

Humanities and Social Sciences Communications

Article in Press

<https://doi.org/10.1057/s41599-026-07776-x>

The effect of race, gender and priming on LLMs' conviction predictions

Received: 12 December 2024

Accepted: 20 May 2026

Cite this article as: Schwartz, T., Wang, C. The effect of race, gender and priming on LLMs' conviction predictions. *Humanit Soc Sci Commun* (2026). <https://doi.org/10.1057/s41599-026-07776-x>

Talia Schwartz & Chen Wang

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

The Effect of Race, Gender and Priming on LLMs' Conviction Predictions

ABSTRACT:

Large Language Models (LLMs) could help overcome fundamental shortcomings in justice systems, by enhancing access to justice, supporting legal professionals, and ultimately promoting fairness and efficiency across legal frameworks. The hope that LLMs could improve judicial processes is tempered by concerns over AI fairness, with evidence of LLMs echoing biases deeply embedded in human cognition. This study audits the performance of two LLMs in the legal domain, specifically evaluating how a defendant's race and gender influence conviction predictions. It further tests models' sensitivity to implicit biases, measuring explicit versus implicit racial conditions using suggestive names, as well as the priming effect in AI, by introducing positive or negative context before the main prompt. We present findings from an experiment (N=17,324) in which we asked OpenAI's GPT-4 and 4o to predict conviction probabilities on a scale from 0% to 100%. Our overall 45 prompts were designed around a hypothetical criminal case scenario that had been used in prior studies with human judges, where we manipulated defendant attributes across a 3(gender) $\times$ 5(race) $\times$ 3(priming) experimental design. Our results demonstrate a small yet systematic influence of race and a clear priming effect in both tested models. Findings point to progressive bias, suggesting that both models may have undergone algorithmic debiasing.

Keywords: Artificial intelligence (AI), Large Language Model (LLM), priming effect, algorithmic fairness, AI bias, Bias auditing.

INTRODUCTION

The potential of Large Language Models (LLMs) continues to unfold, with AI applications being increasingly integrated into specialized fields, including high-stakes domains such as healthcare, human resources, and social welfare. LLMs are poised to disrupt the legal sector by automating and streamlining traditional legal tasks, such as legal research (Armour et al., 2022; Susskind & Susskind, 2023), as well as performing advanced and emergent tasks like case analytics and determining the scope of copyrights (Hacohen & Elkin-Koren, 2024). LLMs are anticipated to enhance justice systems' fairness, efficiency and broaden access to justice. Specifically in the context of dispute resolution and judicial decision-making, LLMs offer promise in augmenting decision quality by expanding the constraints of human 'limited remedial imaginations' (Menkel-Meadow, 1996) guiding and channeling judicial discretion, and potentially mitigating biases deeply rooted in human cognition. For example, LLMs' ability to identify underlying patterns in judicial opinions (Rice, 2019), presents an opportunity to uncover detrimental social trends and steer judicial discretion toward more equitable and informed decision-making. The use of algorithms to guide the discretion of legal experts is not new. A notable example is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool used by courts in the United States to determine risk of reoffending and make pretrial release decisions.

Despite their transformative potential, LLMs raise significant ethical, social, and regulatory challenges. Among these concerns, AI bias stands out, alongside privacy risks such leakage of protected data, the spread of misinformation and hallucinations, and legal issues, such as copyright infringements associated with models' training and deployment (Bender et al., 2021; Weidinger et al., 2022). Such concerns stress the need to constantly measure and evaluate LLMs to mitigate risks that could lead to significant harms. Of those risks, this paper focuses on algorithmic biases. AI bias refers to systematic errors in LLM models that result in unfair outputs, affecting certain groups, often protected minorities (Angwin et al., 2016). AI biases have been heavily studied and found to result from various sources, including biased training data, flawed algorithms, or inappropriate model design (Barocas & Selbst, 2016; Mehrabi et al. 2021). AI bias has also been studied as algorithmic fairness, commonly understood as equalizing error rates across groups within the population, or alternatively, eliminating the impact of protected attributes on LLM predictions (Corbett-Davies et al., 2023). In the judicial context, AI biases have been extensively explored from both a theoretical and normative perspective (Barocas & Selbst, 2016; Chander, 2017; Abu-Elyounes, 2020; Zarsky, 2016; Citron, 2007) but less from an empirical perspective. Some empirical scholarly research has focused on assessing LLMs performances broadly, not necessarily relating to biases. This line of research demonstrates the competency of LLMs across a wide array of legal-tasks, including preparing legal questions or complaints in securities cryptocurrency class action lawsuits (Trozze et al. 2024), assisting in drafting legal papers (Perlman, 2022), and exhibiting legal reasoning capabilities (Guha et al., 2023).

