, pp. 2966–2981.
Lang, K. and A. Kahn-Lang Spitzer (2020). 'Race discrimination: an economic perspective', *Journal of Economic Perspectives*, **34**, pp. 68–89.
Lavanchy, M., P. Reichert, J. Narayanan and K. Savani (2023). 'Applicants' fairness perceptions of algorithm-driven hiring procedures', *Journal of Business Ethics*, **188**, pp. 125–150.
Lindholm, M., R. Richman, A. Tsanakas and M. V. Wüthrich (2024). 'What is fair? Proxy discrimination vs. demographic disparities in insurance pricing', *Scandinavian Actuarial Journal*, **9**, pp. 935–970.
Longoni, C., A. Bonezzi and C. K. Morewedge (2019). 'Resistance to medical artificial intelligence', *Journal of Consumer Research*, **46**, pp. 629–650.
MacLin, O. H. and R. S. Malpass (2001). 'Racial categorization of faces: the ambiguous race face effect', *Psychology, Public Policy, and Law*, **7**, pp. 98–118.
Martin, K. (2019). 'Ethical implications and accountability of algorithms', *Journal of Business Ethics*, **160**, pp. 835–850.
Martin, K. and A. Waldman (2023). 'Are algorithmic decisions legitimate? The effect of process and outcomes on perceptions of legitimacy of AI decisions', *Journal of Business Ethics*, **183**, pp. 653–670.

Martinez, E. and L. Kirchner (2021). 'The secret bias hidden in mortgage-approval algorithms', *TheMarkup*, August 25, https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

McPhail, K., M. Kafouros, P. McKiernan and N. Cornelius (2024). 'Reimagining business and management as a force for good', *British Journal of Management*, **35**, pp. 1099–1112.

Mitchell, T. M. (1997). *Machine Learning*. Machine Learning. New York: McGraw-Hill.

Munoko, I., H. L. Brown-Liburd and M. Vasarhelyi (2020). 'The ethical implications of using artificial intelligence in auditing', *Journal of Business Ethics*, **167**, pp. 209–234.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Patty, J. W. and E. M. Penn (2023). 'Algorithmic fairness and statistical discrimination', *Philosophy Compass*, **18**, art. e12891.

Pennebaker, J. W., R. J. Booth, R. L. Boyd and M. E. Francis (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates.

Peterson, R. S. (1994). 'The role of values in predicting fairness judgments and support of affirmative action', *Journal of Social Issues*, **50**, pp. 95–115.

Pethig, F. and J. Kroenung (2023). 'Biased humans, (un)biased algorithms?', *Journal of Business Ethics*, **183**, pp. 637–652.

Phelps, E. S. (1972). 'The statistical theory of racism and sexism', *American Economic Review*, **62**, pp. 659–661.

Pittman Claytor, C. (2019). 'Are Black consumers a bellwether for the nation?' In *Race in the Marketplace: Crossing Critical Boundaries*, pp. 153–172. Cham: Palgrave Macmillan.

Prince, A. E. R. and D. Schwarcz (2020). 'Proxy discrimination in the age of artificial intelligence and big data', *Iowa Law Review*, **105**, pp. 1257–1318.

Reisman, D., J. Schultz, K. Crawford and M. Whittaker (2018). 'Algorithmic impact assessments: a practical framework for public agency' accountability', *AI Now Institute*, April 9, https://ainowinstitute.org/publications/algorithmic-impact-assessments-report-2

Remedios, J. D., A. L. Chasteen and D. J. Packer (2010). 'Sunny side up: the reliance on positive age stereotypes in descriptions of future older selves', *Self and Identity*, **9**, pp. 257–275.

Selbst, A. and J. Powles (2018). 'Meaningful information and the right to explanation', *International Data Privacy Law*, **7**, pp. 233–242.

Seemann, E. A., W. C. Buboltz, S. M. Jenkins, B. Soper and K. Woller (2004). 'Ethnic and gender differences in psychological reactance: the importance of reactance in multicultural counselling', *Counselling Psychology Quarterly*, **17**, pp. 167–176.

Silberg, J. and J. Manyika (2019). 'Notes from the AI frontier: tackling bias in AI (and in humans)', https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans.

Smith, A. (2018). 'Public attitudes toward computer algorithms', https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/

Speicher, T., M. Ali, G. Venkatadri, F. Nunes Ribeiro, G. Arvanitakis, F. Benevenuto, et al. (2018). 'Potential for discrimination in online targeted advertising', *Proceedings of Machine Learning Research*, **81**, pp. 5–19.

