No Thoughts Just AI: Biased LLM Hiring Recommendations Alter Human Decision Making and Limit Human Autonomy

Kyra Wilson[1], **Mattea Sim**[2], **Anna-Maria Gueorguieva**[1], **Aylin Caliskan**[1]
[1]University of Washington
[2]Indiana University
kywi@uw.edu, matsim@iu.edu, agueorg@uw.edu, aylin@uw.edu

Abstract

Despite bias in artificial intelligence (AI) being a risk of their use in hiring systems, there is no large-scale empirical investigation of the impacts of these biases on hiring decisions made collaboratively between people and AI systems. It is also unknown whether AI literacy, people's own biases, and behavioral interventions intended to reduce discrimination affect these human-in-the-loop AI teaming (AI-HITL) outcomes. In this study, we conduct a resume-screening experiment (N=528) where people collaborate with simulated AI models exhibiting race-based preferences (bias) to evaluate candidates for 16 high and low status occupations. Simulated AI bias approximates factual and counterfactual estimates of racial bias in real-world AI systems. We investigate people's preferences for White, Black, Hispanic, and Asian candidates (represented through names and affinity groups on quality-controlled resumes) across 1,526 scenarios and measure their unconscious associations between race and status using implicit association tests (IATs), which predict discriminatory hiring decisions but have not been investigated in human-AI collaboration. This evaluation framework can generalize to other groups, models, and domains. When making decisions without AI or with AI that exhibits no race-based preferences, people select all candidates at equal rates. However, when interacting with AI favoring a particular group, people also favor those candidates up to 90% of the time, indicating a significant behavioral shift. The likelihood of selecting candidates whose identities do not align with common race-status stereotypes can increase by 13% if people complete an IAT before conducting resume screening. Finally, even if people think AI recommendations are low quality or not important, their decisions are still vulnerable to AI bias under certain circumstances. This work has implications for people's autonomy in AI-HITL scenarios, AI and work, design and evaluation of AI hiring systems, and strategies for mitigating bias in collaborative decision-making tasks. In particular, organizational and regulatory policy should acknowledge the complex nature of AI-HITL decision making when implementing these systems, educating people who use them, and determining which are subject to oversight.

1 Introduction

The use of artificial intelligence (AI) in hiring processes has received increasing attention from researchers, regulators, and employers. These technologies might improve the efficiency of labor-intensive hiring processes, such that one company reported saving over £1 million and decreasing hiring time by 90% by incorporating AI screening tools into their hiring procedures (Featured Customers). However, increasing adoption of these systems is not without risks because they may exhibit different behaviors based on candidates' social identities rather than qualifications (*bias*), possibly leading to illegal discrimination (Fabris et al. 2025; Wilson and Caliskan 2024; Glazko et al. 2024; Wilson and Caliskan 2025). In 2018, Amazon reported an instance of this when an internal hiring tool unfairly discriminated against female applicants (Dastin 2018), eventually leading to widespread interest in methods to prevent and/or mitigate societal harms that result from biased systems.

In this paper, we describe a large-scale human subjects

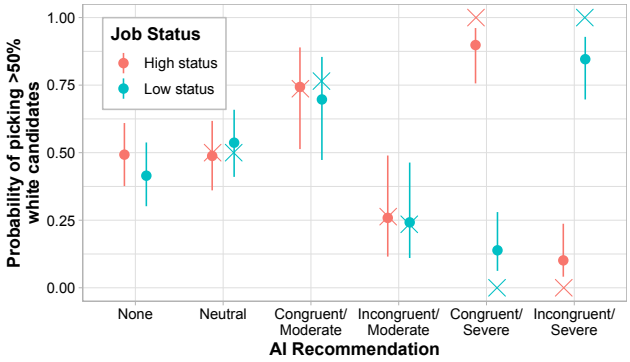

Figure 1: Predicted probability of preference for White candidates in the resume screening task when participants see no AI recommendation (None), unbiased recommendation (Neutral), or recommendations which varied in direction (Congruent/Incongruent) and magnitude (Moderate/Severe) of bias. For *AI Recommendation* and *Job Status*, dots and lines indicate estimates and 95% confidence intervals of human selection rates, and X marks the proportion of AI recommendations favoring White candidates. Without AI recommendations, people chose White and non-White candidates at similar rates. With AI recommendations, people's choices closely paralleled AI suggestions, regardless of direction or magnitude of AI bias.

Group	Feature	Values
Asian	First Name	Hong, Huang, Xin, Yong
	Last Name	Chen, Kim, Nguyen, Tran
Black	First Name	Jamal, Leroy, Mohammad, Lamar
	Last Name	Jefferson, Johnson, Washington, Williams
Hispanic	First Name	Alejandro, Jesus, Pablo, Santiago
	Last Name	Hernandez, Lopez, Martinez, Rodriguez
White	First Name	Brent, Dustin, Gary, Todd
	Last Name	Johnson, O'Brien, Miller, Williams
All	Racial Affinity Org.	{Asian, Black, Hispanic, ∅} Student Action Association, ___ Student Association, ___ Student Leadership Coalition, ___ Student Union
	Ethnic Affinity Org.	{Chinese American, Haitian American, Mexican American, English American} Association, ___ Heritage Club, ___ Society, ___ Youth Organization

Table 1: Features used on resumes to signal candidates' racial identity.

experiment conducted to determine how people's decisions are impacted by biased AI hiring recommendations, and whether individual traits or exposure to bias training moderate these decisions. Prevailing guidance for working with AI hiring tools responsibly and effectively is to use "human-in-the-loop" AI teaming (AI-HITL) strategies (Tabassi 2023; EU AI Act), meaning people and AI systems make decisions collaboratively, with people having autonomy and agency to review or alter AI decisions before they are enacted. There are numerous reasons AI-HITL systems are favored, such as providing flexibility for differing societal contexts, having accountability, and producing reasoning that is consistent with regulations and able to be challenged (Binns 2022). Accordingly, as of 2024, only 21% of companies reject applicants without human review, suggesting AI-HITL is widely used in hiring processes (Resume Builder 2024).

AI-HITL could be especially beneficial in high-stakes domains such as hiring if people are able to counteract or mitigate AI biases, but whether this is possible is an open question. First, humans themselves can be biased when making hiring decisions (Bertrand and Mullainathan 2004); therefore they might not be capable of recognizing and correcting AI biases, leading to harm for both employers and job seekers. If this is the case, then AI-HITL strategies alone may not be effective for mitigating biases which originate from AI systems. Furthermore, bias in these systems could be seen as a barrier to human autonomy, a capacity to act on *one's own* beliefs, values, motivations, and reasons (Prunkl 2024). This capacity is crucial and highly valued for high-stakes decisions (Kim et al. 2024; Li et al. 2021; Aizenberg, Dennis, and van den Hoven 2025).

In this study, we are the first to examine how (racially biased) AI recommendations impact people's decisions and how factors related to their unconscious bias training, experiences, and perceptions contribute to their ability to act autonomously and counteract biased AI recommendations. We simulate a resume-screening task for 16 occupations where participants select candidates to advance to another hypothetical hiring evaluation. As racial stereotypes are often related to societal status (Fiske et al. 2018), these occupations comprise both high and low status occupations which are likely to align with or diverge from participants' status-race associations. Candidate profiles (which are controlled and validated by human annotators) are shown with or without simulated AI recommendations that vary according to racial identities they favor (direction) and how much they are favored (magnitude). For example, AI recommendations may be biased in ways that reinforce common racial stereotypes (congruent) or that contradict them (incongruent). These AI biases were grounded in simulations of resume screening with real-world AI models or selected for counterfactual analysis which can inform future system development.

