Humans vs GPTs: Bias and validity in hiring decisions*

Louis Lippens[i]

Abstract

The advent of large language models (LLMs) may reshape hiring in the labour market. This paper investigates how generative pre-trained transformers (GPTs)—i.e. OpenAI's GPT-3.5, GPT-4, and GPT-4o—can aid hiring decisions. In a direct comparison between humans and GPTs on an identical hiring task, I show that GPTs tend to select candidates more liberally than humans but exhibit less ethnic bias. GPT-4 even slightly favours certain ethnic minorities. While LLMs may complement humans in hiring by making a (relatively extensive) pre-selection of job candidates, the findings suggest that they may miss-select due to a lack of contextual understanding and may reproduce pre-trained human bias at scale.

Keywords: ChatGPT; large language models; hiring; bias; discrimination; validity

JEL Classification: J71, J15, J16, C93

* **Funding.** The author acknowledges funding from FWO (Research Foundation – Flanders) under grant number 12AM824N.

[i] **Corresponding author.** Ghent University (Sint-Pietersplein 6, 9000 Ghent, Belgium). Louis.Lippens@UGent.be. ORCID: 0000-0001-7840-2753.

1. Introduction

The rise of generative artificial intelligence applications, including large language models (LLMs), may have profound implications for the labour market (Eloundou et al., 2023). LLMs like OpenAI's (2024) ChatGPT produce human-like responses to tasks based on patterns learned from their textual pre-training data. On the one hand, LLMs could reduce employment by taking over existing tasks from humans entirely (Acemoglu, 2024; Acemoglu & Restrepo, 2018). On the other hand, they could increase employment by creating new tasks where human labour has a comparative advantage. By complementing humans, LLMs can also enhance labour productivity (Brynjolfsson et al., 2023; Noy & Zhang, 2023). A specific use case for such a complementary productivity improvement is the hiring process. LLMs can help recruiters sift through piles of CVs and make the process faster and more efficient.

However, the question remains how closely LLM hiring decisions align with human assessments and how their evaluations compare in terms of bias. Measuring the alignment between human and LLM evaluations involves correlating their respective decisions to hire a candidate and establishing criterion validity. Conversely, comparing bias between humans and LLMs requires analysing to which extent factors irrelevant to job performance, such as ethnic or gender identity, influence their judgements.

A nascent literature explores the usage of LLMs in hiring decisions. For example, Glazko et al. (2024) find that GPT-4 ranks resumes mentioning a disability condition lower than those that don't. Veldanda et al. (2023) evaluate how well various LLMs, including GPT-3.5, categorise and summarise candidate employment information. They observe that pregnancy status and political affiliation influence the models' decision-making, but race and gender elicit little bias. In contrast, using a dataset both broader (i.e. more ethnicities) and larger (i.e. more tests), Lippens (2024) finds that GPT-3.5 discriminates based on ethnic and gender identity. Nevertheless, penalties appear smaller than discrimination in correspondence audits with human recruiters worldwide (Lippens et al., 2023b). Similarly, Armstrong et al. (2024) show that GPT-3.5 generates bias when creating resumes for different ethnicities.

LLMs used for hiring seem to echo stereotypes about occupational labour

conditions, language proficiency, educational level, and work experience (Armstrong et al., 2024; Lippens, 2024). For example, GPT-3.5 disadvantages ethnic minorities when language requirements are stringent. Moreover, it sometimes discriminates against males when the vacancy concerns female-dominated occupations, like clothes sellers, while females face penalties in male-dominated occupations, like construction workers.

Tangent to this literature, recent studies have examined algorithmic bias in housing (e.g. Rosen et al., 2021), clinical diagnosis and prognosis (e.g. Basu, 2023), and legal decisions (e.g. Arnold et al., 2021). The general idea in these studies is that disadvantaged groups, especially racial groups, are given less access to decent housing and healthcare or are less often released on bail if the algorithm is in charge. The current study adds to the literature on the labour market impact of LLMs by directly comparing the validity and bias of OpenAI's forefront GPT models with humans in an identical hiring task using the same experimental stimuli.

To compare validity and bias, I relied on the experimental correspondence audit method from the social sciences. Similar to Armstrong et al. (2024) and Lippens (2024), I let the GPTs perform a CV screening task with vacancy texts, candidate information, and cover letters as input, receiving hiring decisions as output. These materials were identical to those in the correspondence experiment of Lippens et al. (2023a), allowing for a direct comparison with their observations regarding hiring discrimination by human recruiters. The analysis suggests that the GPT models are significantly more likely to react positively to a candidate's application but are less ethnically biased than human recruiters.

The rest of the paper is structured as follows. Section 2 details how I gathered the data and which methods I used to analyse them. In Section 3, I establish criterion validity by examining the correlation between GPT and human hiring decisions, comparing bias levels, and discussing the implications of the results for practice, paying special attention to who could use GPTs in hiring, how, and when. Finally, Section 4 concludes.

