PNAS NEXUS

PNAS Nexus, 2025, 4, pgaf089
https://doi.org/10.1093/pnasnexus/pgaf089
Advance access publication 12 March 2025
Research Report

Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation

Jiafu An [a], Difang Huang [b,*], Chen Lin [c,*] and Mingzhu Tai [c,*]

[a]Department of Real Estate and Construction, University of Hong Kong, Hong Kong SAR 999077, China
[b]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[c]Faculty of Business and Economics, University of Hong Kong, Hong Kong SAR 999077, China
*To whom correspondence should be addressed: Email: chenlin1@hku.hk (C.L.); Email: difang@amss.ac.cn (D.H.); Email: taimzh@hku.hk (M.T.)
Edited By Jay Van Bavel

Abstract

In traditional decision-making processes, social biases of human decision makers can lead to unequal economic outcomes for underrepresented social groups, such as women and racial/ethnic minorities (1–4). Recently, the growing popularity of large language model (LLM)-based AI signals a potential shift from human to AI-based decision-making. How would this transition affect the distributional outcomes across social groups? Here, we investigate the gender and racial biases of a number of commonly used LLMs, including OpenAI's GPT-3.5 Turbo and GPT-4o, Google's Gemini 1.5 Flash, Anthropic AI's Claude 3.5 Sonnet, and Meta's Llama 3-70b, in a high-stakes decision-making setting of assessing entry-level job candidates from diverse social groups. Instructing the models to score ∼361,000 resumes with randomized social identities, we find that the LLMs award higher assessment scores for female candidates with similar work experience, education, and skills, but lower scores for black male candidates with comparable qualifications. These biases may result in ∼1–3 percentage-point differences in hiring probabilities for otherwise similar candidates at a certain threshold and are consistent across various job positions and subsamples. Meanwhile, many models are biased against black male candidates. Our results indicate that LLM-based AI systems demonstrate significant biases, varying in terms of the directions and magnitudes across different social groups. Further research is needed to comprehend the root causes of these outcomes and develop strategies to minimize the remaining biases in AI systems. As AI-based decision-making tools are increasingly employed across diverse domains, our findings underscore the necessity of understanding and addressing the potential unequal outcomes to ensure equitable outcomes across social groups.

Significance Statement

Our research investigates how AI shifts decision-making processes traditionally influenced by human biases. We analyzed commonly used biases by scoring ∼361,000 resumes across diverse social identities in a high-stakes context—entry-level job hiring. Findings reveal a tendency to favor assessment scores for black male candidates with equivalent qualifications, impacting hiring probabilities by 1–3 percentage points. These biases were more pronounced. While this LLM-based AI system shows promise in reducing gender biases, it does not fully address racial biases. Our work highlights the need for ongoing research in AI, ensuring fair decision-making across all social groups as AI tools become more prevalent.

Introduction

Despite all the efforts by policy makers to promote equal opportunities in the labor market, underrepresented social groups, such as women and racial/ethnic minorities, still face significant gaps in terms of employment rates, income, recognition of attribution, etc. (1, 2). Research has shown that social biases of human decision makers in the recruiting process can be a key driver of such inequality in economic outcomes (3–6). Recently, with the rapid development and increasing popularity of LLM-based generative AIs, it is widely expected that a transition from human to AI-based decision-making can happen shortly (for example, link1 shows survey evidence that over half of firms are investing in AI-based recruiting). This brings up a new question: if generative AIs are used predominantly to replace human recruiters, how would this impact the recruiting outcomes across different social groups, and how would it change the gender/racial distribution in the labor market?

This question is particularly relevant given the intense scrutiny in recent years surrounding the intersection of AI and fairness in decision-making processes, particularly in word embedding and

Competing Interest: The authors declare no competing interests.
Received: June 4, 2024. **Accepted:** February 16, 2025

natural language processing (NLP) tasks (7–12). Since the algorithms are trained on human-generated data, AI agents can inherit and reflect similar behavioral biases as humans (13, 14). This aligns with broader findings in AI research, where language technologies have been shown to perpetuate—and in some cases, amplify—societal biases (8, 9, 15–18). Furthermore, some debiasing methods may mask biases rather than effectively eliminating them (19), reflecting concerns about the limitations of current strategies for mitigating bias in AI systems (20–23). More recently, an expanding body of literature has documented that LLM-based generative AI can exhibit significant social biases, many of which are identified in conversational contexts (24–26).

Meanwhile, scientists also show that large language models (LLMs) can potentially make economic decisions with great rationality (27), and generative agents based on fine-tuning techniques can potentially mitigate the biases of LLMs (28–30). Additionally, developers of LLMs have made enormous efforts to restrict the model from providing inappropriate opinions or judgments (31, 32). There is also anecdotal evidence, suggesting that popular LLMs are likely prodemocratic and left-libertarian orientated, which suggests that they likely value social equality and diversity (for example, see media reports at link2 and link3).

Overall, the direction and magnitude of LLM biases across different social identities need to be better identified and quantified empirically, particularly in the context of high-stakes, real-world decision-makings (33). Our study aims to address this gap by empirically assessing the impact of gender and racial identity on LLM-based evaluations of job candidates. While previous research has examined bias in AI-driven resume filtering using field data, finding that resume writing style and sociolinguistics can be sources of bias (21), our work extends this by utilizing a large-scale randomized experimental design. This approach allows us to quantitatively assess the potential social biases in several widely used, state-of-the-art LLMs.

In this paper, we quantitatively examine whether and how gender and racial identity affect the LLMs' assessment of job candidates. To control for the confounding influences of job candidates' other characteristics that are correlated with their social identity, we utilize an experimental research design that generates a large sample of resumes of entry-level job applicants with randomly drawn work experience, education, and skill sets from real-world distributions of those characteristics. Each resume is assigned a gender and racially distinctive name that reveals the job applicant's social identity. We then instruct a variety of commonly used generative AI models, including OpenAI's GPT-3.5 Turbo and GPT-4o, Google's Gemini 1.5 Flash, Anthropic AI's Claude 3.5 Sonnet, and Meta's Llama 3-70b, to assess each randomized resume using a score ranging from 0 to 100.

Using this randomized experimental research design, we find that, in aggregate, most of the LLMs seem to be "prosocial" when screening resumes: four out of five models we assessed yield significantly higher assessment scores on average for minority (female OR black) job candidates than for white male candidates with otherwise similar characteristics. However, their behaviors vary across different dimensions of social identity: the higher scores for minorities are driven by those for female candidates (both black and white females), but the scores for black males are in fact significantly lower than those for white male candidates by most of the models. The effects are robust across different job positions and applications from different states. They are also robust when we repeat the analyses many times using randomly drawn subsamples.

Our estimations show that the score differences are not only statistically but also economically significant. Assuming that job candidates with a score of 80 (out of 100) or above can be hired, and using GPT-3.5 Turbo's assessments as an example, the AI suggests a ∼35% probability of being hired for the overall sample. However, black (white) female candidates face a 1.7 (1.4) percentage-point greater probability of being hired. This translates to black (white) female candidates having a ∼5% (4%) higher hiring probability compared with the overall population. Meanwhile, black male candidates face a 1.4 percentage-point lower probability of being hired.

Generating resume scores using LLMs

In our experiment, we developed an automated pipeline using LLMs to assess a large sample of randomized resumes for 20 representative entry-level job positions and instruct the AI to generate an assessment score for each of those resumes. The pipeline first parses each job position and the corresponding resumes and then constructs a prompt for the AI by concatenating a brief, position-specific instruction with the resumes. The prompt is subsequently input into the model, generating scores ranging from 0 to 100 for each resume in a single pass. Further details and validation of the pipeline are given in the Supplementary Information.

To assess whether and how social identity influences LLMs' assessments of job candidates, we systematically evaluate its feedback using a randomized experimental design. Specifically, we generate a sample of ∼361,000 fictitious resumes with randomized, real-world applicant characteristics, including work experience, educational background, and skill sets (see Materials and methods and Supplementary Information). Each resume is randomly assigned a gender and racially distinctive name to reveal the applicant's social identity. This method follows the most recent labor economics literature (34–37). If an LLM is socially unbiased, the average score for resumes with different social identities should be similar in large samples, given that all the other characteristics are randomized. Instead, if there are differences in average assessments between different social groups, it suggests that the AI might have potential social biases.