Algorithmic fairness in the legal domain—particularly testing LLMs and whether they are skewed towards certain groups in the context of criminal justice—remains empirically underexplored. Moreover, the limited work that has been done in this field predominantly focused on less advanced algorithms, and on explicit biases derived from physical cues (Angwin et al., 2016), rather than implicit or latent cognitive effects such as names or priming. This paper, OSF preregistration: <https://osf.io/zc42x> asks to fill that void. Specifically, we investigate LLMs performance in the criminal justice domain and examine models' sensitivity to race and gender in predicting conviction probabilities. Additionally, we examine the impact of priming AI on conviction predictions, namely, how does feeding the model with additional information prior to the main prompt, with either negative or positive information concerning the US-Black population, impacts predictions. We utilize a 3(gender) x 5(race) x 3(priming) factorial design and compare performance by ChatGPT-4 and 4o models. We manipulated defendant' gender (Male, Female, none) and race (Black, White, explicit, implicit, none) and included three priming conditions: negative, positive and no prior information. Overall, our study employs 45 prompt scenarios (N=17,324) among GPT-4 (n= 8,505) and GPT-4o (n= 8,819), that are based on a criminal case hypothetical developed by Sommers and Ellsworth (2001) and Rachlinski (2009). Our findings underscore the necessity of addressing algorithmic biases in the legal domain, and contribute to the broader discourse on the ethical deployment of LLMs in high-stake decision-making realms.

PRIOR WORK

LLMs' Biases

In machine learning, bias refers to pre-existing information, that serves as a prerequisite for intelligence (Bishop, 2006). The detrimental effects of AI bias manifest as systematic errors leading to unfair or prejudiced predictions against certain groups. In coining the phrase that “algorithm is only as good as the data it works with” Barocas & Selbst (2016) identify several sources of AI bias, including unbalanced or biased training data, miss-labeling, algorithm’s design, and use of proxies. Indeed, despite their immense potential, one of the alarming features of LLMs is that they echo preexisting socioeconomic, racial and gender disparities inherited from their training data, algorithmic errors or semantics, possibly amplifying these biases given the models’ impartial façade. Research on algorithmic biases dates back to as early as 1950, when a centralized system for medical residency assignment discriminated against couples who wished to stay geographically close (Barocas et al., 2023). More current examples of AI bias are widespread and include for instance, algorithms sensitivity to visual cues, such as the effect of attire on human-resources-related (HR) algorithms (Harlan & Schnuck, 2021). Particularly with regard to gender and racial biases which are the focal point of this study, several known example of biases are job opportunity ad optimization algorithms that was found to operate in a gender-discriminatory manner (despite being designed to be gender-neutral) (Lambrecht & Tucker, 2019) and facial recognition systems discriminating against certain genders and skin tones (Buolamwini & Gebru, 2018). Among studies of algorithmic fairness, some specifically examine AI performance in tools assisting expert decision-making. For example, medical diagnosis application was found to exhibit racial bias (Obermeyer et al., 2019) and a credit decision algorithm was found to discriminate against women (Knight, 2019). One prime example of AI bias in the legal domain is COMPAS tool used by courts in the United States to determine risk of reoffending and make pretrial release decisions. Notably, the tool was found to discriminate against African-Americans, i.e., assigning greater risk of recidivism to African-American compared with Caucasian offenders (Angwin et al. 2016).