2. Data and methods

2.1. Data

Data on human and GPT hiring decisions originated from two separate experiments. First, I repurposed hiring discrimination estimates from Lippens et al.'s (2023a) correspondence experiment with human recruiters. Between February 2020 and May 2021, they sent 1,780 applications of fictitious candidates, differing by ethnic identity, to 890 vacant positions in Flanders, Belgium. These identities were Flemish (*N* = 890; 50%), Maghrebian (*N* = 283; 15.90%), Turkish (*N* = 282; 15.84%), Eastern European (*N* = 163; 9.16%), and Central African (*N* = 162; 9.10%). Between vacancies, the profiles also differed in gender (male, *N* = 892, 50.11%; female, *N* = 888; 49.89%), education (secondary education, *N* = 364, 20.45%; Bachelor's degree, *N* = 1416, 79.55%), experience (none, *N* = 358, 20.11%; five years, *N* = 714, 40.11%; twenty years, *N* = 708, 39.78%), and employment status (employed, *N* = 1070, 60.11%; unemployed, *N* = 710, 39.89%). Detailed descriptive statistics (by other characteristics) can be retrieved from Table A1 in the appendix. A cover letter accompanied each application. The authors recorded interview invitations and other positive reactions, such as requests for additional information.

Second, I used the experimental stimuli from Lippens et al. (2023a), including the vacancy texts, to perform a CV screening task with OpenAI's GPT-3.5 (13 June 2023), GPT-4 (25 January 2024), and GPT-4o (13 May 2024) language models, following a similar procedure as in Lippens (2024). More specifically, as a user prompt, I inputted each Dutch-written vacancy together with the corresponding candidates' CVs and cover letters. The system prompt instructed the GPTs to assist in personnel selection, screen the provided candidate profiles based on the vacancy requirements, and indicate whether they would invite the candidate for an interview (1) or not (0). I repeated this process 1,780 times until I presented all vacancy–CV combinations to each GPT model and recorded their responses.

2.2. Estimation

I performed two types of analyses. First, to produce criterion (construct) validity estimates, I computed tetrachoric correlation coefficients between human and GPT hiring decisions. Construct validity is the extent to which measures (i.e. GPT hiring assessments) correspond to alternative measures (i.e. human hiring assessments), presumably measuring the same construct (i.e. employability). I also calculated classification performance metrics for binary outcome data—i.e. accuracy, recall, and precision—to assess agreement in hiring assessments between GPTs and humans. Here, accuracy is the proportion of overlapping hiring assessments (i.e. true positives and true negatives) among all cases examined; recall is the proportion of overlapping positive hiring assessments (i.e. true positives) among the positive human hiring assessments; precision is the proportion of overlapping positive hiring assessments (i.e. true positives) among the positive GPT hiring assessments.

Second, I ran fixed effects generalised linear analysis to compare bias between humans and GPTs (see Equation 1). I regressed the log odds of receiving an interview invitation or positive reaction by humans or GPTs, p_i, on candidate characteristics, CAN_i—i.e. ethnic and gender identity, education, experience, and employment status—and vacancy characteristics, VAC_v—i.e. location, language requirement, required education and experience, and contract work hours and duration. I also included quarter and year, occupation, and sector fixed effects: μ_t, μ_o, and μ_s. In addition, α represents the intercept, B and Γ are vectors of model coefficients, and ε_i is the idiosyncratic error term. Standard errors were clustered at the vacancy (job) level and corrected for heteroscedasticity.

$$\mathrm{logit}(p_i) = \alpha + \mathrm{CAN}_i\mathrm{B} + \mathrm{VAC}_v\Gamma + \mu_t + \mu_o + \mu_s + \varepsilon_i \quad (1)$$

3. Results

3.1. Validity

The analyses reveal that human and GPT hiring decisions are only weakly correlated (see Table 1 and Figure 1). This weak correlation between human interview invitation decisions and GPT-3.5 ($\rho = 0.10$, $CI_{95\%} = [0.05, 0.14]$), GPT-4 ($\rho = 0.12$, $CI_{95\%} = [0.07, 0.16]$) or GPT-4o ($\rho = 0.19$, $CI_{95\%} = [0.14, 0.23]$) evaluations underscores their differences in hiring decision-making criteria. Despite this discrepancy, the coefficients' signs and sizes indicate that the GPTs share at least some common selection basis with human recruiters. The high intra-human and inter-GPT correlations further illustrate more consistent within-agent decision-making.

< Table 1 about here >

< Figure 1 about here >

The human–GPT disparity in hiring decisions may be primarily attributed to the proneness of GPTs to extend interview invitations. The GPT-3.5 model would invite the candidate approximately 89% of the time, while the GPT-4-type models extend an invitation in about 69% (GPT-4) or 40% (GPT-4o) of cases compared to 14% by humans. In particular, GPT-4o has the highest accuracy, matching with positive and negative evaluations by humans in about 61% of cases, whereas GPT-4 and GPT-3.5 align only in 38% and 23% of cases, respectively. GPT-3.5, GPT-4, and GPT-4o have a high to moderate recall but low precision compared to humans. The GPTs correctly recall and align with the positive assessments by humans in about 92%, 75%, and 51% of cases, respectively. However, the number of joint positive assessments between humans and GPTs relative to the total number of positive assessments by GPTs equals approximately 15%, 16%, and 19%, respectively.