For brevity, we start our analyses by first presenting the results from OpenAI's GPT-3.5 Turbo, one of the most cost-efficient and widely used LLMs recently. We then present the results from the other commonly used LLMs in the last section. Comparing the scores of our sample resumes across social groups in Fig. S2 (also see detailed results in Table S5), we provide preliminary evidence that GPT-3.5 Turbo is not neutral to social identities: it gives higher scores to female candidates but lower scores to black male candidates. We then conduct formal regression analyses with an extensive list of controls and fixed effects based on the specification in Eq. (1) in the Materials and methods section, although we note that the other characteristics should not confound the results as they are randomized in the sample resumes. Specifically, we control for resume characteristics such as the number of skills, work and educational experiences, the starting date of work and education, the number of high-level work experiences, the average duration of each work experience, the mean word count of work experience descriptions, the total length of education, and the highest educational degree. We also control for position fixed effects, state fixed effects, and fixed effects for the latest work title in each resume. The estimated regression coefficients are displayed in Fig. 1 (see detailed results in Table S6). In the first row, we show that the average score of minority (female OR black) candidates is ∼0.099 points higher ($P < 0.05$) than

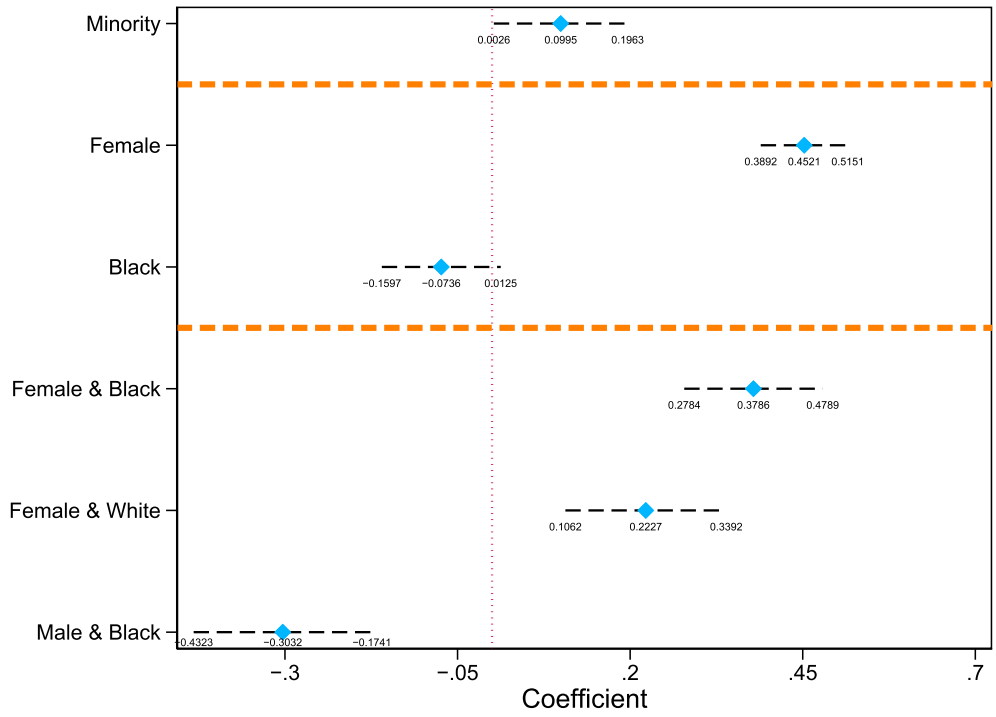

Fig. 1. Regression coefficients: GPT-3.5 Turbo score differences across social groups. This figure presents the regression coefficients that compare the score differences across different social groups by GPT-3.5 Turbo, following the specification in Eq. (1). The bold dashed lines divide the results into three sets depending on different regression specifications. In the first part, we compare the difference in scores between candidates from any minority group (female OR black) and white male candidates (the benchmark "majority" group). In the second part, the comparison is between all females versus all males or between all black versus all white males. In the third part, each minority group is compared with the white male group. The diamonds mark the estimated coefficients, and the dashed lines show the 95% CIs.

that of white male candidates (the benchmark "majority" group). More importantly, in the second and third rows, we find that the average score of female candidates is ∼0.452 points higher ($P <$ 0.001) than that of male candidates, while the average score of black candidates is ∼0.074 points lower ($P < 0.1$) than that of white candidates.

In the last three rows, we further decompose the social groups and demonstrate that, relative to white males, black females score 0.379 points higher, white females score 0.223 points higher, and black males score 0.303 points lower. All these differences are strongly significant, with P values <0.001. These findings suggest that GPT-3.5 Turbo may assign higher scores to both black and white female candidates but lower scores to black male candidates, conditional on otherwise similar characteristics.

Our findings on the intersection of gender and ethnicity align with the widely discussed intersectionality theory in social science (38–40). First, while female candidates receive significantly higher scores (0.452) and black candidates have slightly lower and statistically insignificant scores (−0.075), the intersectional effect for white female candidates (0.223) is smaller than a simple sum of these single-dimensional effects (0.452–0.075 = 0.377). This observation aligns with the first hypothesis proposed by Ghavami and Peplau (41), which suggests that gender-by-ethnic stereotypes contain unique elements that are not merely additive combinations of gender and ethnic stereotypes. This hypothesis is rooted in fundamental intersectionality theory (e.g. 42), which posits that group identities at the intersection of ethnicity and gender are distinct and cannot be fully understood by analyzing each identity dimension independently.

Second, we also find the racial gap to be very different for male candidates (−0.303) versus female candidates (0.379 − 0.223 = 0.156), and only the former is directionally aligned with the racial effect on the overall population (−0.075). This is consistent with (41)'s second hypothesis, which posits that stereotypes of ethnic groups are generally more aligned with stereotypes of the men within those groups than with stereotypes of women. This hypothesis originates from social dominance theory (43) (or the subordinate male target hypothesis), which argues that social stereotypes reflect broader social hierarchies, making racial stereotypes more pronounced among males (the dominant group) than among females.

Third, the bias observed against black male candidates—but not against black female candidates—may also support the gendered race prototype theory (39, 44). This theory proposes that race and sex stereotypes overlap significantly for black men, as black individuals are often stereotyped as more "manly." Given that LLMs exhibit a bias against male candidates, the gendered race prototype could exacerbate the bias against black male candidates.

Robustness checks

Figure 2 shows that the differences in scores between social groups remain robust across different subsamples. The detailed regression results are reported in Tables S7 and S14. First, we compare job positions with higher (above sample median) versus lower (below sample median) shares of female workers according to the national labor statistics from the US Bureau of Labor Statistics (BLS) (see Table S2 for detailed statistics for each position). As shown in Fig. 2a, the scores for female candidates are significantly greater than those for male candidates in both subsamples. Similarly, we compare job positions with higher versus lower shares of black workers in Fig. 2b and find that the scores for black male candidates are also significantly lower in

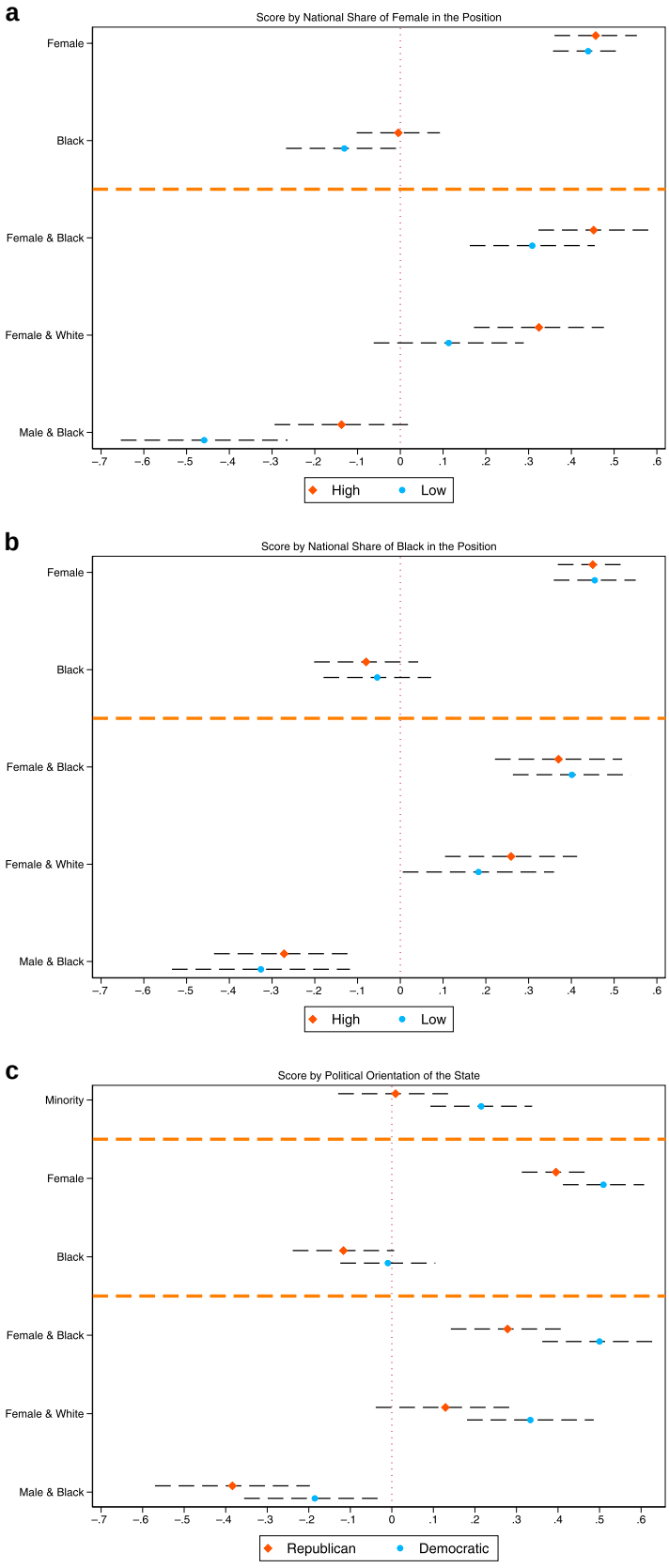

Fig. 2. GPT-3.5 Turbo score differences across social groups. This figure compares the regression coefficients, which represent the cross-social group score differences, across subgroups of candidates. In a), we compare candidates for job positions with high (above the sample median, in diamond) versus low (below the sample median, in circle) shares of female workers according to the national labor statistics from the BLS. In b), the comparison is based on the share of black workers. In c), we compare candidates whose states of residence (randomly assigned and reported in their resumes) are with different political orientations (based on the 2020 presidential election results). Of the five states in our sample (California, Florida, Georgia, New York, and Texas), we consider California and New York to be Democratic states (in diamond), while the other three are Republican (in blue circle). The dashed lines show 95% CIs.