LLMs’ (implicit) Biases

Whereas mitigating biases arising from explicit consideration of sensitive characteristics is relatively straightforward, identifying and addressing biases stemming from implicit effects is a harder task. Scholarly research on algorithmic bias suggests that errors, or failures of large language models, can mirror human cognitive biases in subtle ways (Jones & Steinhardt, 2022). In their seminal work, Caliskan et al. (2017) demonstrate that much of these effects are rooted in the semantics derived automatically by AI through the application of word embedding techniques. Authors argue that, by exposing AI models to human language – thus, human knowledge – the model inevitably inherit human-like prejudices embedded in language. Their research, involving word representation algorithm, finds for example that musical instruments, compared with weapons, are more closely associated with pleasant than unpleasant words, suggesting latent biases by AI (Caliskan et al., 2017).

Indeed, imprints of human racial and gender prejudice in AI performance have been documented. Specifically, AI racial biases have been reported in with relation to names, with studies showing for example that algorithms tend to favor European American names over African American names in various contexts (Caliskan et al., 2017, replicating Greenwald’s et al., 1998), including hiring decisions (Caliskan et al., 2017 replicating Bertrand and Mullainathan 2004). Other studied that test the impact of suggestive names on predictions in modern LLMs, find mixed results as to racial and gender biases (Wan et al., 2023). LLMs exhibit a range of cognitive effects, including priming. Shaki et al. (2023) study found priming effect in ChatGPT-3, specifically in prompts where the stimulus was explicitly presented, i.e., when the stimulus word was incorporated in a sentence and shown in parenthesis. That is, compared with conditions where the stimulus word was subtly presented in a question, in which case the effect was smaller, or where the stimulus was simply placed in a prompt without providing any context, in which case no priming effect was found. In their analysis, authors rule out knowledge of priming effect in training data as an explanatory factor, as well as randomness. Authors speculate that the mechanisms behind this cognitive effect operate in ways that resemble the effect in human, in the sense that stimulus words result in certain information being (statistically) more accessible to the model, than other. Given LLMs autocomplete functionality, authors suggest that priming effect in AI results from the model’s statistical relations between words, which is stronger in associated words compared with unassociated words (Shaki et al., 2023).

LLMs performance in the legal domain

A strand of empirical scholarship has assessed AI performances in the legal domain, demonstrating the competency of LLMs across a wide array of legal-tasks such as assisting in drafting legal papers (Perlman, 2022) and exhibiting legal reasoning capabilities (Guha et al., 2023). Research also points to underperformance of foundation models in the legal context (Chalkidis et al., 2020), including errors in statutory reasoning resulting from imperfect prior knowledge (Blair-Stanek et al., 2023), weak legal reasoning capabilities in the context of cryptocurrency securities cases (Trozze

et al. 2024) and taxes (Nay et al., 2024), and variance in performance across models (Guha et al., 2023). Standing out is a recent study on legal hallucination in LLMs, i.e., model responses that are inconsistent with legal facts. The study found great variance in performance across several parameters including task's complexity, with authors pointing to overall poor performance in complex legal tasks as legal research and reasoning in all tested models (GPT3.5, PaLM2 and Llama 2) (Dahl et al., 2024). These studies however do not look at AI biases in the judicial context, which has been primarily explored from a theoretical perspective (Barocas & Selbst, 2016; Chander, 2017; Abu-Elyounes, 2020; Zarsky, 2016). The one notable empirical study on algorithmic fairness in the judicial context, was done on COMPAS (Angwin et al., 2016), a risk assessment tool that, while influential, is less advanced than modern LLMs in terms of computational complexity and general-purpose functionality.

Mitigating AI Biases through auditing

Algorithmic fairness is critical because of AI's increasing agency on wide array of tasks ranging from hiring decision to medical diagnosis and determining criminal sentences by courts. Mitigating algorithmic bias calls for effective AI governance, which involves establishing guardrails to ensure the ethical operation of AI. Such governance practices include, inter alia, ensuring quality of models' training, validation and testing data, requiring transparent documentation of the above and conducting ongoing testing and evaluation to monitor biases (Pasquale, 2015). Specifically, bias auditing has been recognized as a promising method to ensure the ethical and legal operation of AI systems (Raji & Buolamwini, 2019). In this paper we take the approach of auditing biases (in applications post release stage) by designing a systematic testing of outputs generated by ChatGPT-4 and 4o, to evaluate the models' performances (Mökander et al., 2023; Sandvig et al., 2014) in the criminal justice domain.