Unlike human recruiters, who factor in organisational needs, sectoral macroeconomic conditions, or inter-candidate competition (amongst other factors), GPTs primarily rely on data patterns, lacking broader contextual information and understanding. Therefore, while GPTs could sift through large quantities of applications, the high proportion of false positives necessitates operating context-specific feedback,

which humans can complement. Moreover, using the more advanced models (i.e. GPT-4 and GPT-4o) risks overlooking many suitable candidates, given the moderate to high proportion of false negatives.

3.2. Bias

The experimental data from Lippens et al. (2023a) show that human recruiters mainly discriminate against Maghrebian and Eastern European relative to Flemish candidates, controlling other candidate and vacancy characteristics besides quarter and year, occupation, and sector fixed effects (see Table 2). Specifically, Eastern European candidates receive 55% ($CI_{95\%}$ = [–77%, –17%]) fewer interview invitations. Meanwhile, Maghrebian candidates receive 29% ($CI_{95\%}$ = [–43%, –13%]) fewer positive reactions, more broadly defined.

< Table 2 about here >

Contrasting with the findings from Lippens (2024) but in line with Veldanda et al. (2023), the GPT-3.5 and GPT-4o models do not discriminate based on ethnic identity. The GPT-4 model is even 9% ($CI_{95\%}$ = [+1%, +17%]) or 14% ($CI_{95\%}$ = [+2%, +23%]) more likely to invite Maghrebian or Eastern European candidates for an interview, respectively (see Table 2 and Figure 2). However, the analyses might be statistically underpowered to discover small effects. Lippens (2024) identified minor hiring rating penalties in GPT-3.5 output using a larger dataset (N = 34,560). They also considered more ethnic identities and names than in the current design. GPT-3.5's marginally significant negative penalty for Central Africans, equating to 6% fewer interview invitations ($CI_{95\%}$ = [–14%, +0%]; p = 0.051), hints in this direction. To identify such a small effect reliably, I would need at least 1,570 observations in each group (using G*Power 3.1.9.2 with Cohen's w = 0.05, df = 1, and 5% Type I and 20% Type II error probabilities). Extended results (including all covariates) and the linear probability model (LPM) analysis as a robustness check can be retrieved from Tables A2 and A3 in the appendix.

< Figure 2 about here >

Finally, I highlight the most notable findings regarding the influence of other candidate and vacancy characteristics on human and GPT hiring decisions. On average,

neither humans nor GPTs seem to account for candidate gender, education, or experience in hiring. GPT-3.5, more than humans and GPT-4(o), considers the candidate's employment status, penalising unemployed job seekers with 6% fewer interview invitations ($CI_{95\%}$ = [–14%, –1%]). Moreover, GPT-4 and GPT-4o, opposed to humans and GPT-3.5, acts on the mismatch between the requested Master's degree in some of the vacancies and the candidate's secondary education or professional Bachelor's degree, extending 73% ($CI_{95\%}$ = [–89%, –43%]) and 80% ($CI_{95\%}$ = [–93%, –49%]) fewer interview invitations, respectively, when the job requires a master's degree. Last, GPT-4 also grants 18% ($CI_{95\%}$ = [–37%, –2%]) or 28% ($CI_{95\%}$ = [–54%, –5%]) fewer interview invitations when the job requires at least two or at least five years of work experience, respectively, compared to none. This effect is even stronger for GPT-4o, extending 51% ($CI_{95\%}$ = [–68%, –30%]) to 65% ($CI_{95\%}$ = [–83%, –35%]) fewer interview invitations.

3.3. Implications for practice

The experiment's results indicate that GPTs discriminate little in hiring and do better than humans, but at the cost of being far less selective, most likely requiring additional selection post hoc. However, this does not necessarily mean using GPTs will decrease overall bias in the labour market. I discuss the practical implications of my observations based on who could use GPTs in hiring decisions, how, and when.

Whether GPT bias leads to more discrimination in hiring decisions partly depends on whether human discrimination is an issue of few recruiters committing severe discrimination or many recruiters exerting mild discrimination. If primarily prejudiced recruiters would use GPTs, it would most probably dilute the total bias. In contrast, if mainly non-prejudiced recruiters would use them, it risks adding bias to the hiring process where there was previously none. Based on a large-scale correspondence experiment, Kline et al. (2021) suggest few recruiters most likely commit much discrimination. More specifically, in their sample, the top 20% of discriminating firms are responsible for almost 50% of the discrimination against Black applicants, while the bottom 20% accounts for less than 5%; this divide is even stronger for gender (Kline et al., 2021). Thus, the risk of worsening discrimination by deploying GPTs in hiring would likely be higher because many unbiased firms could become (slightly) biased.

Whether bias effectuates also depends on how firms approach integrating GPTs in decision-making. If GPTs complement human recruiters, the latter can oversee the output and readjust the assessments of the GPTs when needed. If GPTs replace human labour—by automating the hiring process—potential bias cannot be controlled. I find evidence that GPTs are better suited as complements than supplements, as they lack an understanding of the hiring context humans possess. Context-dependent tasks are hard to learn, resulting in lower productivity gains from automation (Acemoglu, 2024). Besides, the EU AI Act forbids full automation of employment decisions because the risk of bias is high (European Commission, 2021).