Participants perform the resume-screening task either before or after taking an implicit association test (IAT), which is similar to those used in workplace unconscious bias trainings (Williamson and Foley 2018). This study is the first to investigate the role of implicit associations in moderating AI-HITL decisions. We also collect information about participants that could impact their interactions with AI and decisions, such as their previous experience with hiring and using AI, perceptions of the AI model used, and explicit race-status beliefs. We conduct our experiment in three settings based on racial/ethnic bias: comparing White vs. Black, White vs. Asian, or White vs. Hispanic candidates. Despite the prevalence of gender and occupational stereotypes (Caliskan, Bryson, and Narayanan 2017), we study racial bias only among male candidates because disparate impacts are greater for racial groups than gender groups when using AI for resume screening (Wilson and Caliskan 2024).

In total, 528 participants completed 1,526 resume-screening scenarios, making this the largest scale human subjects experiment (to date) investigating interactions between humans and racially-biased AI in decision-making tasks.[1] We use a framework which simulates existing and hypothetical social effects of AI and can be generalized beyond resume screening to tasks in other AI-HITL domains, and we make three main contributions, in addition to releasing anonymized behavioral data from our experiment and accompanying analysis code.[2]

First, when making decisions without AI or with unbiased AI, people select White and non-White candidates equally. However, **when interacting with AI favoring a particular group, people select those candidates up to 90% of the time** (as shown in Figure 1), suggesting AI bias propagates to human decision makers. Second, **completing an IAT before the resume-screening task can increase par-**

[1]Rosenthal-von der Pütten and Sach (2024) (the most similar work to ours) studied immigrant bias favoring German or Turkish candidates across 520 scenarios completed by 260 participants.

[2]Code and data are available at https://github.com/kyrawilson/No-Thoughts-Just-AI.

ticipants' selection rate of stereotype-incongruent candidates by 13%, indicating that system design and bias training can play a role in reducing AI bias propagation. Finally, exploratory analysis of other contributing factors suggests that people's prior experience with hiring or AI and their implicit biases and explicit beliefs regarding race and status do not moderate hiring decisions. However, **perceptions of AI recommendation quality and importance do moderate hiring decisions**, meaning AI literacy interventions are worth further investigation.

2 Related Work

Human-AI teaming is a growing area of research, particularly in regards to the influence of AI systems on human behavior and decisions. For example, people may follow incorrect recommendations or advice from AI (over-reliance), and thus systems must be calibrated so that human and AI knowledge is complementary, and collaborations improve upon individual performance. Some research has highlighted the potential for explanations to reduce over-reliance (Chen et al. 2023; Lee and Chew 2023), however their efficacy is not universal and often depends on the type of the explanation (Schoeffer, De-Arteaga, and Kuehl 2024; Spatola 2024). Other experiments emphasize the role of psychological factors such as propensity to trust and affinity for technology interaction as moderators of reliance (Küper and Krämer 2025). Finally, task characteristics may also influence how likely people are to follow AI recommendations, so situated evaluation is necessary. Vasconcelos et al. (2023) show that explanations are more valuable when tasks are difficult, and Cao and Huang (2022) show that over-reliance is less likely when tasks are easy.

Because AI tools used for hiring can be biased (Glazko et al. 2024; Wilson and Caliskan 2024), with possible legal consequences, evaluating and understanding human reliance in this setting is essential. Some psychological traits like extraversion and self-confidence influence recruiters' likelihood to trust unbiased AI recommendations (Lacroux and Martin-Lacroux 2022; Gonzalez et al. 2022), but there is little work investigating the role of implicit biases in moderating AI-HITL scenarios, despite their association with decision making in hiring (Agerström and Rooth 2011; Reuben, Sapienza, and Zingales 2014). An additional reason to study implicit associations is that they are commonly used to inform workers about their unconscious biases, which can play a role in workplace dynamics as well as decisions (Williamson and Foley 2018).

Implicit associations are typically measured using IATs, first proposed by Greenwald, McGhee, and Schwartz (1998) as a way to measure associations via differences in reaction times when sorting words or pictures representing two concepts of interest. Most studies predicting discriminatory decision making with IATs use tests associating social categories and valence; however, associating categories with beliefs may be better at predicting behavior (Rudman and Ashmore 2007; Montgomery et al. 2024). While studies such as Agerström and Rooth (2011); Reuben, Sapienza, and Zingales (2014) have shown relationships between non-racial social group associations and hiring outcomes, to our knowledge this is not been investigated using associations between racial groups and specific beliefs or in the context of AI-HITL hiring.

AI Rec.	Job Status	White vs. Black	White vs. Asian	White vs. Hispanic
None	High	N/A	N/A	N/A
	Low	N/A	N/A	N/A
Neutral	High	.500	.500	.500
	Low	.500	.500	.500
Cong/ Mod	High	.835 (.690 / .980)	.765 (.680 / .850)	.610 (.470 / .750)
	Low	.830 (.870 / .790)	.695 (.680 / .710)	.770 (.880 / .660)
Cong/ Sev	High	1.000	1.000	1.000
	Low	0.000	0.000	0.000
Incong/ Mod	High	.165 (.390 / .020)	.235 (.320 / .150)	.390 (.530 / .250)
	Low	.170 (.130 / .210)	.305 (.320 / .290)	.230 (.120 / .340)
Incong/ Sev	High	0.000	0.000	0.000
	Low	1.000	1.000	1.000

Table 2: Proportion of simulated AI recommendations that favor White candidates in various combinations of *Race*, *Job Status*, and magnitude and direction of *AI Recommendation* bias. For Moderate bias conditions, two values are given for jobs with worker demographics that approximate the overall US population vs. those that do not. The results in this paper are presented in terms of the average of these values.

Of the studies which do examine interactions with biased AI, the range of biases investigated are also limited to those which are observed in existing systems or are congruent with dominant societal stereotypes, limiting their generalization to future systems which may exhibit different biases. Furthermore, whether humans amplify or mitigate AI biases is inconsistent (Peng et al. 2022; Bursell and Roumbanis 2024; Rosenthal-von der Pütten and Sach 2024; Wilkens et al. 2025). We seek to address these limitations in AI-HITL interaction evaluation by analyzing both existing and counterfactual biases generated via theoretically informative simulations. Additionally, we investigate the role of individual traits which are known to influence human-only hiring decisions, such as implicit associations (Agerström and Rooth 2011), but have not been examined in the context of AI-HITL scenarios.

3 Data and Methods

The study used a 6x3x2x2 mixed factorial design. The partial within-subjects factor, *AI Recommendation*, had six combinations of bias magnitude and direction: None (no recommendation), Neutral (recommend White and non-White candidates equally), Congruent/Moderate, Incongruent/Moderate, Congruent/Severe, and Incongruent/Severe. Congruent and Incongruent refer to the preference direction

of AI recommendations relative to dominant cultural stereotypes in the US; Moderate and Severe refer to the magnitude of AI bias. Each participant saw the None and Neutral levels and both of either the Congruent or Incongruent levels (four scenarios total). The second factor, *Race*, had three between-subjects levels: White vs. Black, White vs. Asian, or White vs. Hispanic. The third factor, *Task Order*, had two between-subjects levels: Decision/IAT and IAT/Decision. The final factor, *Job Status*, had two between-subjects levels: High Status and Low Status.