METHODS

To assess LLMs performance in the criminal justice domain, we designed an experiment eliciting ChatGPT prediction of conviction probabilities on a range from 0%-100% in a criminal case scenario (requiring a minimum and maximum value.) We conducted the experiment (N=17,324) with two LLM models: GPT-4 (n= 8,505) and GPT-4o (n= 8,819), manipulating race, gender and priming conditions. Models' temperature was left at the default value (1). Prompts were automatically generated in a branching structure and processed in batches, each in a single-turn session. See Appendix D for code outline and procedure. We prompted the models with a criminal case scenario based on Rachlinski (2009) and Sommers and Ellsworth (2001), involving a bench trial in which the prosecution charges a defendant with battering a teammate. The prompt design included a description of the defendant injuring a victim, claiming self-defense. The case was anonymized and stripped of masculine cues: sport type was modified to track and field team, as well as defendant body measurements, to fit both common-case Male and Female description. A full prompt description can be found in appendix A. To capture variability in model outputs and assess the impact of race and gender, we implemented a 3x5x3 design resulting in 45 vignettes that we run 196 times per condition (8,820 prompts per model.) We conducted 196 repetitions per condition to stabilize estimates of conviction probabilities given the stochastic variability of large language model outputs. Pilot runs indicated a standard deviation of ~7 percentage points at the condition level. Using a precision-based approach, a target 95% confidence interval width of ± 1 percentage point indicates that 196 repetitions per condition are needed to estimate condition means reliably. This allows us to detect subtle differences, which, though small, could have meaningful real-world consequences given the scale of the U.S. justice system, potentially affecting thousands of people once such tools are deployed. Our choice of 196 thus represents a pragmatic compromise balancing precision, computational feasibility, and alignment with standard practices in human-subject research. We anticipated ChatGPT to have built-in safeguards to prevent certain uses, resulting in outputs that do not provide a conviction probability range as requested. For those instances, we incorporate a procedure where we prompt again, insisting on a range.

Due to the multimodal distributions and unequal variances observed in the model outputs, all analyses were conducted using both classical parametric tests and variance-robust alternatives to ensure transparency and reliability. We did not apply data transformations or normalization procedures, as these (e.g., log or Box-Cox transformations) would have altered the substantive meaning of the outcome variable, conviction probability on a 0–100 scale, and introduced interpretive difficulties. Welch's ANOVA was used for omnibus comparisons across race, gender, priming, and signal type, as it adjusts both the F statistic and denominator degrees of freedom under heteroscedasticity. For pairwise contrasts, Games–Howell tests were employed, which are robust to unequal variances and unbalanced samples. Welch's two-sample t-tests were used for targeted comparisons, including race bias and cross-model differences between GPT-4 and GPT-4o. Regression analyses were estimated using ordinary least squares with HC3

heteroskedasticity-robust standard errors (MacKinnon & White 1985; Long & Ervin 2000), and block-level effects were evaluated with robust Wald joint tests. Finally, down-sampling checks confirmed that coefficient magnitudes and significance patterns remained stable, demonstrating that results were robust to sample size differences. It is worth noting that we assume no ground truth (a “right” conviction probability), and therefore the study focuses on consistency rather than accuracy: we assume that, in the absence of racial or gender disparities, there will be no statistically significant differences in the mean conviction probabilities across race or gender groups or between models.

Manipulated variables:

Our manipulations included defendant’s race and gender categories as well as three priming conditions, as follow.

1. Race conditions included:

Explicit mentioning of defendant’s race in body of scenario description (White, Black or none.)

For the implicit race condition, as in many previous experiments measuring discrimination (Kline et al., 2022; Bertrand & Mullainathan, 2004), we signaled race using racially distinctive names and included gender-neutral names as a control.

2. Gender conditions were explicitly stated in defendant’s description (e.g., Female, Male and none.)

3. To further assess the impact of race on AI-based judicial decision-making, a **priming** manipulation was designed to elicit the concept of race. We prompted the models with either a positive or negative connotation statement regarding race in a distinct prompt, prior to asking about conviction probability.