Spence, J. T., R. Helmreich and J. Stapp (2013). 'A short version of the attitudes toward women scale (AWS)', *Bulletin of the Psychonomic Society*, **2**, pp. 219–220.

Spencer, S. J., M. P. Zanna and G. T. Fong (2005). 'Establishing a causal chain: why experiments are often more effective than mediational analyses in examining psychological processes', *Journal of Personality and Social Psychology*, **89**, pp. 845–851.

Srinivasan, R. and G. Sarial-Abi (2021). 'When algorithms fail: consumers' responses to brand harm crises caused by algorithm errors', *Journal of Marketing*, **85**, pp. 74–91.

Sweeney, L. (2013). 'Discrimination in online ad delivery', *Communications of the Association of Computing Machinery*, **56**, pp. 44–54.

Tajfel, H. and J. C. Turner (1986). 'The social identity theory of intergroup behavior'. In S. Worchel and W. G. Austin (eds), *Psychology of Intergroup Relation*, pp. 7–24. Chicago, IL: Nelson-Hall.

Tajfel, H. and J. C. Turner (1979). 'An integrative theory of intergroup conflict'. In W. G. Austin and S. Worchel (eds), *The Social Psychology of Intergroup Relations*, pp. 33–47. Monterey, CA: Brooks/Cole.

Tilcsik, A. (2020). 'Statistical discrimination and the rationalization of stereotypes', *American Sociological Review*, **86**, pp. 93–122.

Vaillant, Y., F. Vendrell-Herrero, O. F. Bustinza and Y. Xing (2025). 'HRM algorithms: moderating the relationship between chaotic markets and strategic renewal', *British Journal of Management*, **36**, pp. 529–545.

Wang, X., Y. C. Wu, X. Ji and H. Fu (2024). 'Algorithmic discrimination: examining its types and regulatory measures with emphasis on US legal practices', *Frontiers in Artificial Intelligence*, **7**, art. 1320277.

West, S. M., M. Whittaker and K. Crawford (2019). *Discriminating Systems: Gender, Race, and Power in AI*. AI Now Institute.

Yan, C., Q. Chen, X. Zhou, X. Dai and Z. Yang (2024). 'When the automated fire backfires: the adoption of algorithm-based HR decision-making could induce consumer's unfavorable ethicality inferences of the company', *Journal of Business Ethics*, **190**, pp. 841–859.

Youssef, A., M. Abramoff and D. Char (2023). 'Is the algorithm good in a bad world, or has it learned to be bad? The ethical challenges of "Locked" versus "Continuously Learning" and "Autonomous" versus "Assistive" AI tools in healthcare', *The American Journal of Bioethics*, **23**, pp. 43–45.

Zafar, M. B., I. Valera, M. Gomez-Rodriguez and K. P. Gummadi (2017). 'Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment', https://arxiv.org/abs/1610.08452

Zliobaite, I. (2015). 'A survey on measuring indirect discrimination in machine learning', https://arxiv.org/abs/1511.00148

Zliobaite, I. (2017). 'Measuring discrimination in algorithmic decision making', *Data Mining and Knowledge Discovery*, **31**, pp. 1060–1089.

Gülen Sarial-Abi is a Professor of Marketing, Copenhagen Business School, Frederiksberg, Denmark. Her research explores how financial scarcity, digitalization, and social crises influence consumer behavior, and how interventions can improve decision-making and well-being. Her publications appeared in top-tier journals including *Journal of Marketing, Journal of Consumer Research, Journal of Consumer Psychology, Journal of the Academy of Marketing Science, and International Journal of Research in Marketing*.

Verdiana Giannetti isAssociate Professor of Marketing at Leeds University Business School, University of Leeds, Leeds, UK. Dr. Giannetti's research interests lie in the broad domain of empirical marketing strategy, with a particular focus on product innovation. More specifically, her current projects focus on (1) international new product

launch decisions and on (2) the antecedents and consequences of product-harm crises. Her work appeared in journals including *Journal of the Academy of Marketing Science*, *Marketing Letters*, *Psychology & Marketing*, *Journal of International Marketing and International Marketing Review*.

Supporting Information

Additional supporting information can be found online in the Supporting Information section at the end of the article.

Supporting Information