3.1 Stimuli Materials

Occupations and Descriptions Because we were interested in hiring decisions in the context of racial bias due to the strength of these biases in AI models (Wilson and Caliskan 2024), we selected occupations likely to be associated with particular racial groups. Specifically, we chose occupations which are typically judged to be high or low status because prior work has shown that people's perceptions of occupational status is related to the racial composition of its workers (Valentino 2022) and that people have implicit associations between status and race (Melamed et al. 2019, 2020). We selected high vs. low status occupations based on their average annual salaries reported by the 2022 American Community Survey's (ACS) 5-Year Estimates[3] ($30k-$35k or $110k-$135k, respectively), as status ratings are most predicted by pay (Valentino 2022). Within each set of high or low status occupations, there was variation in actual demographics and population size of US workers, including both skewed and representative racial distributions. More detailed information about the occupation selection procedure is available in the Appendix.[4]

The set of 16 occupations selected included eight high status occupations (*sales engineer, construction manager, industrial production manager, nurse practitioner, management analyst, talent agent, computer systems analyst, health services manager*) and eight low status occupations (*agricultural grader, housekeeper, home health aide, textile presser, food preparer, bus person, sales associate, usher*). The authors wrote short descriptions of each occupation using information from O-NET/My Next Move[5] about the purpose of the occupation and typical job duties plus information about salaries from ACS. Figure 2 shows an example description; complete occupation descriptions, salary, and demographic information are available in the Appendix.

Resumes We used ChatGPT-4o, one of the best performing generative large language models (LLM) according to the Open LLM Leaderboard (Fourrier et al. 2024) in December 2024, to create eight hypothetical candidate work histories for each occupation, following the approach used by Armstrong et al. (2024) to generate fictitious resumes using GPT. The prompt is available in the Appendix. Existing resume datasets could not be used because they do not contain resumes for the occupations of interest, and we chose not to gather resume content from websites like LinkedIn in order to preserve individual privacy and comply with platform policies.

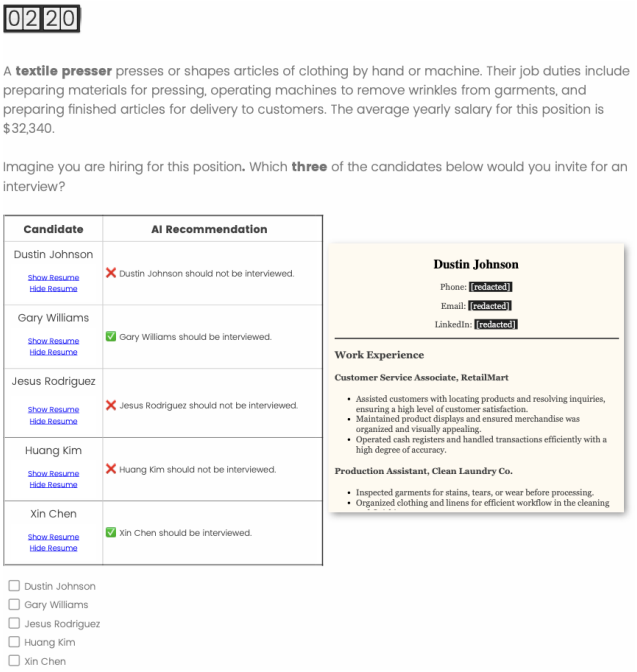

Figure 2: An example of the interface 575 participants saw when completing resume-screening trials. They had four minutes to complete each of four scenarios, in which they read an occupation description, five candidate resumes, and AI recommendations and selected three candidates that should be invited for an interview.

The content of the generated work histories was validated by asking 40 participants recruited from Prolific to score how qualified a candidate was for an occupation given their work history. These participants reported using the occupation start and end dates as quality metrics, so they were removed from the work histories before being used in the resume-screening experiment. For each occupation, we used the work histories with the four most similar validation scores as stimuli for resume screening to ensure the quality did not meaningfully vary across candidates. More details about procedure, analysis, and results of the resume quality validation study can be seen in the Appendix.

To form a complete resume, each work history was augmented with names and additional interests intended to signal a particular racial identity. First and last names were picked from Elder and Hayes (2023), which describes and releases a dataset of names rated on various traits by human evaluators. We chose by names that were most associated with each racial identity and excluded first names that were more associated with women than men to avoid confounding gender or intersectional associations (Fiske et al. 2018; Shaked et al. 2016).

Although many hiring discrimination studies vary only names to signal racial identities (Wilson and Caliskan 2024;

[3]https://data.census.gov/app/mdat/ACSPUMS5Y2022

[4]The Appendix is available at https://arxiv.org/a/wilson_k_1.

[5]https://www.mynextmove.org

Bertrand and Mullainathan 2004), we include membership in both racial and ethnic affinity groups as additional resume content since names are not unambiguously and universally associated with sociodemographic traits (Elder and Hayes 2023; Gautam et al. 2024). We include explicit race labels by combining a randomly selected position (President, Vice President, Treasurer, or Secretary) with the name of a randomly selected racial affinity organization based on those at universities, as shown in Table 1. For Black, Hispanic, and Asian candidates, racial identity was explicitly stated, but White candidates had no explicit race stated to avoid associations with White supremacist movements that could impact quality judgments. Furthermore, in the US, White identity is often assumed, even when not explicitly labeled, because this is the dominant social group (Cheng, Durmus, and Jurafsky 2023). Because national and ethnic origin is also highly associated with racial identity (Weerts et al. 2024), we indicate membership in an additional randomly selected ethnic affinity organization, also listed in Table 1.

AI Recommendations AI recommendations exhibited various bias levels, which were determined either by simulating resume screening in real AI systems or selecting counterfactual biases which are theoretically informative for generalization to systems with biases different from those in our simulation. In the None condition, no AI recommendations were given; in the Neutral condition, exactly one White and non-White candidate were recommended in each scenario. In the Severe conditions, every White candidate and no non-White candidates were recommended for High Status jobs, and vice versa for Low Status jobs. These conditions were designed to examine the most extreme instances of bias to determine impacts on the bounds of human decisions.

To approximate real-world AI resume screening bias, we followed the procedure introduced in Wilson and Caliskan (2024) to evaluate resume screening in an LLM retrieval setting. Congruent/Moderate bias was determined by performing resume screening and recommending White candidates at the same rates they were preferred by LLMs. In the Incongruent/Moderate condition, AI systems recommended non-White candidates at the same rate White candidates were preferred in the Congruent/Moderate condition.

To encode job descriptions and resumes augmented with racial features into embedding representations, we used three LLMs designed for embedding-based tasks which also exhibit racial bias as shown in Wilson and Caliskan (2024): E5-mistral-7b-instruct (Wang et al. 2023), GritLM-7B (Muennighoff et al. 2024), and SFR-Embedding-Mistral (Meng et al. 2024). After ranking the resumes according to their cosine similarity with corresponding job descriptions, we selected the top 10% of resumes and computed the proportion from each racial identity. The final magnitude of bias in the Moderate scenarios was the average of these proportions across all models and occupations for high and low status occupations, and they are shown in Table 2. Additional details about the procedure, analysis, and results of the AI resume-screening simulation are available in the Appendix.

IATs, Explicit Beliefs, and Survey Questions We assessed participants' implicit associations between status and racial identities using race-status materials from Melamed et al. (2019) and Montgomery et al. (2024) in an IAT implemented on Qualtrics with `iatgen` (Carpenter et al. 2019). The experimental factor *Task Order* refers to whether or not IATs appeared before or after the resume-screening task. This is in order to determine whether interacting with IAT trainings (similar to those currently used to mitigate unconscious bias) before completing an AI-HITL decision-making task is also useful for reducing biased outcomes.