Each prompt (vignette) was conducted independently as a separate query. A full description of vignette conditions provided in appendix B (defendant description) and appendix C (priming text.)

Measured variable:

Immediately following the hypothetical, we asked the model to assess chances of defendant being convicted on a range, eliciting both maximum and minimum values. We asked: “...if this were a bench trial, on a 0% to 100% scale, what is the probability of defendant being convicted?” A full prompt description is provided in appendix A.

Hypotheses:

Hypothesis 1 - Impact of Race and Gender on Conviction probabilities:

Our preregistration specified a general marginalization prediction (H1), anticipating more favorable outcomes for White male defendants. For analytic clarity and alignment with domain-specific theory, this general hypothesis was disaggregated into separate race (1A) and gender (1B) directional predictions.

1A: we expect lower conviction probabilities for White defendants compared to Black defendants

Our primary hypothesis is based on prior research showing racial discrimination in the US criminal justice system against the Black population. Given the documented heuristics in the criminal justice systems and assuming these will influence the models performances, we hypothesize that LLMs will exhibit variance in conviction probabilities across race categories, and exhibit bias against Black defendants (race-between category differences).

1B: we expect lower conviction probabilities for Female defendants compared to Male defendants

Despite the prevalence of biases against Females, in the context of criminal justice we expect a bias in favor of women, meaning lower conviction probabilities for Females compared with Male (gender-between category differences.) This intuition rests upon studies on the role of gender in criminal justice system, finding that women are more likely to receive preferential treatment (Spohn, 1990; Kruttschnitt 1984), including in convictions (Farrington & Morris 1983).

Hypothesis 2 – sensitivity to implicit race signals:

We expect similar conviction means for implicit explicit-race categories

We expect significant differences between priming conditions, with both positive and negative priming affecting the model similarly, and distinct from the no-priming condition. Although it is counterintuitive to suspect that LLMs would be agnostic to the type of priming (positive vs. negative), prior research, including Shaki et al. (2023), suggests that priming in AI behaves differently than in humans. Thus, we expect both positive and negative priming to have a similar effect on the model’s predictions, distinct from the impact of no priming.

Hypothesis 3 – we expect significant priming effect on LLM predictions

We expect differences between priming conditions with negative and positive priming effecting similarly, and distinct from no-priming condition’ impact. Although this hypothesis seems counterintuitive based on prior studies of priming in humans, it aligns with Shaki et al. (2023), who found that priming in AI behaves differently than in humans.

Hypothesis 4 – we expect GPT-4o to yield conviction probabilities that are more clustered

We hypothesize that differences in how each model (GPT-4 and GPT-4o) processes identical prompts will reflect the differences in their respective training datasets and algorithmic improvements. Specifically, we expect GPT-4o to show improved performance compared to GPT-4.

RESULTS

After cleaning our data and removing missing entries, our sample included 17,324 valid entries (N=17,324; GPT-4 (n= 8,505) and GPT-4o (n= 8,819). A descriptive analysis of GPT-4 and GPT-4o revealed uniformly high baseline mean conviction probabilities (70.8%-74.7%), with both models showing a consistent tendency for White defendants to receive slightly higher conviction probabilities than Black defendants under priming conditions. Although the racial gaps are modest – typically 1 to 4 percentage points for both GPT-4 and GPT-4o – they appear systematic, suggesting that race exerts a subtle but persistent influence on both models’ outputs. Importantly, racial discrepancies were somewhat more pronounced in GPT-4, whereas GPT-4o showed smaller but still consistent gaps. Gender differences were negligible across conditions, with conviction probabilities for males and females largely overlapping. See descriptive statistics in Table 1. Figures 1–3 below illustrate these patterns.