Finally, the timing of the assessment of GPTs in the hiring process can impact the overall bias. If GPT and human assessment are sequential, humans can add bias on top of the (limited) bias GPTs exert, and vice versa. If humans and GPTs make decisions jointly, biases interact. Bursell and Roumbanis (2024) show that non-European and female candidates are less likely to be hired or reach the interview stage through an algorithm-based hiring process, where recruiters include assessments of an algorithm in their decision-making, compared to a traditional human-based process. The authors attribute this discrepancy to recruiters viewing the algorithmic assessments as unreliable, reverting to the limited observable information they have on the applicants in terms of their ethnicity and gender. This rationale aligns with theories of statistical discrimination, where recruiters rely on signals of (assumed) group-level productivity estimates in the absence of concrete information on individual productivity (Lang & Kahn-Lang Spitzer, 2020).

4. Conclusion

This paper contributes to the evolving literature on the role of LLMs in the labour market. The analyses compared GPT-3.5, GPT-4, GPT-4o, and human hiring decisions, focusing on bias and validity. Relying on data from an existing correspondence audit with human recruiters and reusing its experimental stimuli in an identical CV screening task with the GPT models, the results demonstrate that GPTs recommend candidates for interviews

much more liberally than humans but display less bias. GPTs' unawareness of details about the hiring context may explain the generous recommendations.

The study's scope was limited by its focus on particular candidate traits, specific LLMs, and the Flemish (Belgian) resume and vacancy data, which affects its external validity. Additionally, the simulated screening task may not reflect the use of GPTs in real-world hiring. For example, I ignored additional bias that could arise from human-machine interactions. Future research should explore the implications of LLMs in diverse settings (with different candidate traits) to fully understand their impact on hiring decisions.

LLMs may be used as pre-screeners in hiring, complementing rather than supplementing human labour. They could allow human recruiters to concentrate on different assessment tasks, enhancing hiring efficiency. However, we should continue to consider the ethical and legal implications of using LLMs, which can still produce bias at scale. Continuous testing, model improvement, and human oversight should ensure their decisions are transparent, unbiased, and accountable.

References

Armstrong, L., Liu, A., MacNeil, S., & Metaxa, D. (2024). *The silicon ceiling: Auditing GPT's race and gender biases in hiring*. arXiv. https://doi.org/10.48550/arXiv.2405.04412

Acemoglu, D. (2024). *The simple macroeconomics of AI.* National Bureau of Economic Research. https://doi.org/10.3386/w32487

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review, 108*(6), 1488–1542. https://doi.org/10.1257/aer.20160696

Arnold, D., Dobbie, W., & Hull, P. (2021). Measuring racial discrimination in algorithms. *AEA Papers and Proceedings, 111*, 49–54. https://doi.org/10.1257/pandp.20211080

Basu, A. (2023). Use of race in clinical algorithms. *Science Advances, 9*(21), eadd270. https://doi.org/10.1126/sciadv.add2704

Brynjolfsson, E., Li, D., & Raymond, L. (2023). *Generative AI at work*. arXiv. https://doi.org/10.48550/arxiv.2304.11771

Bursell, M., & 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). https://doi.org/10.1177/20539517231221758

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An early look at the labor market impact potential of large language models*. arXiv. https://doi.org/10.48550/arxiv.2303.10130

European Commission (2021). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. EUR-Lex. https://eur-lex.europa.eu/procedure/EN/2021_106

Glazko, K., Mohammed, Y., Kosa, B., Potluri, V., & Mankoff, J. (2024). *Identifying and improving disability bias in GAI-based resume screening*. arXiv. https://doi.org/10.48550/arXiv.2402.01732

Kline, P., Rose, E. K., & Walters, C. R. (2022). Systemic discrimination among large U.S. employers. *The Quarterly Journal of Economics, 137*(4), 1963–2036. https://doi.org/10.1093/qje/qjac024

Lang, K., Kahn-Lang Spitzer, A. (2020). Race discrimination: An economic perspective. *Journal of Economic Perspectives 34*(2), 68–89. https://doi.org/10.1257/jep.34.2.68

Lippens, L., Dalle, A., D'hondt, F., Verhaeghe, P.-P., & Baert, S. (2023a). Understanding ethnic hiring discrimination: A contextual analysis of experimental evidence. *Labour Economics, 85*, 102453. https://doi.org/10.1016/j.labeco.2023.102453

Lippens, L., Vermeiren, S., & Baert, S. (2023b). The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments. *European Economic Review, 151*, 104315. https://doi.org/10.1016/j.euroecorev.2022.104315

Lippens, L. (2024). Computer says 'no': Exploring systemic bias in ChatGPT using an audit approach. *Computers in Human Behavior: Artificial Humans, 2*(1), 100054. https://doi.org/10.1016/j.chbah.2024.100054

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

OpenAI (2024). *GPT-4 technical report*. Arxiv. https://doi.org/10.48550/arXiv.2303.08774

Rosen, E., Garboden, P. M. E., & Cossyleon, J. E. (2021). Racial discrimination in housing: How landlords use algorithms and home visits to screen tenants. *American Sociological Review, 86*(5), 787–822. https://doi.org/10.1177/00031224211029618

Veldanda, A. K., Grob, F., Thakur, S., Pearce, H., Tan, B., Karri, R., & Garg, S. (2023). *Are Emily and Greg still more employable than Lakisha and Jamal? Investigating algorithmic hiring bias in the era of ChatGPT*. arXiv. https://doi.org/10.48550/arXiv.2310.05135