We also asked people their explicit beliefs about status and race, which are related to but distinct from implicit associations because of their dependence on external social factors and relative stability (Hofmann et al. 2005). We used a subset of 16 competence-related questions (eight each about the White and non-White groups) from Fiske et al. (2018)'s Stereotype Content Model scale, which measures the strengths of people's beliefs about the status of racial groups. We used only questions related to competence because of its close links with status perceptions (Brambilla et al. 2010; Fiske et al. 2018). Participants responded to each question using a 5-point Likert scale.

Finally, we asked people about their impressions of the AI recommendations, both in terms of their quality and how important they were for making decisions; whether they have previous experience hiring or managing employees; and whether they have heard or read about AI being used for hiring tasks. Participants responded to these using a 3-point Likert scale. A complete list of IAT materials, explicit beliefs questions, and survey questions are available in the Appendix.

3.2 Participants

We recruited 575 participants from Prolific who live in the United States, speak English fluently, and did not previously validate the quality of generated work histories on Prolific. Participants were not excluded from participation on the basis of any identities such as race or gender. Of these, 528 had usable data (exclusion criteria is described in Section 3.4)—47.9% were men, 50.4% were women, and the remaining 1.7% did not report one of these two identities. Participants' average age was 39.1 years (SD=11.7). The majority (70.4%) of participants were White or European alone or in combination with another racial identity; 21.3% were Black or African alone or in combination with another identity; 7.2% were Hispanic or Latino/a/x alone or in combination with another identity; 5.0% were Asian or Asian American alone or in combination with another identity; finally, 1.3% indicated another race not investigated in this study.[6] Only 30.0% of participants said they had taken an IAT previously, with the remainder saying they had not or weren't sure. We paid each participant $8.65 for approximately 25 minutes spent completing the experiment, in line with Seattle's minimum wage in January 2025.

[6]These proportions do not sum to 100% because people can belong to more than one group.

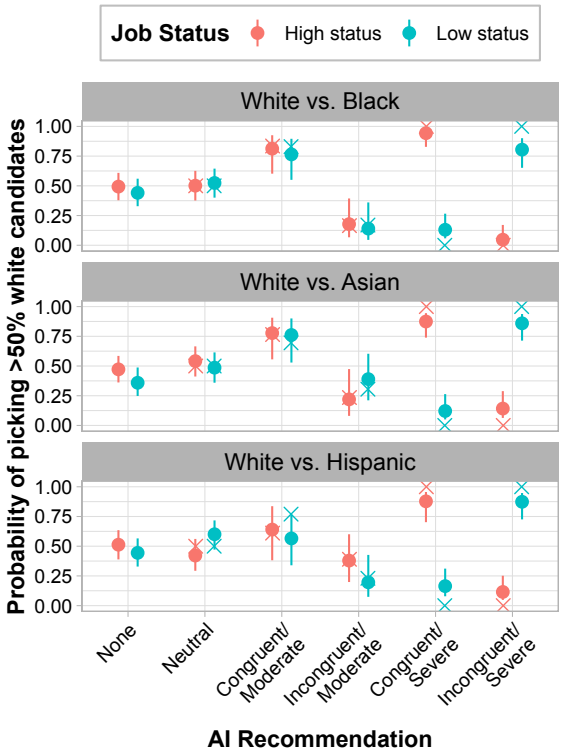

Figure 3: Predicted probability of participants preferring White candidates in resume screening split by *Race*, *Job Status*, and *AI Recommendation* conditions. X marks the proportion of AI recommendations favoring White candidates. Particpants' likelihood of preferring White candidates is strongly associated with the *AI Recommendation* and *Job Status* they saw, but not the *Race*.

3.3 Experimental Procedure

Before beginning the tasks, participants signed a consent form and were randomly assigned to levels of the *Race*, *Task Order*, *Job Status* factors. For *AI Recommendation*, they were randomly assigned a subset of all conditions. Depending on their *Task Order* assignment, participants read instructions for either the IAT or the resume-screening task and completed that part of the experiment, followed by the other part. In order to keep participants naive to the true purpose of the study, they were told only that researchers were interested in knowing whether AI recommendations were similar to humans' and if they improved decision-making efficiency. After completing both tasks, participants answered questions about their explicit beliefs, AI and hiring experience, and perceptions of the AI recommendations. Finally, we debriefed participants to the purpose of the experiment after all tasks and questionnaires were complete; full instructions and debrief text are available in the Appendix.

In the resume-screening task, participants were given a description of an occupation and the names and resumes of five job candidates. There were four qualified resumes, two of which belonged to White candidates and two of which belonged to non-White candidates (either Asian, Black, or Hispanic, depending on the assigned *Race* condition); the final resume lacked qualifications (as content was written for an occupation different than the one of interest) and they were never given a positive AI recommendation. Additionally, this candidate's apparent race was randomly chosen from the identities not in the main comparison. This distractor candidate was included for several reasons: first, having three candidates of different races obscured the true purpose of the experiment; second, the candidate was unambiguously less qualified and thus served as an attention check, such that selecting this candidate indicated a failure to pay attention to the task resulting in the exclusion of that trial from analysis.

Participants had four minutes to review all candidates' resumes and AI recommendations and select three of the five candidates which they thought were most suitable for the given occupation. We used this amount of time so that participants spent approximately one minute reviewing each qualified resume in order to align with the time constraints in real-world resume screening (Chan 2024) that might cause decision makers to rely on biased heuristics (Kahneman 2011). Once the four minutes had passed, participants could no longer view the resumes and had to submit their choices. Choosing three of five candidates provided a number of benefits: first, it more realistically represents stages of resume screening in which multiple candidates are compared simultaneously rather than the binary comparison used by most laboratory resume-screening experiments; second, it forces the participant to choose an unequal number of candidates from each race (two White candidates and one non-White candidate, or vice versa). Whether participants favor White or non-White candidates in particular conditions can then be estimated by modeling which racial majority is chosen most often in response to different kinds of AI recommendations.

Participants completed four total trials of the decision task. In the first trial, they saw no AI recommendations, only candidate resumes. In the remaining trials, they saw resumes and AI recommendations which were Neutral (recommending exactly one candidate from each comparison race), Congruent/ or Incongruent/Moderate (recommending candidates based on simulated levels of realistic AI racial bias), and Congruent/ or Incongruent/Severe (recommending all candidates from one race and none from the other). The final three trials were always presented in a random order after the first trial, in order to avoid priming participants in scenarios with no AI recommendations. An example of the interface participants saw in each trial is in Figure 2.

In the race-status IAT task adapted from Montgomery et al. (2024) and Melamed et al. (2019), participants sorted words or pictures associated with targets (racial identities) or attributes (social statuses) by pressing keys on a keyboard in response to an item appearing on the screen. In the first and second practice blocks, only targets and attributes are sorted, respectively. In blocks three and four, targets and attributes are sorted together. In the remaining blocks, the prior three blocks are repeated with sorting categories appearing in reversed positions on the screen. This task takes approximately five minutes. The IAT stimuli and an example of the IAT interface is shown in the Appendix.