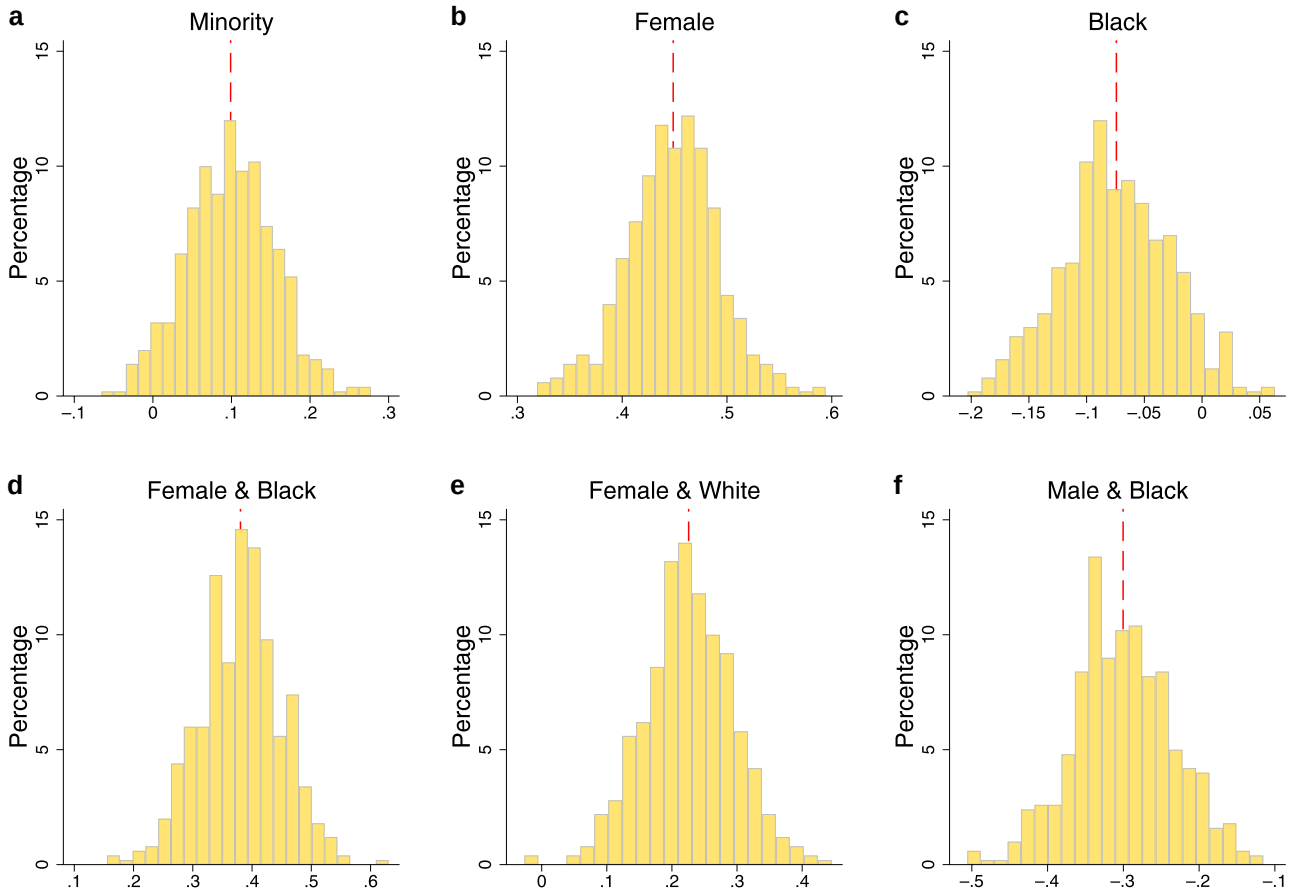

Fig. 3. Regression coefficients using random sampling, GPT-3.5 Turbo score. This figure plots the frequency distribution of the estimated regression coefficients under random sampling. We replicate the baseline regression by randomly drawing a 20% subsample each time. We repeat this random sampling 500 times and summarize the distribution of the estimated regression coefficients in this figure. The vertical dashed line represents the baseline regression coefficients for the full sample. Panel a) focuses on Minority, b) on Female, c) on Black, d) on Female and Black, e) on Female and White, and f) on Male and Black.

both subsamples. Figure S5 plots the coefficients for the other LLMs.

Next, we compare candidates from states with different political orientations. In our randomized sample, each fictitious resume includes a randomly assigned state of residence, drawn from the five most populous states in the entry-level job market: California, Florida, Georgia, New York, and Texas. We categorize these states into two groups based on their political orientation from the 2020 presidential election results—for example, California and New York as Democratic states, and Florida, Georgia, and Texas as Republican states. We then compare the score differences between the two subsamples of candidates, whose states of residence fall into one of these two groups. As shown in Fig. 2c, the profemale and antiblack male behavior of the LLM is found in both state groups, although the model seems to be more prosocial when assessing candidates from democratic states.

In an additional robustness check, we replicate our baseline regression analyses that compare the score differences across different social groups by randomly drawing a 20% subsample each time. We repeat this random sampling 500 times and summarize the distribution of the estimated regression coefficients in Fig. 3. Of the 500 random draws, the mean and median coefficient for black female candidates (which represents the difference in scores compared with white male candidates) are ~0.38, and this coefficient is >0.20, with a >99% probability (Fig. 3d). A robustly positive distribution of coefficients for white female candidates is shown in Fig. 3e, while the coefficient for black male candidates is robustly negative according to the distribution presented in Fig. 3f. This robustness check rules out the possibility that our findings are driven by a small number of observations in the sample. We repeat the same exercise for the other LLMs in Fig. S7.

In our baseline prompt to GPT-3.5 Turbo, we provide a very brief and simple instruction that only specifies the title of the job position. This simple instruction is supposed to simulate an untrained recruiter without any deep knowledge about the position. We also conduct two robustness checks that use representative, detailed job descriptions in the prompt. Specifically, for each position, we use a representative, real-world job posting downloaded from Indeed, a mainstream job search website, and add this job posting to the prompt to the GPT. We consider a version of the prompt with the equal opportunity claim and a version without this claim. As shown in Table S8, the results remain robust under both versions. More details are provided in the Supplementary Information.

Resume quality and LLM biases

Economic theories (45) suggest that human decision makers' social biases can demonstrate very different patterns when they are screening candidates with high versus low quality. Given this, it is also an interesting question to ask whether and how the LLMs' social biases can vary by resume quality. To answer this question, we first conduct quantile regressions to estimate

how candidate social identities affect their scores at the 10th, 25th, 50th, 75th, and 90th percentiles of sample resume distribution. As shown in Fig. S3 and Table S9, the bias of GPT-3.5 Turbo toward white female increases with resume quality, but its bias against black male decreases with quality. Table S15 presents the results for other LLMs. In particular, we observe a much stronger bias toward white female candidates with top 10% resume scores (i.e. those in the 90% percentile), but the strongest bias against black male candidates is found in the bottom 10%.

We supplement the quantile regression analyses by further conducting a linear regression analysis that directly interacts candidate social identities with a resume quality indicator that is predicted by all other observable resume characteristics. The results are presented in Tables S10 and S16. Lastly, we also compare in Tables S11 and S17 (also plotted in Fig. S6) by what extent the LLMs make use of information revealed from observable resume characteristics other than social identities, and how the extent of this information usage varies across social groups. We find that the LLM makes usage of quality-relevant information in the resumes by a similar extent regardless of the candidate's social identity. This suggests that LLMs may have potentially addressed the "attention discrimination" that can occur from decision-making by humans that have constrained capacity of processing information.

Real consequences on the probability of hiring

We further explore the real consequences of these documented differences across social groups, with a focus on examining the extent to which a candidate's social identity influences his/her probability of being hired. Specifically, we assume that a job candidate can be hired if he/she receives a score from the LLM that is equal to or above a certain threshold. Based on this simple criterion, we can estimate the probability of hiring for each social group under different score thresholds. We consider thresholds of 60, 65, 70, 75, 80, and 85, given that the most frequent scores assigned by the LLMs are integer multiples of five (e.g. see Fig. S1 for the distribution of the scores assigned by GPT-3.5 Turbo). Under each threshold, we analyze a linear probability model using a similar specification as in Eq. (1). The dependent variable is a dummy indicator of whether a candidate can be hired under that corresponding threshold, and the key explanatory variable is the corresponding social identity.