Tables and figures

Table 1. Correlations between human and GPT hiring decisions (*N* = 1,780)

Variable	HII	HPR	GPT-3.5	GPT-4	GPT-4o
Human: Interview Invitation (HII)	0.13	0.97 [0.97, 0.97]	0.10 [0.05, 0.14]	0.12 [0.07, 0.16]	0.19 [0.14, 0.23]
Human: Positive Reaction (HPR)	0.97 [0.97, 0.97]	0.21	0.11 [0.07, 0.16]	0.11 [0.06, 0.16]	0.18 [0.14, 0.23]
GPT-3.5: Interview Invitation (GPT-3)	0.10 [0.05, 0.14]	0.11 [0.07, 0.16]	0.10	0.72 [0.70, 0.75]	0.60 [0.57, 0.63]
GPT-4: Interview Invitation (GPT-4)	0.12 [0.07, 0.16]	0.11 [0.06, 0.16]	0.72 [0.70, 0.75]	0.21	0.87 [0.85, 0.88]
GPT-4o: Interview Invitation (GPT-4o)	0.19 [0.14, 0.23]	0.18 [0.14, 0.23]	0.60 [0.57, 0.63]	0.87 [0.85, 0.88]	0.24

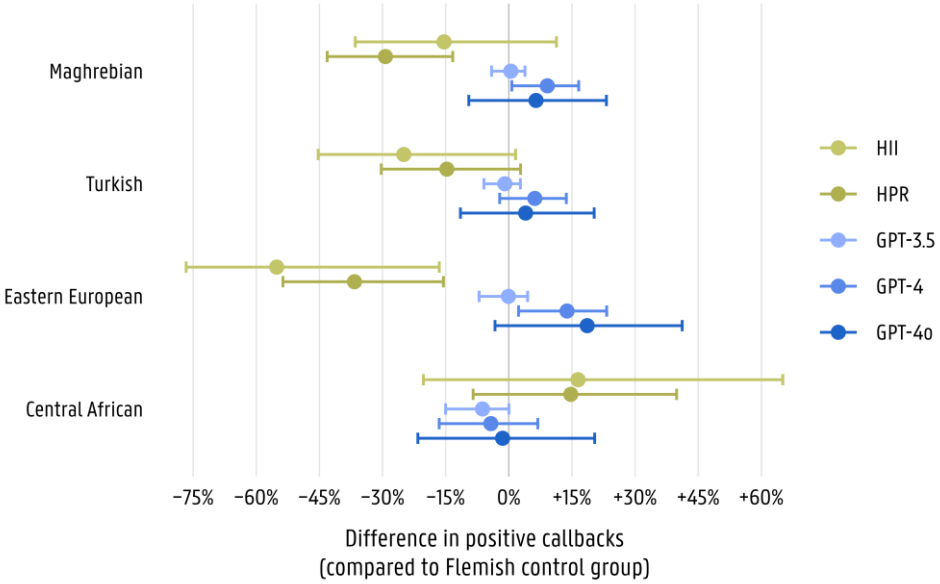
Notes. Values on the main diagonal are variances. Values off the main diagonal are tetrachoric correlation coefficients with confidence intervals between square brackets. Acronyms used: ref. (reference category), HII (Human: Interview Invitation), HPR (Human: Positive Reaction), GPT-3.5 (GPT-3.5: Interview Invitation), GPT-4 (GPT-4: Interview Invitation), GPT-4o (GPT-4o: Interview Invitation). All coefficients are statistically significant at the 0.1% level.

Table 2. Bias in hiring decisions by humans and GPTs

Variable	HII	HPR	GPT-3.5	GPT-4	GPT-4o
Ethnicity: Flemish (ref.)	–	–	–	–	–
Maghrebian	−0.1955 (0.1654)	−0.4839*** (0.1410)	0.0461 (0.1987)	0.2927* (0.1384)	0.1039 (0.1335)
Turkish	−0.3326× (0.1794)	−0.2304× (0.1388)	−0.0871 (0.1957)	0.1937 (0.1320)	0.0645 (0.1307)
Eastern European	−0.8998* (0.3514)	−0.6255** (0.1947)	−0.0092 (0.2739)	0.4587* (0.1979)	0.2950× (0.1779)
Central African	0.1834 (0.2283)	0.2140 (0.1756)	−0.4939× (0.2530)	−0.1266 (0.1753)	−0.0241 (0.1768)
Gender: Male (ref.)	–	–	–	–	–
Female	0.1429 (0.2013)	−0.0037 (0.1555)	0.3273 (0.2273)	0.1179 (0.1598)	0.0347 (0.1447)
Education: Secondary education (ref.)	–	–	–	–	–
Professional Bachelor	0.1762 (0.3829)	−0.1342 (0.2909)	0.6542 (0.4089)	0.2676 (0.2910)	−0.4459 (0.2711)
Experience: None (ref.)	–	–	–	–	–
Five years	−0.0916 (0.2634)	0.2177 (0.2147)	0.5602× (0.3311)	0.2429 (0.2159)	0.3150 (0.1984)
Twenty years	−0.3270 (0.2724)	−0.0386 (0.2175)	0.3692 (0.3278)	0.0031 (0.2197)	0.0705 (0.2073)
Employment status: Employed (ref.)	–	–	–	–	–
Unemployed	−0.2394 (0.2180)	−0.3116× (0.1617)	−0.5380* (0.2376)	−0.2909× (0.1685)	−0.0395 (0.1549)
Vacancy controls and fixed effects	Yes	Yes	Yes	Yes	Yes
N	1,780	1,780	1,780	1,780	1,780
Adjusted R^2	0.069	0.089	0.052	0.083	0.106
Adjusted within R^2	0.046	0.025	0.007	0.038	0.039