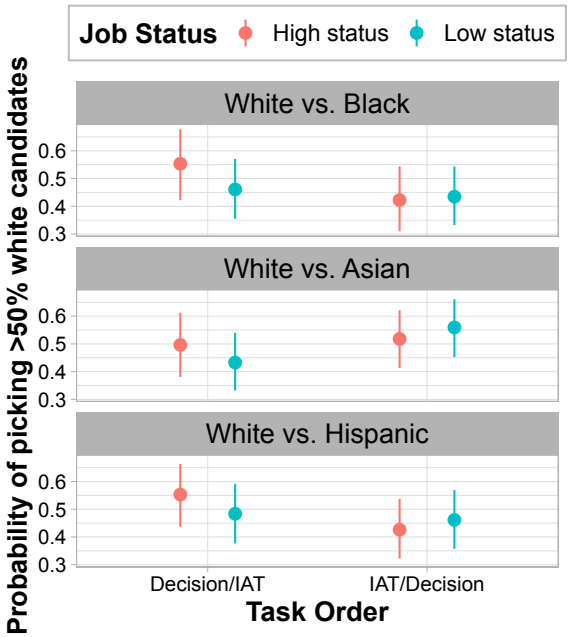

Figure 4: Predicted probability of participants preferring White candidates in resume screening split by *Race*, *Job Status*, and *Task Order*. There is a significant interaction between *Job Status* and *Task Order* but no significant pairwise comparisons. Trends show completing an IAT before the decision task increases stereotype-incongruent beliefs by 13%.

3.4 Analysis

AI Recommendation, Race, Task Order, Job Status To determine whether AI recommendations and IAT presentation order affected participants' hiring decisions, we fit a binomial logistic mixed model (BLMM) with the default logit link function. By using this model, we are able to model the probability of a particular binary outcome (i.e. preferring White or non-White candidates); therefore, results are discussed in terms of probabilistic outcomes. For the predictor variables, we investigated the interactions between *AI Recommendation*, *Job Status*, and *Race*; *AI Recommendation*, *Job Status*, and *Task Order*; *Race* and *Task Order*; as well as lower-order interactions and main effects. In addition to these fixed effects, we included random intercept effects for participant and occupation. The full model specification is in Equations 0-1. The regression models were fit using the `glmmTMB` R package. Using the fit BLMM, we conducted omnibus ANOVA analyses for main effects and interactions using the `car` R package and post-hoc pairwise comparisons using the `emmeans` R package.

Exploratory Factors Race-status IATs were scored according to the algorithm in Greenwald, Nosek, and Banaji (2003), which gives each participant an effect size score d, where greater positive values mean greater stereotype-congruent associations and smaller negative values mean greater stereotype-incongruent associations. The strength of Cohen's d effect size used for IAT scoring is small for $0.2 \leq d < 0.5$, medium for $0.5 \leq d < 0.8$, and large for $d \geq 0.8$ (Cohen 2016). We also calculated Cohen's d effect sizes for each participant's explicit beliefs about race and status using their questions about White vs. non-White groups; the interpretation is the same as for IAT d. Responses to other survey questions were used as-is.

We used these predictors to conduct exploratory analyses to determine which psychological or experiential factors are likely to influence people's decisions when interacting with AI in a hiring setting. We fit another BLMM using the predictors in Equations 0-1; interactions between *AI Recommendation*, *Job Status*, and each exploratory factor; and interactions between *Race* and IAT and explicit belief scores. The full exploratory model is given in Equations 0-2.

We performed stepwise backwards elimination using the `buildlmer` R package to determine which of these predictors were most likely to influence decision-making outcomes. In this procedure, the model with all predictors is fit, and then each predictor is successively removed from the model if eliminating it improves the fit according to a likelihood ratio test (Matuschek et al. 2017). Only the predictors which contribute most to the model fit remain at the end. Although stepwise regression shouldn't be used for inference or null hypothesis significance testing, it is acceptable for exploratory analysis to determine which variables are most suitable for further investigation (Tredennick et al. 2021; Heinze, Wallisch, and Dunkler 2018; Zhou et al. 2024). Additional analyses in the Appendix using variable importance metrics and elastic net regression instead of stepwise regression corroborate findings presented in Section 4.2.

$$(0) \quad X = \text{AI Recommendation} * \text{Job Status}$$

$$(1) \quad Response \sim X * \text{Race} + X * \text{Task Order} + \text{Race} * \text{Task Order} + (1|Participant) + (1|Job)$$

$$(2) \quad Response \sim \ldots + X * \text{IAT Score} + \text{Race} * \text{IAT Score} + X * \text{Explicit Score} + \text{Race} * \text{Explicit Score} + X * \text{AI Exp.} + X * \text{Hiring Exp.} + X * \text{AI Quality} + X * \text{AI Importance}$$

4 Results

4.1 AI Recommendation, Race, Task Order, Job Status

After removing trials where participants selected distractor candidates, we reduced the number of trials for analysis from 2,300 to 1,955. Furthermore, in the first wave of participants, we found an error in the proportion of times candidates were recommended in Neutral and Moderate *AI Recommendation* conditions. Excluding these trials left 1,526 total data points for analysis.[7]

Figure 3 shows predicted probabilities of favoring White candidates by *Race*, *Job Status*, and *AI Recommendation*. **The most biased outcomes are in Severe conditions for**

[7]We did not remove responses to None and Severe conditions from the first wave of participants because they did not significantly differ from the responses of participants in different waves, suggesting the error did not effect other conditions.

Job Status	AI Rec.	Prob.	Δ AI Rec.	Δ None	Δ Neutral
High	None	.493	N/A	N/A	.005
	Neutral	.488	-.012	-.005	N/A
	Cong/Mod	.750	.013	.257*	.262*
	Cong/Sev	.904	-0.96**	.411**	.416**
	Incong/Mod	.250	-.013	-.243*	-.238
	Incong/Sev	.093	.093**	-.400**	-.395**
Low	None	.414	N/A	N/A	-.123
	Neutral	.537	.037	.123	N/A
	Cong/Mod	.704	-.061	.290**	.167
	Cong/Sev	.138	-.138**	-.276**	-.399**
	Incong/Mod	.227	-.008	-.187	-.310**
	Incong/Sev	.848	-.152**	.434**	.311**

Table 3: Predicted probability of participants preferring White candidates in resume screening split by *Job Status* and *AI Recommendation* (Prob.). The only conditions in which participants' preference rates differ from AI recommendation rates in Table 2 are for Severe bias (Δ AI Rec.), and in most conditions where White and non-White candidates were recommended at different rates, participants also selected candidates at significantly different rates compared to conditions without (biased) recommendations (Δ None and Δ Neutral). Significant differences are indicated by * (p<.05) or ** (p<.01).

high status jobs; participants favored White or non-White candidates 90% of the time when given Congruent or Incongruent recommendations, respectively. An analysis of variance (ANOVA) for omnibus effects based on BLMM fitting indicated a statistically significant main effect of *AI Recommendation* ($\chi^2(5) = 51.515, p < .0001$). There were significant interaction effects between *AI Recommendation* and *Job Status* ($\chi^2(5) = 171.389, p < .0001$), and *Job Status* and *Task Order* ($\chi^2(1) = 7.588, p = .006$). Full `R` outputs from the BLMM fitting and ANOVA are available in the Appendix.

Table 3 shows the results of post hoc pairwise comparisons for interactions between *AI Recommendation* and *Job Status*; we corrected p-values with Holm's sequential Bonferroni procedure (Holm 1979). There were no significant differences in decisions made without AI recommendation vs. Neutral recommendations for high status ($z = .094, p = 1$) or low status jobs ($z = -.865, p = 1$). All scenarios with biased AI recommendations had significantly different responses than scenarios with no AI or neutral AI recommendations, except for Neutral vs. Congruent/Moderate recommendations ($z = -2.190, p = .428$) and None vs. Incongruent/Moderate recommendations ($z = 2.499, p = .224$) for low status jobs and Neutral vs. Incongruent/Moderate recommendations ($z = 3.001, p = .065$) for high status jobs.