Taking the results from GPT-3.5 Turbo as an example, Fig. 4 shows that score differences among candidates from different social groups can lead to nonnegligible differences in their probability of being hired. For example, with an 80-point cutoff, which implies an average hiring probability of ~35%, black female applicants could expect a 1.7 percentage-point greater probability of being hired than white male applicants, while white female applicants have a 1.4 percentage-point higher probability. Given that there are 11.3 million black women and 58.4 million white women in the US labor force in 2023 (source: US BLS), our estimates suggest that over 190,000 black women and nearly 820,000 white women could see improved job opportunities if all employers were to adopt LLM-based AI tools for resume filtering, assuming the social biases we identified apply to all job positions. Intriguingly, black male applicants face a 1.4 percentage-point lower probability of being hired than their white male peers with otherwise similar characteristics. With 10.5 million black men in the labor force, this translates to nearly 150,000 jobs negatively affected for this group. These findings highlight the importance of understanding and addressing the disparities that arise from the use of LLM assessment scores in hiring decisions. We repeat the exercise for the other four models, and the results are presented in Fig. S8.

Results from other commonly used LLMs

So far, our analyses are based on the scoring by OpenAI's GPT-3.5 Turbo, one of the most widely used and cost-efficient LLMs especially during the earlier period when LLM-based AIs were introduced to the general public. More recently, a number of other models have been introduced and become more and more commonly used in many sectors. In this section, we further examine the potential social biases of four other models that are now prevailing in the market, including (i) OpenAI's GPT-4o, (ii) Google's Gemini 1.5 Flash, (iii) Anthropic AI's Claude 3.5 Sonnet, and (iv) Meta's Llama 3-70b. All four models are the very recent versions released by their developers as of the time we conduct our analyses (July 2024). The score distribution for each of these four additional models is reported in Figs. S4 and S5 and Table S12.

Before presenting the detailed results from the additional models, we compare the key features for the five prominent models in Table S20. As shown in the table, these models differ across several technical dimensions, including parameter size (e.g. ~70 billion for Llama 3-70B versus more than 200 billion for Gemini 1.5 Flash), context window size (e.g. about 2k tokens for Llama 3-70b versus 100k for Claude 3.5 Sonnet), training data sources, and typical use cases. For example, GPT-4o is generally suited for conversational agents, Gemini 1.5 Flash excels at multimodal content creation, Claude 3.5 Sonnet is optimized for professional coding, and Llama 3-70b, being open-source, is more suitable for private, customized deployment and research.

All these models employ a range of debiasing strategies—including reinforcement learning from human feedback, adversarial training, and fairness constrains—to address ethnical and fairness concerns. Despite these variations in debiasing strategies, the intersectional gender–racial biases observed under our setting framework remain qualitatively consistent and quantitatively similar across models, suggesting the existence of systematic mechanisms driving these biases that have yet to be fully addressed by current debiasing techniques. Interestingly, although Claude is generally marketed as being optimized for safety and ethics, our tests reveal that its gender and racial biases in resume evaluations are comparable to those of other models. This underscores the importance of independently auditing the ethical performance of AI tools using carefully designed, close-to-real-world research experiments, rather than relying solely on developers' public claims.

Moreover, it is important to note that open-source LLMs (e.g. Llama 3-70b used in our study) can differ significantly from proprietary models (e.g. the other models examined) in various aspects. In Table S21, we outline the key differences between these two types of models. Notably, open-source LLMs are significantly less costly to access, offering users full control over their development and customization, while allowing data to remain in a local, private environment. Perhaps more importantly, open-source LLMs provide much greater transparency compared with proprietary models. With access to the sources code, users can independently audit training data and algorithms, facilitating the development of ethical solutions and fostering trust. These advantages suggest that open-source LLMs are likely to see broader adoption in human resource (HR)-related business practices, such as the resume screening tasks examined in this study. However,

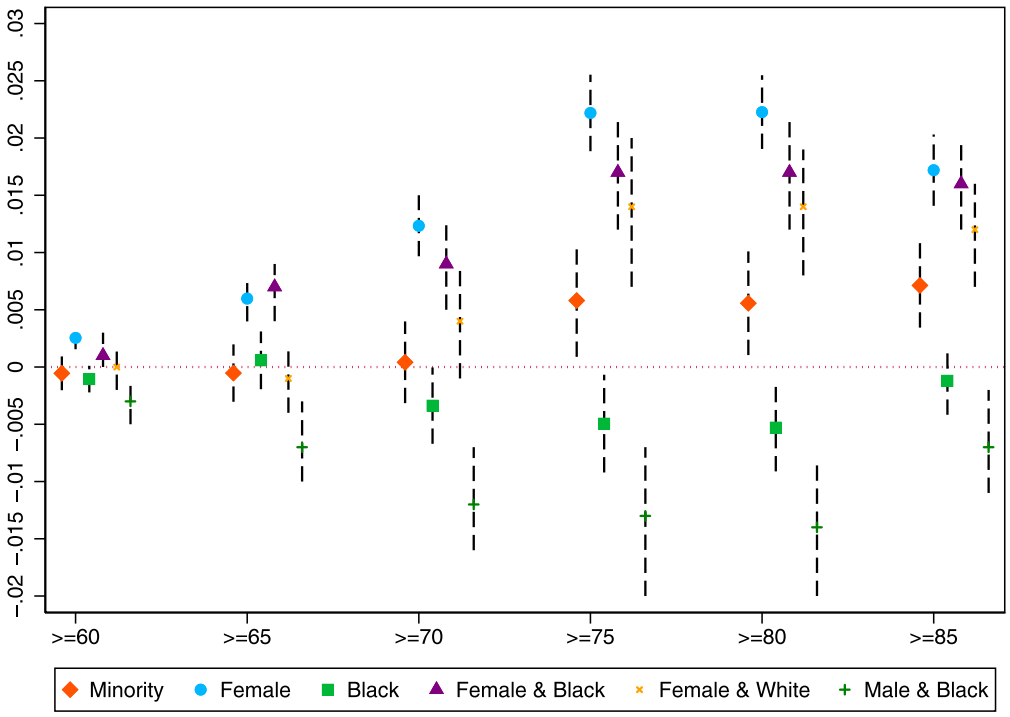

Fig. 4. Estimated differences in the probability of being hired, based on GPT-3.5 Turbo score. This figure presents the estimated probability of being hired by a certain minority social group relative to the benchmark group (minority versus nonminority, female versus male, and the three intersection groups versus white male). Under each score threshold, we assume that a candidate with a score equal to or above the threshold could be hired, and we regress the corresponding probability of hiring for each candidate on the social identity using the same regression specification as in Eq. (1). The regression coefficients reported in this figure are also presented in panel (a) of Table S18. The dashed gray lines show 95% CIs.

our tests show no obvious differences in the resume scoring outcomes between the open-source model (Llama 3-70b) and the proprietary models. This finding underscores the need for more customized, debiasing-oriented fine-tuning if practitioners aim to use open-source AI tools to ensure fair and equitable recruiting decisions.

Next, we discuss more details of our results. In Fig. 5, we compare how candidate social identities affect the resume assessments by each of these four models. The regression results are presented in Table S13. Consistent with our baseline finding from GPT-3.5 Turbo, all these other four models also strongly favor female candidates when scoring their resumes, even conditioning on otherwise similar characteristics. Such profemale preferences are strong for both black and white candidates. For example, on average, Gemini 1.5 Flash gives 0.956-point higher score to a black female candidate than an otherwise similar white male, and its score premium for a white female candidate is about 0.827. Meanwhile, similar to GPT-3.5 Turbo, most of these other models also demonstrate a significant bias against black male candidates. The only exception is Llama 3-70b: although its scoring for black male candidates is still lower than that for white male, the difference is not statistically significant. In aggregate, as the magnitudes of the profemale and antimale biases for the black candidates vary, the bias for the overall black group turns out to be negative by GPT-4o and Claude 3.5 Sonnet, insignificant by Gemini 1.5 Flash, and positive by Llama 3-70b.

We also replicate all other analyses for these four additional models. The results, as presented in Figs. S6–S8 and Tables S14–S18, suggest that the social biases by all models remain consistent under all our robustness checks, although the ways those biases are affected by resume quality can vary across different LLM models.

Assessing more ethnicity groups

So far, our analyses regarding LLMs' social biases in ethnicity focus on the comparison between black versus white candidates. We further extend our analyses to include two more ethnicity groups: Asian and Hispanic. As shown in Fig. S9a and Table S19, GPT-3.5 Turbo has very strong bias against Asian and Hispanic candidates, regardless of their gender. These biases are partially cured in the more recent models (see Fig. S9b–e and Table S19): scoring for Asian males is no longer significantly different from that for white males by GPT-4o and Gemini 1.5 Flash, and the bias against Hispanic males is eliminated by Llama 3-70b. In addition, these models may "overcure" the discrimination against Asian and Hispanic for quite a few subgroups: all four models have significantly higher scoring for Asian and Hispanic female candidates than white male, and GPT-4o and Gemini 1.5 Flash also score Hispanic male candidates higher than white male. These findings suggest that developers of the more recent models are likely aware of the bias against certain ethnicity groups and have addressed the issue, but they seem to have overreacted and pushed the bias toward the opposite direction.