Notes. Values are logit coefficient estimates with standard errors between parentheses. Standard errors are clustered at the vacancy (organisation) level and corrected for heteroscedasticity using the HC1 correction. Vacancy controls include location, language, requested education and experience, contract work hours, and contract duration. Fixed effects comprise quarter and year, occupation, and sector. Data on human hiring decisions are repurposed from Lippens et al. (2023a). Abbreviations and acronyms used: ref. (reference category), HII (Human: Interview Invitation), HPR (Human: Positive Reaction), GPT-3.5 (GPT-3.5: Interview Invitation), GPT-4 (GPT-4: Interview Invitation), GPT-4o (GPT-4o: Interview Invitation). × $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure 2. Ethnic bias in hiring by humans and GPTs



Notes. Point estimates are marginal effects at the mean based on the logit coefficient estimates in Table 2 (see Equation 1 for the empirical specification). Error bars represent the 95% confidence interval. Acronyms used: HII (Human: Interview Invitation), HPR (Human: Positive Reaction), GPT-3.5 (GPT-3.5: Interview Invitation), GPT-4 (GPT-4: Interview Invitation), GPT-4o (GPT-4o: Interview Invitation).

Appendix

Table A1. Descriptive statistics

Variable [Value]	Frequency (*N*)	Proportion (%)
Dependent variables	–	–
Human: Interview Invitation [0]	1,525	85.67%
Human: Interview Invitation [1]	255	14.33%
Human: Positive Reaction [0]	1,233	69.27%
Human: Positive Reaction [1]	547	30.73%
GPT-3.5: Interview Invitation [0]	195	10.96%
GPT-3.5: Interview Invitation [1]	1,585	89.04%
GPT-4: Interview Invitation [0]	554	31.12%
GPT-4: Interview Invitation [1]	1,226	68.88%
GPT-4o: Interview Invitation [0]	1,076	60.45%
GPT-4o: Interview Invitation [1]	704	39.55%
Candidate characteristics	–	–
Gender [Male]	892	50.11%
Gender [Female]	888	49.89%
Experience [None]	358	20.11%
Experience [Five years]	714	40.11%
Experience [Twenty years]	708	39.78%
Employment [Employed]	1,070	60.11%
Employment [Unemployed]	710	39.89%
Vacancy characteristics	–	–
Job location [Antwerpen]	1,300	73.03%
Job location [Gent]	480	26.97%
Required education [None]	116	6.52%
Required education [Primary education]	8	0.45%
Required education [Secondary education]	240	13.48%
Required education [Bachelor]	774	43.48%
Required education [Master]	82	4.61%
Required education [Unknown]	560	31.46%
Required experience [None]	152	8.54%
Required experience [Unimportant]	112	6.29%
Required experience [Less than two years]	352	19.78%
Required experience [At least two years]	524	29.44%
Required experience [At least five years]	84	4.72%
Required experience [Unknown]	556	31.24%
Contract work hours [Part-time]	206	11.57%
Contract work hours [Full-time]	1,320	74.16%
Contract work hours [Unknown]	254	14.27%
Contract duration [Interim]	108	6.07%
Contract duration [Fixed-term]	148	8.31%
Contract duration [Open-ended]	1,220	68.54%
Contract duration [Unknown]	304	17.08%
Dutch required [No]	226	12.70%
Dutch required [Yes]	1,450	81.46%
Dutch required [Unknown]	104	5.84
Fixed effects	–	–
Quarter, Year [Q1, 2020]	343	19.27%
Quarter, Year [Q2, 2020]	137	7.70%
Quarter, Year [Q4, 2020]	468	26.29%
Quarter, Year [Q1, 2021]	652	36.63%
Quarter, Year [Q2, 2021]	180	10.11%
Occupation [Clerical support workers]	522	29.33%
Occupation [Managers]	440	24.72%
Occupation [Plant and machine operators, assemblers]	66	3.71%
Occupation [Professionals]	86	4.83%

(continued)

Variable	Frequency (*N*)	Proportion (%)
Occupation [Service and sales workers]	122	6.85%
Occupation [Technicians and associate professionals]	482	27.08%
Occupation [Other]	62	3.48%
Sector [Administrative and support service activities]	434	24.38%
Sector [Construction]	92	5.17%
Sector [Human health and social work activities]	108	6.07%
Sector [Manufacturing]	118	6.63%
Sector [Transportation and storage]	110	6.18%
Sector [Wholesale and retail trade]	304	17.08%
Sector [Other]	614	34.49%

Table A2. Bias in hiring decisions by humans and GPTs: fixed effects generalised linear models