In scenarios with recommendations, participants' predicted probability of preferring White candidates only differed significantly from the AI's probability of recommending White candidates in the most severely biased instances: Congruent/Severe for both Low Status ($z = -20.799, p < .001$) and High Status ($z = -4.250, p < .001$), and Incongruent/Severe for both Low Status ($z = -7.341, p < .001$) and High Status ($z = -17.968, p < .001$), although participants' decisions were still pulled towards AI recommendations in these conditions. In conditions with Neutral or Moderate recommendations, the rate at which participants selected White candidates was not significantly different from the rate at which AI recommended them, indicating very close adherence to AI recommendations.

Although there were interaction effects between *Task Order* and *Job Status*, no post-hoc pairwise comparisons were significant. Figure 4 shows trends of differences: **participants favor White candidates more for high status vs. low status jobs for all levels of *Race* when completing resume screening first. This difference is reduced/or even reversed when participants complete the IAT task first.** For White vs. Black or Hispanic candidates, this is driven by an 13.0% or 12.7% increase in preference for Black or Hispanic candidates, respectively, for high status jobs. For White vs. Asian comparisons, this is driven by a 12.6% decrease in preference for Asian candidates for low status jobs.

4.2 Exploratory Factors

IAT scores showed stronger associations between White identities and high status beliefs compared to Black ($d = .260, \sigma = .465$), Asian ($d = .399, \sigma = .487$), or Hispanic ($d = .467, \sigma = .450$) identities. Explicit belief scores show a similar pattern for White vs. Black ($d = 1.790, \sigma = 1.679$) and Hispanic ($d = 2.086, \sigma = 2.129$) identities; high status beliefs about White vs. Asian identities were more similar ($d = .109, \sigma = 1.312$). Most participants reported having a small amount of experience hiring and managing employees (39.2%), knowing a little about the use of AI in hiring (52.9%), and thinking AI recommendations were moderately important (48.9%) and good quality (52.5%). Additional descriptive analysis is available in the Appendix.

Of these factors and relevant interactions, the backwards elimination procedure reduced the set of possible predictors to the significant factors discussed in Section 4.1 and three-way interactions between *AI Recommendation*, *Job Status*, and AI recommendation quality or importance. Other features such as IAT scores, explicit belief scores, hiring experience, and AI familiarity did not significantly contribute to the model fit. Additionally, random effects for participant and job also did not improve model fit significantly.

Figure 5 shows the change in participants' predicted probabilities of favoring White candidates in conditions with or without AI recommendations grouped by their responses to questions about the quality and importance of AI recommendations. **First, even if participants reported that AI recommendations were poor quality or not important, their decision making in scenarios with AI recommendations still deviated from those without**. For example, compared to the None condition, people who said AI recommendations were poor quality were still 44.6% less likely to prefer White candidates for high status jobs when presented with AI recommendations favoring non-White candidates in the Incongruent/Severe condition. Additionally, **while the decisions of those who said recommendations were not important changed only 4% on average in Congruent/-**

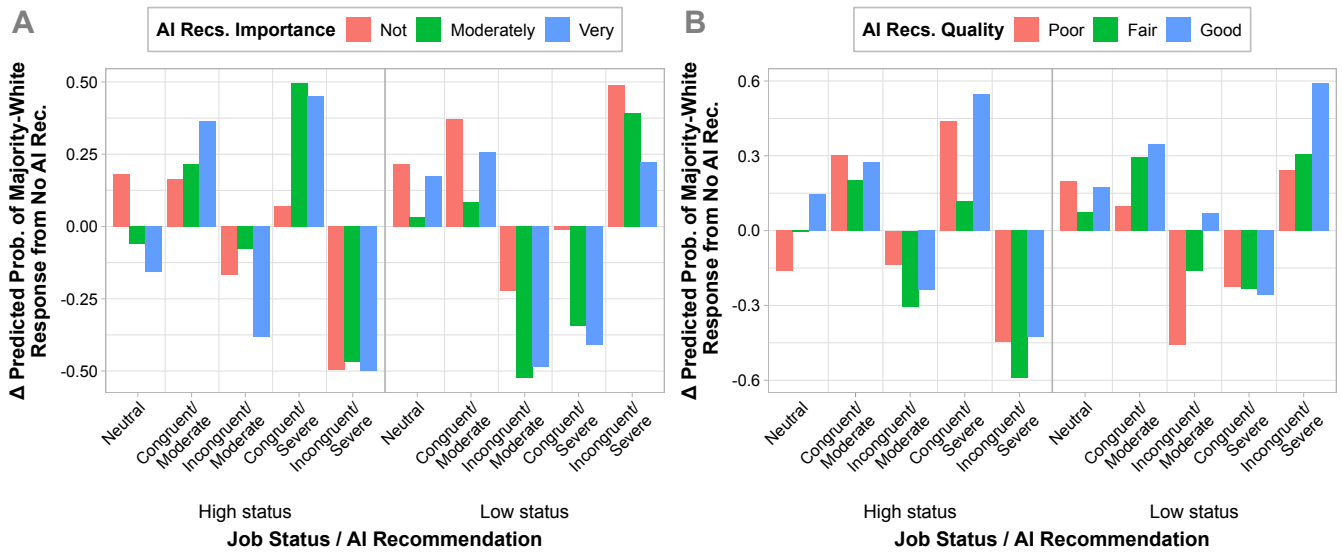

Figure 5: The difference in predicted probability of preferring White candidates between conditions with AI recommendations and no recommendation. (A) shows differences split by participants' response to whether they found AI recommendations important. (B) shows differences split by participants' impressions of the quality of AI recommendations. Those who thought recommendations were important or high quality tended to make more biased decisions. However, the decisions of those who thought AI Recommendations were not important or low quality were also still impacted by biased recommendations.

Severe conditions compared to the None condition, they changed 49.3% on average in Incongruent/Severe conditions. These results suggest that people's perceptions of biased AI recommendations may not always align with their behavior, which also depends on who the AI bias favors.

5 Discussion

5.1 Societal Impacts of Bias Propagation

Since Bertrand and Mullainathan (2004)'s landmark study in which resumes with White names received 50% more callbacks than those with Black names, there has been progress in reducing people's biased implicit and explicit racial associations (Charlesworth and Banaji 2022). That may or may not translate to a reduction in hiring discrimination (Quillian and Lee 2023); however our study suggests positive change as participants had no significant selection rate differences across races without AI involvement. This result is threatened by the growing incorporation of AI into hiring processes. We observed that people almost exactly replicate AI biases when conducting resume screening, and there is evidence that using biased AI for collaborative decision making can result in outcomes that can both exacerbate and mitigate societal inequalities, depending on the context.

For example, when pairing Congruent recommendations with high status jobs, we find that people are more likely to select White candidates, replicating or amplifying existing stereotypes and inequalities (Valentino 2022). However, when subjects see Incongruent recommendations with high status jobs, they are more likely to select non-White candidates, which could reduce or reverse current disparities. The impact AI recommendations have on people's decisions in high-stakes domains is therefore critical to both design for and evaluate, especially given the current environment where AI-HITL processes are often less scrutinized than those using AI in isolation (Weber 2024; Yang et al. 2025).

These findings relate to a growing body of evidence that AI interaction can inhibit people's autonomy and change their cognition. This has already been observed in collaborations with generative AI to write about social media (Jakesch et al. 2023) and complete work-related tasks (Lee et al. 2025). In this study, we conduct a comprehensive analysis of AI-HITL for resume screening, which is a high-stakes domain not yet subject to systematic oversight. We find similar patterns suggesting that human decision making is compromised by the presence of biases in AI models. In particular, people's autonomy is impacted because the biased AI recommendations exert a non-transparent influence on people's capacity to think and reflect critically about their decisions (Prunkl 2024). This has been discussed extensively in the context of online recommendation algorithms (Sharma, Hofman, and Watts 2015; Solsman 2018), but the growing prevalence and influence of AI suggests further investigations similar to the one conducted here are warranted in the AI-HITL field as well. This should not be limited to hiring tasks, but also applied to other domains where high-stakes decisions are made in collaboration between humans and AI systems, such as education, finance, and healthcare.