Discussion

This paper sheds light to a crucial question over the development and widespread application of LLMs: if LLM-based generative AIs are used predominantly used to replace humans in high-stakes decision-making processes, such as resume filtering by job recruiters, how would this affect the recruiting outcomes across different social groups, and what impact could it have on gender and racial distribution in the labor market? We consider a setting in which social biases of human decision makers have long been

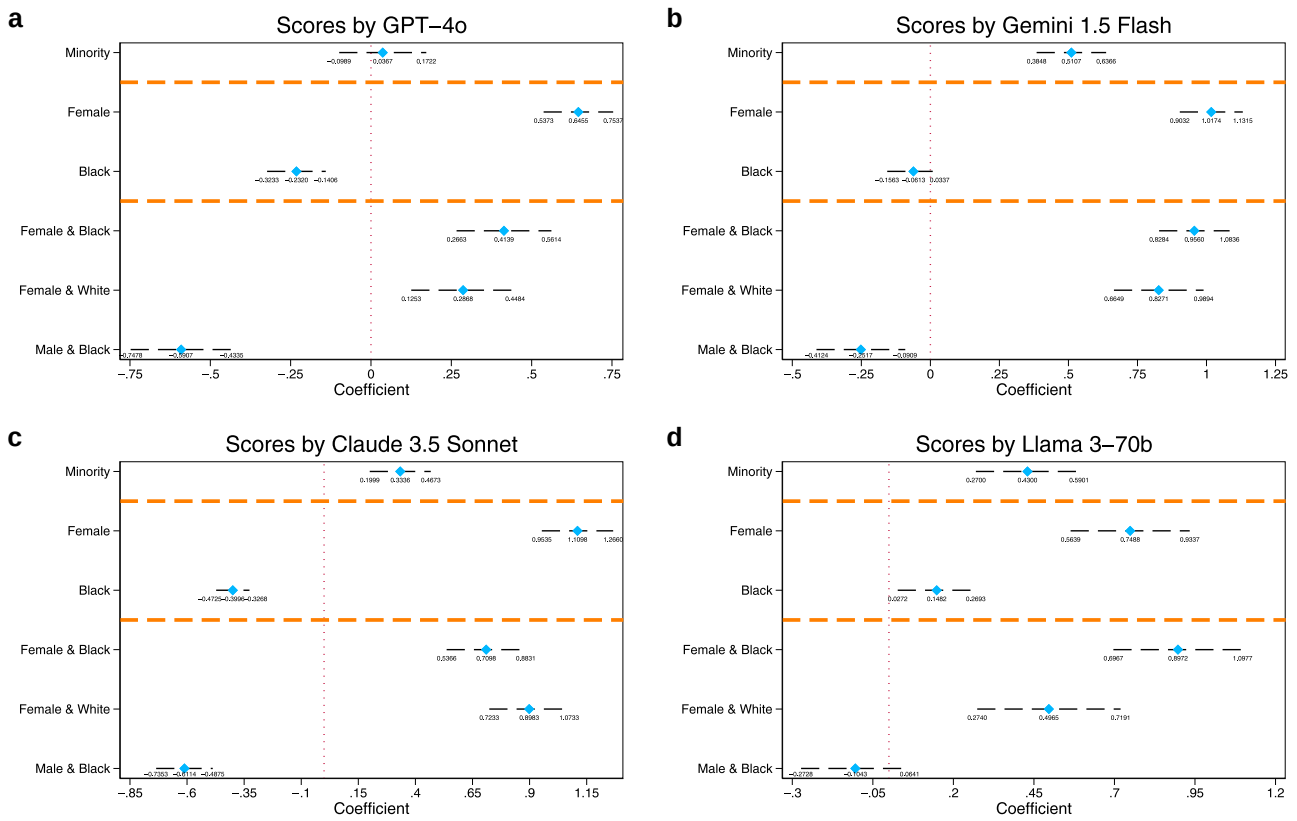

Fig. 5. Regression coefficients: score differences across social groups, four other models. This figure presents the regression coefficients that compare the other four models' score differences across different social groups, following the specification in Eq. (1). a)–d) present the results for GPT-4o, Gemini 1.5 Flash, Claude 3.5 Sonnet, and Llama 3-70b, respectively. The bold dashed lines divide the results into three sets depending on different regression specifications. In the first part, we compare the difference in scores between candidates from any minority group (female OR black) and white male candidates (the benchmark "majority" group). In the second part, the comparison is between all females versus all males or between all black versus all white males. In the third part, each minority group is compared with the white male group. The diamonds mark the estimated coefficients, and the dashed lines show the 95% CIs.

documented by social scientists: HR screening of job candidates and hiring decisions. Employing a randomized experimental research design that generates a large sample of resumes with randomly drawn social identity but real-world characteristics in other dimensions (e.g. work experiences, education, skill sets, etc.), we instruct a number of commonly used LLMs, including OpenAI's GPT-3.5 Turbo and GPT-4o, Google's Gemini 1.5 Flash, Anthropic AI's Claude 3.5 Sonnet, and Meta's Llama 3-70b, to assess each of those randomized resumes and score it. Our results show that these LLMs exhibit bias in favor of both black and white female candidates, while most display bias against black male candidates. These patterns remain consistent across various robustness checks. The biases at other minority ethnicity groups, such as Asian or Hispanic, vary across different models. At a certain threshold, its biases can lead to a 1–3 percentage-point difference in the probability that candidates from a certain social group can be hired, compared with their otherwise similar peers.

This study is related to the growing body of literature that examines the potential social biases of LLMs (8–30, 33). While most existing research focuses on biased or "toxic" expressions in AI narratives or on methodologies for fine-tuning these biases, we investigate a more realistic scenario, allowing us to quantitatively infer the social stance of LLMs in high-stakes decision-making processes. Our work builds upon and extends prior research on bias in AI-driven resume filtering (21) by focusing on state-of-the-art LLMs and using a large-scale randomized experimental design.

More specifically, our findings contribute to and complement the growing body of literature examining biases in AI-driven recruitment processes. Many studies (e.g. 46, 47) have raised concerns that leveraging LLMs in hiring can perpetuate biases through training data and flawed algorithmic designs, despite their acknowledgment of the potential benefits of AI-based recruitment, such as improved recruitment quality, reduced costs, and enhanced efficiency. Additionally, HR officers have expressed concerns regarding the potential for AI systems to introduce biases into hiring decisions (48). Experimental studies provide empirical support for these concerns. Glazko et al. (49), for example, find that GPT-4 exhibits biases against resumes that explicitly indicate the applicant has a disability. Other scholars document that LLMs also exhibit religious and political biases in hiring (50).

Of particular relevance to our study is the emerging body of literature examining gender- and race-related biases in LLMs used for recruitment (e.g. 51–55), many relying on smaller, less representative datasets or older LLMs and word-embedding algorithms. More importantly, most of those studies focus on gender and race biases independently, without considering the intersectionality across social identities. We contribute to these works by explicitly evaluating how LLMs exhibit very different bias patterns across the intersections between gender and race. From this standpoint, we also contribute to the recent literature on detecting intersectional stereotypes in AI. While most of the existing studies focus on methods for identifying such stereotypes in word embeddings or probing language models (e.g. 56–61), our approach contributes

to this literature by employing a directly quantifiable, close-to-real-world experimental setting. This methodology allows us to uncover intersectional social biases in a way that is more grounded in practical applications.

Closest to our study are (61–63). Wilson and Caliskan (61), for instance, analyze 500 publicly available resumes and identify statistically significant model bias against black men. However, their study utilizes the older LLM Mistral-7B-v0.1 (or its fine-tuned variants), which limits the applicability of their findings to more current models that are widely used in the market. Building on this, Armstrong et al. (62) employ a more updated model, OpenAI's GPT-3.5, in a similar study. Examining the intersection of gender and race, they observe small but statistically significant differences in ratings. Another related study is by Bloomberg (63), which further finds that the intersectional biases of GPT-3.5 can vary by job. Our research extends and complements these earlier efforts by analyzing a much larger dataset of over 36,000 resumes and utilizing five of the most recent and widely used LLMs from four different developers. Our findings provide a more nuanced understanding of biases in these models, revealing significant patterns of bias related to gender, race, and their intersection. Interestingly, our results show that while LLMs display favorable biases toward some underrepresented groups (e.g. women), they continue to exhibit pronounced biases against others (e.g. black men). This complexity underscores the need for comprehensive, multidimensional approaches to studying and mitigating bias in AI systems, as highlighted in recent surveys (22). It also offers fresh insights into the interpretability of data-driven algorithmic decisions that have significant impacts on individuals' lives (33).

Our findings also align with broader concerns in the field about AI systems' potential to perpetuate or even amplify societal biases (11, 12). However, our results also paint a more nuanced picture, with the LLMs displaying bias in favor of some underrepresented groups (e.g. females) while exhibiting bias against others (e.g. black males). It also offers fresh insights into the interpretability of data-driven algorithmic decisions that have significant impacts on individuals' lives (33). Besides, our work has implications for ongoing efforts to debias AI systems (7, 20). The intricate pattern of biases we observed suggests that simplistic debiasing methods may be insufficient. Instead, more nuanced, context-aware approaches may be necessary. This supports recent calls for more holistic and intersectional strategies to address social biases in AI (23).