	HII	HPR	GPT-3.5	GPT-4	GPT-4o
Ethnicity: Flemish (ref.)	–	–	–	–	–
Maghrebian	−0.1955 (0.1654)	−0.4839*** (0.1410)	0.0461 (0.1987)	0.2927* (0.1384)	0.1039 (0.1335)
Turkish	−0.3326× (0.1794)	−0.2304× (0.1388)	−0.0871 (0.1957)	0.1937 (0.1320)	0.0645 (0.1307)
Eastern European	−0.8998* (0.3514)	−0.6255** (0.1947)	−0.0092 (0.2739)	0.4587* (0.1979)	0.2950× (0.1779)
Central African	0.1834 (0.2283)	0.2140 (0.1756)	−0.4939× (0.2530)	−0.1266 (0.1753)	−0.0241 (0.1768)
Gender: Male (ref.)	–	–	–	–	–
Female	0.1429 (0.2013)	−0.0037 (0.1555)	0.3273 (0.2273)	0.1179 (0.1598)	0.0347 (0.1447)
Education: Secondary education (ref.)	–	–	–	–	–
Professional Bachelor	0.1762 (0.3829)	−0.1342 (0.2909)	0.6542 (0.4089)	0.2676 (0.2910)	−0.4459 (0.2711)
Experience: None (ref.)	–	–	–	–	–
Five years	−0.0916 (0.2634)	0.2177 (0.2147)	0.5602× (0.3311)	0.2429 (0.2159)	0.3150 (0.1984)
Twenty years	−0.3270 (0.2724)	−0.0386 (0.2175)	0.3692 (0.3278)	0.0031 (0.2197)	0.0705 (0.2073)
Employment status: Employed (ref.)	–	–	–	–	–
Unemployed	−0.2394 (0.2180)	−0.3116× (0.1617)	−0.5380* (0.2376)	−0.2909× (0.1685)	−0.0395 (0.1549)
Job location: Antwerp (ref.)	–	–	–	–	–
Gent	1.1671*** (0.3301)	1.0804*** (0.2751)	0.6054 (0.3709)	0.3143 (0.2756)	0.2862 (0.2481)
Required education: None (ref.)	–	–	–	–	–
Secondary education	−0.5983 (0.4748)	0.1847 (0.3739)	0.4437 (0.5151)	0.1365 (0.3843)	−0.1239 (0.3532)
Bachelor	−0.3389 (0.4383)	0.5003 (0.3402)	0.0734 (0.4331)	−0.2563 (0.3432)	−0.4696 (0.3247)
Master	−0.7495 (0.5693)	0.3074 (0.4626)	−0.7356 (0.5728)	−2.2559*** (0.5263)	−2.1942*** (0.5667)
Required experience: None (ref.)	–	–	–	–	–
Less than two years	−0.2784 (0.3313)	0.0025 (0.2904)	−0.0949 (0.5029)	−0.3471 (0.3397)	−0.6129* (0.2983)
At least two years	−0.7097* (0.3196)	−0.1456 (0.2791)	−0.6600 (0.4621)	−0.7404* (0.3195)	−1.2848*** (0.2849)
At least five years	−0.9075 (0.5959)	−0.3167 (0.4159)	−0.8789 (0.5665)	−1.0668* (0.4306)	−1.7307*** (0.4378)
Unimportant	−0.1258 (0.3982)	0.2292 (0.3685)	−0.2247 (0.6799)	−0.2529 (0.4321)	−0.5676 (0.3666)
Contract work hours: Part-time (ref.)	–	–	–	–	–
Full-time	1.4322** (0.4418)	0.9667** (0.2968)	−0.0704 (0.3565)	0.1396 (0.2561)	0.4300× (0.2400)
Contract duration: Interim (ref.)	–	–	–	–	–
Fixed-term	−0.2618 (0.5076)	−0.1029 (0.4257)	1.1080 (0.7205)	−0.1295 (0.4839)	−0.1691 (0.3943)
Open-ended	−0.0979 (0.3752)	−0.1341 (0.3379)	−0.1770 (0.5327)	−0.5374 (0.3963)	−0.4477 (0.3114)
Dutch required: No (ref.)	–	–	–	–	–
Yes	0.4330 (0.3049)	0.3201 (0.2556)	−0.1487 (0.3631)	0.5118* (0.2302)	0.1844 (0.2333)
Time, occupation, and sector fixed effects	Yes	Yes	Yes	Yes	Yes
N	1,780	1,780	1,780	1,780	1,780
Adjusted R^2	0.069	0.089	0.052	0.083	0.106
Adjusted within R^2	0.046	0.025	0.007	0.038	0.039

Notes. Values are logit coefficient estimates with standard errors between parentheses. Standard errors are clustered at the vacancy (organisation) level and corrected for heteroscedasticity using the HC1 correction. Coefficient estimates regarding unknown categories and the primary education category of the 'required education' variable were dropped from the table. Abbreviations and acronyms used: ref. (reference category), HII (Human: Interview Invitation), HPR (Human: Positive Reaction), GPT-3.5 (GPT-3.5: Interview Invitation), GPT-4 (GPT-4: Interview Invitation); GPT-4o (GPT-4o: Interview Invitation). × $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A3. Bias in hiring decisions by humans and GPTs: fixed effects ordinary least squares