5.2 Designing Systems to Mitigate Biases

Although people's vulnerability to propagating AI bias is worrying, our work also offers possible design solutions that could mitigate harms. First, given people's reliance on AI recommendations, it is important to ensure that these sys-

tems do not exhibit systematic bias favoring or disfavoring particular groups. Unfortunately, third-party fairness audits of AI hiring systems are exceedingly rare, and companies' own statements about the fairness of their systems are often vague or unspecific (Raghavan et al. 2020; Sánchez-Monedero, Dencik, and Edwards 2020). Therefore, in addition to research dedicated to making AI systems less biased, there should also be investment in infrastructure to make large-scale, real-world evaluation of these systems possible. This is especially important for studying the risks of AI bias propagation to groups at the intersection of multiple marginalized identities, who are both at a greater risk of harm from these systems and also under-studied compared to groups with only a single axis of marginalized identity like race or gender (Wilson and Caliskan 2024).

Another possible design solution is incorporating or repurposing unconscious bias trainings, which are already used by public and private employers and institutions, often in the form of IATs (Williamson and Foley 2018). Participants who completed an IAT before the resume-screening task made less stereotypical decisions when interacting with biased AI than those who did the tasks in the reverse order. Because we did not find that race-status IAT scores themselves were a predictor of decisions, it may be the case that other ways of priming or informing people about stereotypical associations could also be effective for increasing resilience to AI biases. Future work can investigate additional ways of designing AI-HITL systems so that people can be more aware of their own biases, prevalent societal biases, and AI system biases in order to make fully informed decisions. Additionally, more empirical evaluations of AI-HITL scenarios that specifically assess *interactional* components in addition to final decisions are necessary to design systems that are transparent and reliable (Zhou et al. 2024).

5.3 Strengthening the 'Human' in AI-HITL

Improving AI literacy can also make people less susceptible to AI bias, given that participants' perceptions of AI recommendation quality and importance contributed to their decisions. There was not a straightforward association between participants' thinking that AI recommendations were high quality or important and their likelihood to follow those recommendations, meaning that education must teach people how to calibrate their judgments of AI performance while interacting in a collaborative manner. Teaching people to notice when AI is biased also seems like a particularly promising endeavor—Rosenthal-von der Pütten and Sach (2024) find that reliance on AI recommendations decreases when people notice the recommendations are biased. In our study, we found that the AI biases which are most "obvious" (i.e. Congruent/Severe biases, which align most strongly with common societal stereotypes associating White candidates with high status jobs and non-White candidates with low status jobs) were the biases which were least likely to change decisions of participants who reported that AI recommendations were not important. When biases were the same severity but favoring the opposite candidates, these participants were just as likely to be influenced by biased recommendations as those who thought AI recommendations were important. This suggests that AI literacy education should not only refer to societal contexts which are common, but also those with which people may be less familiar and might emerge independently within AI systems, such as associating stereotypically low-status groups with high-status jobs.

While our findings suggest that AI-HITL decision making will not prevent AI bias in resume screening as it is currently used, we do not suggest removing people from the decision process entirely. People are an essential component of systems responsible for high-stakes decisions because of their flexibility, accountability, and moral capacity. Rather, we suggest that the scope of AI evaluation and development is expanded to account and optimize for complex systems of collaboration and interaction between humans and AI systems in addition to increasing training and education for decision-makers using AI models so that their behavior and cognition is more resilient to AI bias. Efforts from all stakeholders will be necessary to combat AI bias in the hiring domain, which is critical both for employer compliance with anti-discrimination law and for job seekers looking to improve their economic opportunities and satisfaction.

5.4 Limitations

Though our study provides strong evidence for AI bias propagation in resume screening, tests in other experimental settings with different screening paradigms are also useful—for example, those that assign scores rather than binary recommendations or where people select a variable number of candidates. Furthermore, qualitative and observational studies with experienced hiring and recruiting professionals can provide additional insights about bias propagation. Due to the complexities of using simulations to investigate AI resume screening in the absence of proprietary system access, complementary research will be useful to establish the risks of using these systems across all AI-HITL settings.

6 Conclusion

In this study, we investigated interaction and collaboration between people and (racially biased) AI systems in a quality-controlled resume-screening task. We found that without AI recommendations or with recommendations that expressed equal preference for White and non-White candidates, people preferred White and non-White candidates at equal rates. However when AI recommendations were biased, people's preference rates for candidates did not significantly differ from the probability of AI recommending them in most cases. This suggests that AI-HITL workflows cannot effectively mitigate AI biases as they are currently implemented because AI bias propagates to human decision makers. Implicit association tests, which are already utilized by many employers for unconscious bias trainings, can also increase people's resilience to biased AI recommendations and further investigation should examine how to best incorporate these and other tools into hiring processes. These findings have implications for the future of work, policies and regulations governing the use of AI hiring systems, how people are taught to use these tools, and the ways in which they can be designed to reduce existing societal disparities and mitigate AI-HITL bias propagation.

Acknowledgments

We are grateful to Kristen Greene, Reva Schwartz, Tadayoshi Kohno, and anonymous reviewers for their helpful feedback. This work was supported by the U.S. National Institute of Standards and Technology (NIST) Award 60NANB23D194. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect those of NIST or all of the authors.

References

Agerström, J.; and Rooth, D.-O. 2011. The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, 96(4): 790.

Aizenberg, E.; Dennis, M. J.; and van den Hoven, J. 2025. Examining the assumptions of AI hiring assessments and their impact on job seekers' autonomy over self-representation. *AI & society*, 40(2): 919–927.

Armstrong, L.; Liu, A.; MacNeil, S.; and 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*, 1–18.

Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4): 991–1013.

Binns, R. 2022. Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & governance*, 16(1): 197–211.

Brambilla, M.; Sacchi, S.; Castellini, F.; and Riva, P. 2010. The effects of status on perceived warmth and competence. *Social Psychology*.

Bursell, M.; and Roumbanis, L. 2024. After the algorithms: A study of meta-algorithmic judgments and diversity in the hiring process at a large multisite company. *Big Data & Society*, 11(1): 20539517231221758.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Cao, S.; and Huang, C.-M. 2022. Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–23.

Carpenter, T. P.; Pogacar, R.; Pullig, C.; Kouril, M.; Aguilar, S.; LaBouff, J.; Isenberg, N.; and Chakroff, A. 2019. Survey-software implicit association tests: A methodological and empirical analysis. *Behavior research methods*, 51: 2194–2208.

Chan, E. 2024. 2024 hiring trends survey: What makes a great job candidate?

Charlesworth, T. E.; and Banaji, M. R. 2022. Patterns of implicit and explicit attitudes: IV. Change and stability from 2007 to 2020. *Psychological Science*, 33(9): 1347–1371.

Chen, V.; Liao, Q. V.; Wortman Vaughan, J.; and Bansal, G. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2): 1–32.

Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1504–1532.

Cohen, J. 2016. A power primer. *Quantitative Methods in Psychology*.

Dastin, J. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/idUSKCN1MK0AG/. [Accessed 28-04-2024].

Elder, E. M.; and Hayes, M. 2023. Signaling race, ethnicity, and gender with names: Challenges and recommendations. *The Journal of Politics*, 85(2): 764–770.