Our inferences also have important implications for practitioners and policy makers. As human capital is a key driver of productivity and economic growth (64–67), a well-functioning labor market that aligns diverse talents with roles that maximize returns is crucial (68, 69). Moreover, labor market outcomes and distributions significantly affect the effectiveness of economic policies (70, 71). With the rapid acceleration of LLM-based AI applications in recruitment process (e.g. a Gartner survey [link4] revealed that only 15% of the HR leaders have no plans to integrate generative AI to their HR processes), it is increasingly urgent to understand the direction and extent to which AI may influence social distribution within the labor market. In fact, there have been concerns that AI may even increase the risk of discrimination in recruiting (for example, see the report of link5). On the contrary, there is also anecdotal evidence that some models can sometimes be "too woke" regarding Diversity, Equality, and Inclusion issues (for example, see the discussion by Business Insider: link6). For policy makers, as the decision-making rule of the AI is a black box, they cannot directly judge from the algorithm whether and how the AI takes the candidates' social identities into its decision-making. Using a revealed-preference method, our study provides clear and quantifiable evidence to both parties regarding the social biases of generative AI from multiple dimensions.

Our key finding suggests a bias in favor of female job candidates by commonly used LLM models in the market, raising the important question of what potential sources drive this observed female "advantage." One possibility is that the models' training corpora—particularly the sizable portion derived from the internet—may reflect biased social values from a skewed subset of the population that is overrepresented online. Another potential source of bias stems from the later stages of model training, specifically during debiasing processes. For example, crowd workers (i.e. "annotators") hired by the developers may introduce biased human feedback during reinforcement learning, and developers' debiasing algorithms may be overly sensitive to certain social harms. Both channels have been discussed in the existing studies. For instance, Abdurahman et al. (72), Atari et al. (73), and Zewail et al. (74)suggest that both skewed training data and debiasing procedures can lead LLMs to favor Western, Educated, Industrialized, Rich, and Democratic social values. Similarly, Zhou and Sanfilippo (75) demonstrate that LLMs trained in different countries exhibit gender biases in varying ways and display differing levels of political correctness.

While much more careful future work is needed to clearly disentangle these two potential sources of LLM biases, we conjecture that the debiasing stage likely plays a more prominent role in generating the "female advantage" observed in this paper. Specifically, if this pattern were primarily driven by the training data, we would expect to see a similar bias *toward* female in both the pretrained base models and the human-feedback-tuned models. However, studies such as (76) and (77) have shown that social opinions by the pretrained base models are more aligned with the social values of internet contributors, while human-feedback-tuned models align more closely with the social values of crowd workers hired by the AI developers. Importantly, the former group (internet contributors) tend to be less educated and politically right leaning, making them less likely to favor female candidates in the job market. Indeed, pretrained models have been found to exhibit human-like gender biases *against* women (e.g. 78), whereas fine-tuned models such as GPT-4o demonstrate reduced less and can even perform as moral experts (79). Furthermore, Crockett and Messeri (80) argue that the internet-based training sets of LLMs are inherently limited to those who have the ability to speak online, overrepresenting younger individuals. Notably, survey evidence suggests that younger generations display more pronounced gender biases (e.g. see the report by Ipsos' annual International Women's Day: link7, and a report by Forbes: link8). Additionally, Chen et al. (81) found that OpenAI's GPT-4 became less willing to answer sensitive questions about social biases between March 2023 and June 2023. Since OpenAI does not modify the training dataset of a released model, this increased cautiousness and sensitivity to harmful content likely stems from iterative debiasing methods applied postrelease. In summary, these findings collectively suggest that the higher scores assigned to female job candidates by LLMs are more likely attributable to the debiasing stage. This may result from biased human feedback during data filtering or reinforcement learning, or from tuning algorithms that are oversensitive to certain social harms.

This study has several limitations. First, all LLMs are fundamentally a product of their training corpora (82, 83) and, as such, can only reflect the values and biases embedded in the underlying population (84). Consequently, our findings are likely not representative of non-US audiences. Since English-based content is disproportionately represented in the training data of the LLMs employed in our

experiments, and given that our experimental context is US-centric, our results inherently carry both cultural and linguistic limitations (85). To provide further context, recent research by Tao et al. (86) rigorously compared the cultural values of widely used LLMs, including GPT-4 and GPT-3.5 Turbo, against nationally representative survey data from the World Values Survey. They found that LLMs predominantly reflect cultural values akin to English-speaking and Protestant European countries. While cultural prompting and updates in newer models may reduce this bias to some extent, it does not entirely eliminate it. With these considerations in mind, we caution readers that our findings are relevant only to an English-speaking context, specifically within the United States, and may not extend to other linguistic or cultural settings. We have incorporated a discussion of these limitations in the revised manuscript to reflect this important caveat.

Second, in our research design, gender and racial identities are inferred solely from the fictitious names of job candidates. As a result, our study cannot assess potential LLM biases related to other factors, such as variations in work experience, educational attainment, and skill sets that may arise from different social backgrounds.

Third, as we focus on the screening of job applications for entry-level positions, for which the job requirements and candidates' backgrounds are much simpler and more homogeneous than those of higher-level positions, our results may have limited external validity and may not necessarily extend to recruiting decisions for high-skill jobs or management positions. Relatedly, to make a clean inference regarding the role of social identities in affecting AI decision-making, our sample resumes are designed to follow a standardized format that only includes information about a candidate's work experiences, education, and skill sets. Although all these characteristics are drawn from representative, real-world distributions, the sample results are likely simpler and more homogeneous than the actual ones.

Fourth, our baseline prompts to the AIs are designed to include only brief information about the job, and we do not conduct any fine-tuning. Thus, the assessments by AIs under our research design can be less sophisticated than those by professional and well-trained human recruiters. Lastly, our experiment is only conducted at one snapshot of time. As AIs continue to evolve and learn, the way they consider social identities in decision-making can also evolve. It is also important to note that this research does not develop any novel AI methods, nor does it propose solutions to address AI biases. Instead, our focus is on assessing the social biases present in existing AI tools, which provides valuable insights for researchers in their future AI development endeavors.

A potential extension to future work is to better understand how dynamic changes in training sets over time may change the LLMs' "social" stance. As the extent of left versus right leaning in our society is dynamically changing, the information fed to the model and thus the model's social stance may also vary over time. If the relationship between social norms and the social stance of AIs can be dynamically mapped, this would shed more light on the social consequences of applying LLMs in real-world, high-stakes decision-making.

Materials and methods

Experimental design

A key challenge in identifying social biases in the recruiting process (and many other real-world activities) is that candidates' social identities can be correlated with other features essential to recruiting decisions (23). For instance, even when a certain social group appears to have a lower probability of being employed on average, it is difficult to disentangle whether this is due to recruiters' bias against the group identity or because the candidates' skills are less well matched with the specific job.

To address this challenge, we employ an experimental research design that generates a large sample of fictitious resumes with randomized yet real-world applicant characteristics, including educational background, work experience, and skill sets. Each resume is assigned a gender and racially distinctive name to reveal the job applicant's social identity. This method has been widely adopted in recent economic research to detect discrimination in the labor market (34–37). The underlying concept is straightforward. By considering two large samples of such resumes with different social identities and given that job applicants' characteristics are randomized, the law of large numbers implies that the average quality (and thus the probability of being employed) of the two groups of applicants should be similar if the recruiters who screen the resumes do not exhibit social biases. However, if we observe differences in the recruiters' average assessments (and recruiting decisions) for these two groups of applicants, this outcome would indicate the presence of social biases among the decision makers. We provide the detailed experimental procedures as follows.

Resume creation

To construct a sample of randomized resumes, we first need a set of representative job positions and real-world job applicants targeting those positions. We begin with a sample of randomly collected entry-level job postings from the mainstream job search websites Indeed and Snagajob. We apply a number of filtering criteria to ensure that those postings were representative (Table S1 presents the distribution of the filtered job postings by industry, state, and position). From this sample of job postings, we select 20 representative job positions with high/medium/low shares of female or black workers based on the corresponding position's national labor statistics reported by the US BLS. The gender and racial statistics for those 20 selected job positions are reported in Table S2. More details are discussed in Supplementary Note S1.

For each position, we randomly download 1,250 real-world resumes from the job search websites. This gives us a total of 25,000 resumes from which we extract key characteristics, including (i) work experience (e.g. job title, employer name, textual description of job responsibilities, and length of employment for each experience), (ii) education experience (e.g. degree title, school name, and length of education for each experience), (iii) skills (key words), and (iv) the state of residence. We summarize a number of key features of these characteristics in Table S3.

Based on the distributions of those real-world characteristics, we generate a large sample of randomized fictitious resumes. For computational efficiency, we focus on the five states with the highest number of real-world job applicants: California, Florida, Georgia, New York, and Texas. For each position-state pair, we generate a sample of fictitious resumes 40 times larger than the number of real-world resumes we downloaded. For each fictitious resume, characteristics such as work experience, education, and skills are each randomly drawn from the corresponding real-world distributions we obtained earlier. We then randomly assign a gender and racial distinctive name to each resume. In summary, we generated a total of ~361,000 randomized resumes for our baseline analyses, with an equal distribution among the four gender and racial combinations (black females, white females, black males, and white males). In our extended

analyses, we also further expand the sample to include two more ethnicity groups: Asian and Hispanic. More details are given in Supplementary Note S2.

LLM scoring

We employed a number of most commonly used LLMs to evaluate a large sample of randomized resumes. The LLMs we use include: (i) OpenAI's GPT-3.5 Turbo, (ii) OpenAI's GPT-4o, (iii) Google's Gemini 1.5 Flash, (iv) Anthropic AI's Claude 3.5 Sonnet, and (v) Meta's Llama 3-70b. A summarization of each model's basic information is reported in Table S20. We leveraged each LLM's application programming interface to facilitate rapid and scalable interactions, enabling us to treat each LLM as a distinct entity and conduct the resume-scoring task repeatedly and independently, thereby enhancing the reliability of our findings.