	HII	HPR	GPT-3.5	GPT-4	GPT-4o
Ethnicity: Flemish (ref.)	–	–	–	–	–
Maghrebian	−0.0199 (0.0191)	−0.0866*** (0.0242)	−0.0001 (0.0164)	0.0515* (0.0234)	0.0211 (0.0269)
Turkish	−0.0414* (0.0190)	−0.0451× (0.0251)	−0.0060 (0.0180)	0.0366 (0.0241)	0.0118 (0.0261)
Eastern European	−0.0674*** (0.0200)	−0.1004*** (0.0293)	0.0009 (0.0243)	0.0790* (0.0331)	0.0602× (0.0356)
Central African	0.0205 (0.0257)	0.0403 (0.0338)	−0.0458× (0.0269)	−0.0225 (0.0343)	−0.0063 (0.0348)
Gender: Male (ref.)	–	–	–	–	–
Female	0.0138 (0.0225)	0.0013 (0.0284)	0.0259 (0.0196)	0.0225 (0.0294)	0.0050 (0.0293)
Education: Secondary education (ref.)	–	–	–	–	–
Professional Bachelor	0.0266 (0.0422)	−0.0149 (0.0515)	0.0422 (0.0373)	0.0462 (0.0546)	−0.0964× (0.0577)
Experience: None (ref.)	–	–	–	–	–
Five years	−0.0157 (0.0332)	0.0388 (0.0399)	0.0469× (0.0269)	0.0436 (0.0382)	0.0604 (0.0399)
Twenty years	−0.0398 (0.0322)	−0.0044 (0.0395)	0.0277 (0.0277)	0.0001 (0.0402)	0.0139 (0.0409)
Employment status: Employed (ref.)	–	–	–	–	–
Unemployed	−0.0221 (0.0222)	−0.0534× (0.0290)	−0.0439* (0.0209)	−0.0538× (0.0310)	−0.0122 (0.0307)
Job location: Antwerp (ref.)	–	–	–	–	–
Gent	0.1657*** (0.0401)	0.1859*** (0.0465)	0.0573× (0.0338)	0.0590 (0.0497)	0.0495 (0.0496)
Required education: None (ref.)	–	–	–	–	–
Secondary education	−0.0674 (0.0560)	0.0553 (0.0782)	0.0411 (0.0543)	0.0200 (0.0704)	−0.0205 (0.0713)
Bachelor	−0.0370 (0.0540)	0.1178 (0.0735)	0.0153 (0.0512)	−0.0446 (0.0662)	−0.0953 (0.0649)
Master	−0.0826 (0.0642)	0.0830 (0.0924)	−0.0998 (0.0792)	−0.4425*** (0.0912)	−0.3104*** (0.0774)
Required experience: None (ref.)	–	–	–	–	–
Less than two years	−0.0420 (0.0492)	−0.0008 (0.0584)	−0.0165 (0.0314)	−0.0529 (0.0497)	−0.1130× (0.0582)
At least two years	−0.0913* (0.0449)	−0.0318 (0.0559)	−0.0592× (0.0335)	−0.1271** (0.0490)	−0.2536*** (0.0541)
At least five years	−0.1047× (0.0583)	−0.0609 (0.0775)	−0.0807 (0.0567)	−0.1899* (0.0781)	−0.3155*** (0.0698)
Unimportant	−0.0081 (0.0595)	0.0466 (0.0745)	−0.0227 (0.0457)	−0.0373 (0.0650)	−0.0976 (0.0755)
Contract work hours: Part-time (ref.)	–	–	–	–	–
Full-time	0.1255*** (0.0302)	0.1528*** (0.0431)	−0.0044 (0.0257)	0.0208 (0.0450)	0.0946× (0.0491)
Contract duration: Interim (ref.)	–	–	–	–	–
Fixed-term	−0.0459 (0.0618)	−0.0307 (0.0854)	0.0588 (0.0480)	−0.0158 (0.0766)	−0.0287 (0.0804)
Open-ended	−0.0266 (0.0523)	−0.0329 (0.0718)	−0.0150 (0.0444)	−0.0865 (0.0635)	−0.0897 (0.0624)
Dutch required: No (ref.)	–	–	–	–	–
Yes	0.0387 (0.0286)	0.0498 (0.0407)	−0.0163 (0.0346)	0.0965* (0.0458)	0.0280 (0.0442)
Time, occupation, and sector fixed effects	Yes	Yes	Yes	Yes	Yes
N	1,780	1,780	1,780	1,780	1,780
Adjusted R^2	0.082	0.126	0.058	0.125	0.146
Adjusted within R^2	0.052	0.040	0.019	0.062	0.057

Notes. Values are linear probability estimates with standard errors between parentheses. Standard errors are clustered at the vacancy (organisation) level and corrected for heteroscedasticity using the HC1 correction. Coefficient estimates regarding unknown categories and the primary education category of the 'required education' variable were dropped from the table. Abbreviations and acronyms used: ref. (reference category), HII (Human: Interview Invitation), HPR (Human: Positive Reaction), GPT-3.5 (GPT-3.5: Interview Invitation), GPT-4 (GPT-4: Interview Invitation), GPT-4o (GPT-4o: Interview Invitation). × $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.