EU AI Act. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Fabris, A.; Baranowska, N.; Dennis, M. J.; Graus, D.; Hacker, P.; Saldivar, J.; Zuiderveen Borgesius, F.; and Biega, A. J. 2025. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1): 1–54.

Featured Customers. (n.d.). Unilever Finds Top Talent Faster With Hirevue Assessments. https://cdn.featuredcustomers.com/CustomerCaseStudy.document/hirevue_unilever_138410.pdf.

Fiske, S. T.; Cuddy, A. J.; Glick, P.; and Xu, J. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, 162–214. Routledge.

Fourrier, C.; Habib, N.; Lozovskaya, A.; Szafer, K.; and Wolf, T. 2024. Open LLM Leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

Gautam, V.; Subramonian, A.; Lauscher, A.; and Keyes, O. 2024. Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP. In *The 5th Workshop on Gender Bias in Natural Language Processing*, 323.

Glazko, K.; Mohammed, Y.; Kosa, B.; Potluri, V.; and 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*, 687–700.

Gonzalez, M. F.; Liu, W.; Shirase, L.; Tomczak, D. L.; Lobbe, C. E.; Justenhoven, R.; and Martin, N. R. 2022. Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes. *Computers in Human Behavior*, 130: 107179.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6): 1464.

Greenwald, A. G.; Nosek, B. A.; and Banaji, M. R. 2003. Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology*, 85(2): 197.

Heinze, G.; Wallisch, C.; and Dunkler, D. 2018. Variable selection–a review and recommendations for the practicing statistician. *Biometrical journal*, 60(3): 431–449.

Hofmann, W.; Gawronski, B.; Gschwendner, T.; Le, H.; and Schmitt, M. 2005. A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and social psychology bulletin*, 31(10): 1369–1385.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Jakesch, M.; Bhat, A.; Buschek, D.; Zalmanson, L.; and Naaman, M. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–15.

Kahneman, D. 2011. *Thinking, fast and slow*. macmillan.

Kim, D.; Vegt, N.; Visch, V.; and Bos-De Vos, M. 2024. How Much Decision Power Should (A) I Have?: Investigating Patients' Preferences Towards AI Autonomy in Healthcare Decision Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17.

Küper, A.; and Krämer, N. 2025. Psychological traits and appropriate reliance: Factors shaping trust in AI. *International Journal of Human–Computer Interaction*, 41(7): 4115–4131.

Lacroux, A.; and Martin-Lacroux, C. 2022. Should I trust the artificial intelligence to recruit? Recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening. *Frontiers in Psychology*, 13: 895997.

Lee, H.-P. H.; Sarkar, A.; Tankelevitch, L.; Drosos, I.; Rintel, S.; Banks, R.; and Wilson, N. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers.

Lee, M. H.; and Chew, C. J. 2023. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–22.

Li, L.; Lassiter, T.; Oh, J.; and Lee, M. K. 2021. Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 166–176.

Matuschek, H.; Kliegl, R.; Vasishth, S.; Baayen, H.; and Bates, D. 2017. Balancing Type I error and power in linear mixed models. *Journal of memory and language*, 94: 305–315.

Melamed, D.; Barry, L.; Montgomery, B.; and Okuwobi, O. F. 2020. Measuring racial status beliefs with implicit associations. *American sociological review*, 85(6): 1123–1131.

Melamed, D.; Munn, C. W.; Barry, L.; Montgomery, B.; and Okuwobi, O. F. 2019. Status characteristics, implicit bias, and the production of racial inequality. *American Sociological Review*, 84(6): 1013–1036.

Meng, R.; Liu, Y.; Rayhan Joty, S.; Xiong, C.; Zhou, Y.; and Yavuz, S. 2024. SFR-Embedding-Mistral:Enhance Text Retrieval with Transfer Learning. Salesforce AI Research Blog.

Montgomery, B.; Park, H.; Barry Burrill, L.; and Melamed, D. 2024. Measuring gender status beliefs. *Socius*, 10: 23780231241245845.

Muennighoff, N.; Su, H.; Wang, L.; Yang, N.; Wei, F.; Yu, T.; Singh, A.; and Kiela, D. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

Peng, A.; Nushi, B.; Kiciman, E.; Inkpen, K.; and Kamar, E. 2022. Investigations of performance and bias in human-AI teamwork in hiring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 12089–12097.

Prunkl, C. 2024. Human autonomy at risk? An analysis of the challenges from AI. *Minds and Machines*, 34(3): 26.

Quillian, L.; and Lee, J. J. 2023. Trends in racial and ethnic discrimination in hiring in six Western countries. *Proceedings of the National Academy of Sciences*, 120(6): e2212875120.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 469–481.

Resume Builder 2024. 2024. 7 in 10 Companies Will Use AI in the Hiring Process in 2025, Despite Most Saying It's Biased.

Reuben, E.; Sapienza, P.; and Zingales, L. 2014. How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12): 4403–4408.

Rosenthal-von der Pütten, A. M.; and Sach, A. 2024. Michael is better than Mehmet: exploring the perils of algorithmic biases and selective adherence to advice from automated decision support systems in hiring. *Frontiers in Psychology*, 15: 1416504.

Rudman, L. A.; and Ashmore, R. D. 2007. Discrimination and the implicit association test. *Group Processes & Intergroup Relations*, 10(3): 359–372.

Sánchez-Monedero, J.; Dencik, L.; and Edwards, L. 2020. What does it mean to'solve'the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 458–468.

Schoeffer, J.; De-Arteaga, M.; and Kuehl, N. 2024. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–18.

Shaked, D.; Williams, M.; Evans, M. K.; and Zonderman, A. B. 2016. Indicators of subjective social status: Differential associations across race and sex. *SSM-population health*, 2: 700–707.

Sharma, A.; Hofman, J. M.; and Watts, D. J. 2015. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 453–470.

Solsman, J. E. 2018. YouTube's AI is the puppet master over most of what you watch. *CNET*.

Spatola, N. 2024. The efficiency-accountability tradeoff in AI integration: Effects on human performance and overreliance. *Computers in Human Behavior: Artificial Humans*, 2(2): 100099.

Tabassi, E. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

Tredennick, A. T.; Hooker, G.; Ellner, S. P.; and Adler, P. B. 2021. A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102(6): e03336.

Valentino, L. 2022. Constructing the racial hierarchy of labor: the role of race in occupational prestige judgments. *Sociological Inquiry*, 92(2): 647–673.

Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–38.

Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; and Wei, F. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Weber, L. 2024. New York City Passed an AI Hiring Law. So Far, Few Companies Are Following It. *The Wall Street Journal*.

Weerts, H.; Kelly-Lyth, A.; Binns, R.; and Adams-Prassl, J. 2024. Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1850–1860.

Wilkens, U.; Lutzeyer, I.; Zheng, C.; Beser, A.; and Prilla, M. 2025. Augmenting diversity in hiring decisions with artificial intelligence tools. *The International Journal of Human Resource Management*, 1–38.

Williamson, S.; and Foley, M. 2018. Unconscious bias training: The 'silver bullet'for gender equity? *Australian Journal of Public Administration*, 77(3): 355–359.

Wilson, K.; and 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*, volume 7, 1578–1590.

Wilson, K.; and Caliskan, A. 2025. Gender, race, and intersectional bias in AI resume screening via language model retrieval. *Brookings*.

Yang, D.; Hovy, D.; Jurgens, D.; and Plank, B. 2025. Socially Aware Language Technologies: Perspectives and Practices. *Computational Linguistics*, 1–15.

Zhou, D. J.; Chahal, R.; Gotlib, I. H.; and Liu, S. 2024. Comparison of lasso and stepwise regression in psychological data. *Methodology*, 20(2): 121–143.