The process is based on an automated pipeline that parses the 20 entry-level job positions and feeds the randomized resumes to each corresponding position. For each resume, each LLM is instructed to provide a score on a 0–100 scale. Resume-scoring task sessions were conducted separately for each LLM at a temperature setting of 0.0 (0 being the most concentrated and 2 being the most random, with the exception of the Llama-3-70b, which was set to 0.01 due to technical constraints), with all test items delivered in a single chat session. This consistent temperature setting ensures a fair comparison across all LLMs in our sample and strikes a balance between deterministic and stochastic model outputs. The detailed settings and prompts for the generative AI models are given in Supplementary Note S3. Figure S4 shows the distribution of the scores by GPT-3.5 Turbo.

Regression model

To study the relationship between an applicant's social identity (i.e. whether the applicant belongs to a minority or majority group) and the LLM score for the resume, we estimate Eq. (1) using a regression specification:

$$Y_{ij} = \alpha_j + \gamma_k + \beta \cdot Identity_{ij} + \theta \mathbf{X}_{ij} + \varepsilon_{ij}, \quad (1)$$

where the dependent variable Y_{it} is the LLM score for the resume of applicant i, sent to job posing j. $Identity_{it}$ is a set of variables indicating the candidate's social identity, which can be an indicator that equals one if the candidate falls into any minority group (female OR black; in this case, the omitted group is white male), two separate indicators for female (relative to male) and black (relative to white), respectively, or a set of indicators for black female, white female, and black male, respectively (also with white male to be the omitted group). $\mathbf{X}_{it}$ is a vector of other resume characteristics, including the number of skills, work and educational experiences, the starting date of work and education, the number of high-level work experiences, the average duration of each work experience, the mean word count of work experience descriptions, the total length of education, and the highest educational degree. In all analyses, we exclude outlier observations with these characteristics falling in the top or bottom 1% of the sample distribution. The summary statistics of these characteristics are reported in Table S4. We also control for job position fixed effects, candidate's state fixed effects, and fixed effects for the title of the applicant's last job. SEs are adjusted for potential heteroscedasticity by position. In this specification, our null hypothesis states that the LLM has no social bias when screening job candidates. If the coefficient β is significantly different from zero, we can reject this hypothesis.

We also conduct quantile regressions to compare how the LLMs' social biases vary by resume quality. Using a similar specification as in the baseline analyses, we estimate how candidate social identities affect their scores at the 10, 25, 50, 75, or 90% percentile of sample resume distribution. To do so, we use a residual score that controls for the job position and state of residence fixed effects as the dependent variable in the quantile regression model. We do not control for any other resume characteristics in these regressions because those characteristics could be directly related to the resume quality, which is exactly what we would like to compare in this analysis.

We supplement the quantile regression analyses by further conducting a linear regression analysis that directly interacts candidate social identities with a resume quality measure. We construct the resume quality measure by the following steps. First, we regress the resume score on all the observable resume characteristics other than social identity (i.e. the same set of characteristics we used as control variables in the baseline regression) in a univariate regression, and estimate the regression coefficient for each characteristic. Second, we compute the predicted value of resume score using a linear function based on those estimated regression coefficients and the corresponding characteristics for each resume. This predicted score value can be used as the resume quality measure, as a higher predicted value reflects a combination of stronger resume characteristics. If the interaction between this resume quality measure and a social identity indicator is positive, it suggests that the LLM's bias toward (against) a candidate with that social identity increases (decreases) with their quality.

Lastly, we also compare how the LLM makes use of information revealed from observable resume characteristics for candidates from different social groups. For each social group, we regress the sample resume scores generated by the LLM on all resume characteristics other than social identity (i.e. the same set of characteristics we used as control variables in the baseline regression), and we report the corresponding R^2. A higher R^2 suggests that for this subsample, the LLM better utilizes the observable resume characteristics when scoring.

Acknowledgments

The authors appreciate the excellent research assistant work provided by Ting Wang at the University of Hong Kong.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

C.L. and M.T. acknowledge funding support from the major program of the National Natural Science Foundation of China (no. 72193284021). J.A. acknowledges funding support from the University of Hong Kong Start-up Fund for New Professoriate Staff. D.H. acknowledges financial support from the National Natural Science Foundation of China (grant no. T2293771).

Author Contributions

J.A. was involved in conceptualization and writing—review and editing. D.H. was involved in conceptualization, supervision, and writing—review and editing. C.L. contributed to project

administration. M.T. was involved in conceptualization, formal analysis, supervision, and writing—original draft.

Preprints

This manuscript was posted on a preprint: http://arxiv.org/abs/2403.1528.

Data Availability

Data and code for replication are deposited in the public repository at here: https://osf.io/4dahv/.

References

1 André C, Causa O, Soldani E, Sutherland D, Unsal F. 2023. *Promoting gender equality to strengthen economic growth and resilience*. Paris: OECD Economics Department.
2 Sarsons H. 2017. Recognition for group work: gender differences in academia. *Am Econ Rev*. 107:141–145.
3 Charles KK, Guryan J. 2008. Prejudice and wages: an empirical assessment of Becker's the economics of discrimination. *J. Pol. Econ*. 116:773–809.
4 Bertrand M, Duflo E. 2017. Field experiments on discrimination. In: *Handbook of economic field experiments*. Vol. 1. Amsterdam: North-Holland. p. 309–393.
5 Becker GS. 1957. *The economics of discrimination*. Chicago: University of Chicago Press.
6 Aigner DJ, Cain GG. 1977. Statistical theories of discrimination in labor markets. *Ind Labor Relat Rev*. 30:175–187.
7 Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Adv Neural Inf Process Syst*. 29:1–9.
8 Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*. 356(6334):183–186.
9 Garg N, Schiebinger L, Jurafsky D, Zou J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci U S A*. 115(16):E3635–E3644.
10 Blodgett SL, Barocas S, Daumé H III, Wallach H. 2020. Language (technology) is power: A critical survey of "bias" in NLP, *arXiv, arXiv:2005.14050*, preprint: not peer reviewed.
11 Bender EM, Gebru T, McMillan-Major A, Shmitchell S. 2021. On the dangers of stochastic parrots: can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York (NY): Association for Computing Machinery. p. 610–623.
12 Crawford K. 2021. *The atlas of AI: power, politics, and the planetary costs of artificial intelligence*. New Haven (CT): Yale University Press.
13 Horton JJ. 2023. Large language models as simulated economic agents: what can we learn from homo silicus?, *arXiv, arXiv:2301.07543*, preprint: not peer reviewed.
14 Park JS, *et al.* 2023. Generative agents: interactive simulacra of human behavior, *arXiv, arXiv:2304.03442*, preprint: not peer reviewed.
15 Ouyang L, *et al.* 2022. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst*. 35:27730–27744.
16 Touvron H, *et al.* 2023. Llama: open and efficient foundation language models, *arXiv, arXiv:2302.13971*, preprint: not peer reviewed.
17 Zhang A, Yuksekgonul M, Guild J, Zou J, Wu J. 16 November 2023. ChatGPT exhibits gender and racial biases in acute coronary syndrome management. *medRxiv 23298525* https://doi.org/10.1101/2023.11.14.23298525, preprint: not peer reviewed.
18 Equality Now. ChatGPT-4 reinforces sexist stereotypes by stating a girl cannot "handle technicalities and numbers" in engineering. [accessed 2023 Mar 23]. https://www.equalitynow.org/news_and_insights/chatgpt-4-reinforces-sexist-stereotypes/.
19 Gonen H, Goldberg Y. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, *arXiv, arXiv:1903.03862*, preprint: not peer reviewed.
20 Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. 2021. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 54(6):1–35.
21 Deshpande KV, Pan S, Foulds JR. 2020. Mitigating demographic bias in AI-based resume filtering. In Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization. New York: Association for Computing Machinery. p. 268–275.
22 Gallegos IO, *et al.* 2024. Bias and fairness in large language models: a survey. *Comput Linguist*. 50:1097–1179.
23 Hanna A, Denton E, Smart A, Smith-Loud J. 2020. Towards a critical race methodology in algorithmic fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York: Association of Computing Machinery. p. 501–512.
24 Lee MH, Montgomery JM, Lai CK. 2024. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In The 2024 ACM Conference on Fairness, Accountability, and Transparency. New York: Association of Computing Machinery. p. 1321–1340.
25 Kotek H, Dockum R, Sun D. 2023. Gender bias and stereotypes in large language models. In Proceedings of the ACM Collective Intelligence Conference. New York: Association of Computing Machinery. p. 12–24.
26 Bai X, Wang A, Sucholutsky I, Griffiths TL. 2024. Measuring implicit bias in explicitly unbiased large language models, *arXiv, arXiv:2402.04105*, preprint: not peer reviewed.
27 Chen Y, Liu TX, Shan Y, Zhong S. 2023. The emergence of economic rationality of GPT. *Proc Natl Acad Sci U S A*. 120:e2316205120.
28 Glaese A, *et al.* 2022. Improving alignment of dialog agents via targeted human judgments, *arXiv, arXiv:2209.14375*, preprint: not peer reviewed.
29 Bai Y, *et al.* 2022. Constitutional AI: harmlessness from AI feedback, *arXiv, arXiv:2212.08073*, preprint: not peer reviewed.
30 Savage N. 2023. How robots can learn to follow a moral code. *Nature*. 7729:59–64.
31 OpenAI. How should AI systems behave, and who should decide? [accessed 2023 Feb 16]. https://openai.com/blog/how-should-ai-systems-behave.
32 OpenAI. Snapshot of ChatGPT model behavior guidelines. 2022. [accessed 2025 Jan 5]. https://cdn.openai.com/snapshot-of-chatgpt-model-behavior-guidelines.pdf.
33 Stoyanovich J, Van Bavel JJ, West TV. 2020. The imperative of interpretable machines. *Nat Mach Intell*. 2(4):197–199.
34 Agan A, Starr S. 2018. Ban the box, criminal records, and racial discrimination: a field experiment. *Q J Econ*. 133:191–235.
35 Neumark D, Burn I, Button P. 2019. Is it harder for older workers to find jobs? New and improved evidence from a field experiment. *J. Pol. Econ*. 127:922–970.

36 Kline P, Rose EK, Walters CR. 2022. Systemic discrimination among large US employers. *Q J Econ.* 137:1963–2036.
37 Bertrand M, Mullainathan S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am Econ Rev*. 94:991–1013.
38 Crenshaw K. 2013. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In: *Feminist Legal Theories*. London: Routledge. p. 23–51.
39 Browne I, Misra J. 2003. The intersection of gender and race in the labor market. *Annu Rev Sociol.* 29(1):487–513.
40 Rosette AS, de Leon RP, Koval CZ, Harrison DA. 2018. Intersectionality: connecting experiences of gender with race at work. *Res Organ Behav*. 38:1–22.
41 Ghavami N, Peplau LA. 2013. An intersectional analysis of gender and ethnic stereotypes: testing three hypotheses. *Psychol Women Q.* 37(1):113–127.
42 Cole ER. 2009. Intersectionality and research in psychology. *Am Psychol.* 64(3):170–180.
43 Sidanius J, Pratto F. 1999. Social dominance theory. In: *Handbook of theories of social psychology*. New York: SAGE Publications Ltd. p. 418–438.
44 Johnson KL, Freeman JB, Pauker K. 2012. Race is gendered: how covarying phenotypes and stereotypes bias sex categorization. *J Pers Soc Psychol.* 102(1):116–131.
45 Bartoš V, Bauer M, Chytilová J, Matějka F. 2016. Attention discrimination: theory and field experiments with monitoring information acquisition. *Am Econ Rev*. 106:1437–1475.
46 Bogen M, Rieke A. 2018. *Help wanted: an examination of hiring algorithms, equity, and bias*. California: Upturn.
47 Chen Z. 2023. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanit Soc Sci Commun.* 10(1):1–12.
48 Aydin O, Karaarslan E, Narin NG. 2024. Artificial intelligence, vr, ar and metaverse technologies for human resources management, *arXiv, arXiv:2406.15383*, preprint: not peer reviewed.
49 Glazko K, Mohammed Y, Kosa B, Potluri V, Mankoff J. 2024. Identifying and improving disability bias in GPT-based resume screening. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. New York: Association of Computing Machinery. p. 687–700.
50 Beatty D, Masanthia K, Kaphol T, Sethi N. 2024. Revealing hidden bias in AI: lessons from large language models, *arXiv, arXiv:2410.16927*, preprint: not peer reviewed.
51 Gerszberg NR. 2024. Quantifying gender bias in large language models: when ChatGPT becomes a hiring manager [PhD thesis]. Massachusetts Institute of Technology.
52 Lippens L. 2024. Computer says 'no': exploring systemic bias in ChatGPT using an audit approach. *Comput Hum Behav Artif Hum.* 2(1):100054.
53 Morehouse K, Pan W, Contreras JM, Banaji MR. 2024. Bias transmission in large language models: evidence from gender-occupation bias in GPT-4. ICML 2024 Next Generation of AI Safety Workshop.
54 Veldanda AK, *et al.* 2023. Investigating hiring bias in Large Language Models. R0–FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Large Foundation Models at NeurIPS 2023.
55 Wang Z, *et al.* 2024. Jobfair: a framework for benchmarking gender hiring bias in large language models, *arXiv, arXiv:2406.15484*, preprint: not peer reviewed.
56 Guo W, Caliskan A. 2021. Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York: Association of Computing Machinery. p. 122–133.
57 Haim A, Salinas A, Nyarko J. 2024. What's in a name? Auditing large language models for race and gender bias, *arXiv, arXiv:2402.14875*, preprint: not peer reviewed.
58 Kirk HR, *et al.* 2021. Bias out-of-the-box: an empirical analysis of intersectional occupational biases in popular generative language models. *Adv Neural Inf Process Syst*. 34:2611–2624.
59 Ma W, Chiang B, Wu T, Wang L, Vosoughi S. 2023. Intersectional stereotypes in large language models: dataset and analysis. In Findings of the Association for Computational Linguistics: EMNLP 2023. USA: Association for Computational Linguistics. p. 8589–8597.
60 Omrani Sabbaghi S, Wolfe R, Caliskan A. 2023. Evaluating biased attitude associations of language models in an intersectional context. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. New York: Association of Computing Machinery. p. 542–553.
61 Wilson K, Caliskan A. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York: Association of Computing Machinery. p. 1578–1590.
62 Armstrong L, Liu A, MacNeil S, Metaxa D. 2024. The silicon ceiling: auditing GPT's Race and gender biases in hiring. In Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. New York: Association of Computing Machinery. p. 1–18.
63 Yin L, Alba D, Nicoletti L. OpenAI's GPT is a recruiter's dream tool: tests show there's racial bias. Bloomberg [accessed 2024 Mar 8]; https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination.
64 Barro RJ. 1991. Economic growth in a cross section of countries. *Q J Econ.* 106(2):407–443.
65 Nelson RR, Phelps ES. 1966. Investment in humans, technological diffusion, and economic growth. *Am Econ Rev*. 56(1/2):69–75.
66 Romer PM. 1990. Endogenous technological change. *J Polit Econ.* 98(5, Part 2):S71–S102.
67 Becker GS, Murphy KM, Tamura R. 1990. Human capital, fertility, and economic growth. *J Polit Econ.* 98(5, Part 2):S12–S37.
68 Murphy KM, Shleifer A, Vishny RW. 1991. The allocation of talent: implications for growth. *Q J Econ.* 106(2):503–530.
69 Acemoglu D. 1995. Reward structures and the allocation of talent. *Eur Econ Rev*. 39(1):17–33.
70 Blanchard O. *Monetary policy and unemployment*. The US Euro-Area, and Japan, 2005. p. 9–15.
71 Friedman BM, Woodford M, editors. 2010. *Handbook of monetary economics*. Elsevier.
72 Abdurahman S, *et al.* 2024. Perils and opportunities in using large language models in psychological research. *PNAS Nexus*. 3(7):245.
73 Atari M, Xue MJ, Park PS, Blasi D, Henrich J. 2024. Which humans? *PsyarXiv* https://osf.io/preprints/psyarxiv/5b26t, preprint: not peer reviewed.
74 Zewail A, Figueroa A, Graham J, Atari M. 2024. Moral stereotyping in large language. *Psyarxiv*. https://osf.io/preprints/psyarxiv/t9×8r, preprint: not peer reviewed.
75 Zhou KZ, Sanfilippo MR. 2023. Public perceptions of gender bias in large language models: cases of ChatGPT and Ernie, *arXiv, arXiv:2309.09120*, preprint: not peer reviewed.
76 Rozado D. 2024. The political preferences of LLMs, *arXiv, arXiv:2402.01789*, preprint: not peer reviewed.
77 Santurkar S, Durmus E, Ladhak F, Lee C, Liang P. 2023. Whose opinions do language models reflect? In Proceedings of the 40th International Conference on Machine Learning. New York: Association of Computing Machinery. p. 29971–30004.

78 Schramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat Mach Intell*. 4(3): 258–268.
79 Dillion D, Mondal D, Tandon N, Gray K. 2024. Large language models as moral experts? GPT-4o outperforms expert ethicist in providing moral guidance. *PsyarXiv* https://osf.io/preprints/psyarxiv/w7236, preprint: not peer reviewed.
80 Crockett M, Messeri L. 2023. Should large language models replace human participants? *PsyarXiv* https://osf.io/preprints/psyarxiv/4zdx9, preprint: not peer reviewed.
81 Chen L, Zaharia M, Zou J. 2023. How is ChatGPT's behavior changing over time?, *arXiv, arXiv:2307.09009*, preprint: not peer reviewed.
82 Varnum ME, Baumard N, Atari M, Gray K. 2024. Large language models based on historical text could offer informative tools for behavioral science. *Proc Natl Acad Sci U S A*. 121(42): e2407639121.
83 Bail CA. 2024. Can generative AI improve social science? *Proc Natl Acad Sci U S A*. 121(21):e2314021121.
84 Capraro V, *et al*. 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*. 3(6):191–208.
85 Henrich J, Heine SJ, Norenzayan A. 2010. The weirdest people in the world? *Behav Brain Sci*. 33(2–3):61–83.
86 Tao Y, Viberg O, Baker RS, Kizilcec RF. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*. 3(9): 346–355.