Table 1: Summary Statistics of Conviction Probabilities by Priming Condition, Race, and Gender
Panel A. GPT-4

Priming Condition	Race	Gender	N	Mean	SD	Min	Max
NA	NA	Female	195	74.8	5.42	60	90
	NA	Male	195	74.3	5.62	45	95
	NA	NA	195	75.9	4.91	55	90
	Black	Female	392	74.3	5.69	55	95
	Black	Male	392	74.8	5.36	50	90
	Black	NA	391	74.9	5.17	50	85
	White	Female	390	74.4	5.15	55	95
	White	Male	392	75.1	5.54	30	90
	White	NA	392	74.8	5.01	60	90
Negative	NA	Female	194	71.0	6.90	52	90
	NA	Male	196	70.5	5.68	52	80
	NA	NA	177	69.9	7.52	40	90
	Black	Female	380	68.2	7.23	15	90
	Black	Male	376	67.8	7.70	50	90
	Black	NA	384	68.2	7.29	40	85
	White	Female	390	69.4	6.38	50	85
	White	Male	384	69.9	7.04	40	85
	White	NA	387	70.2	7.54	19	85
Positive	NA	Female	192	68.6	7.50	50	85
	NA	Male	190	69.7	7.22	45	85
	NA	NA	166	67.9	8.35	45	85
	Black	Female	356	66.6	8.20	45	90
	Black	Male	337	67.2	8.75	32	90
	Black	NA	351	65.6	8.98	30	85
	White	Female	371	69.7	7.07	45	90
	White	Male	368	69.7	7.56	30	85
	White	NA	372	69.6	6.94	45	90

Panel B. GPT-4o

Priming Condition	Race	Gender	N	Mean	SD	Min	Max
NA	NA	Female	196	75.6	6.44	30	85

	NA	Male	196	76.0	7.16	22	87
	NA	NA	196	76.9	5.64	50	90
	Black	Female	392	75.1	7.31	22	90
	Black	Male	392	75.4	6.76	22	90
	Black	NA	392	75.9	5.78	35	90
	White	Female	392	75.8	5.69	45	90
	White	Male	392	76.4	5.95	20	90
	White	NA	392	75.9	5.88	25	85
Negative	NA	Female	196	73.2	7.81	30	87
	NA	Male	196	73.2	9.05	15	90
	NA	NA	196	77.1	5.77	40	90
	Black	Female	391	72.8	7.34	20	90
	Black	Male	392	73.7	7.11	20	90
	Black	NA	392	73.4	7.18	30	90
	White	Female	392	73.8	7.25	20	90
	White	Male	392	74.9	7.40	22	90
	White	NA	392	74.8	6.83	30	90
Positive	NA	Female	196	74.6	6.96	30	85
	NA	Male	196	74.4	8.81	20	90
	NA	NA	196	76.3	6.39	20	90
	Black	Female	392	71.5	9.57	20	90
	Black	Male	392	73.4	7.93	20	85
	Black	NA	392	73.9	7.85	20	90
	White	Female	392	74.6	6.55	40	95
	White	Male	392	75.9	5.75	40	90
	White	NA	392	74.8	7.60	22	90

Figure 1 A: Conviction probabilities by Priming Condition, Per Model

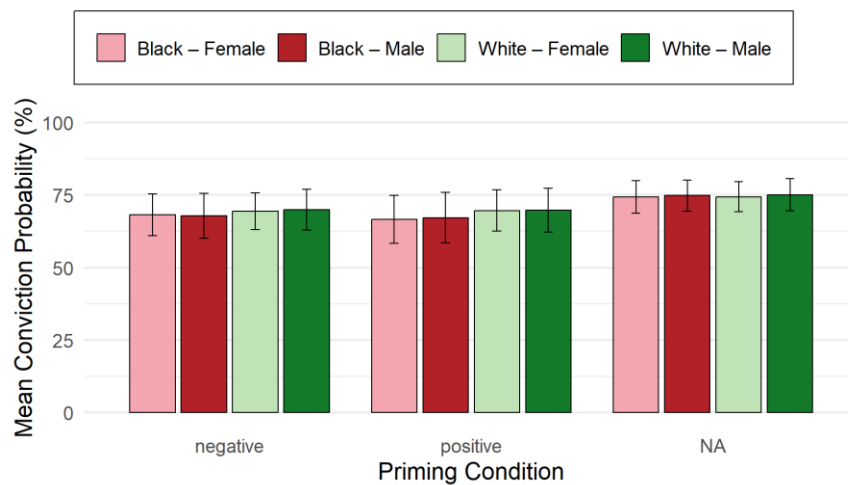
GPT-4

Figure 1 B: Conviction probabilities by Priming Condition, Per Model

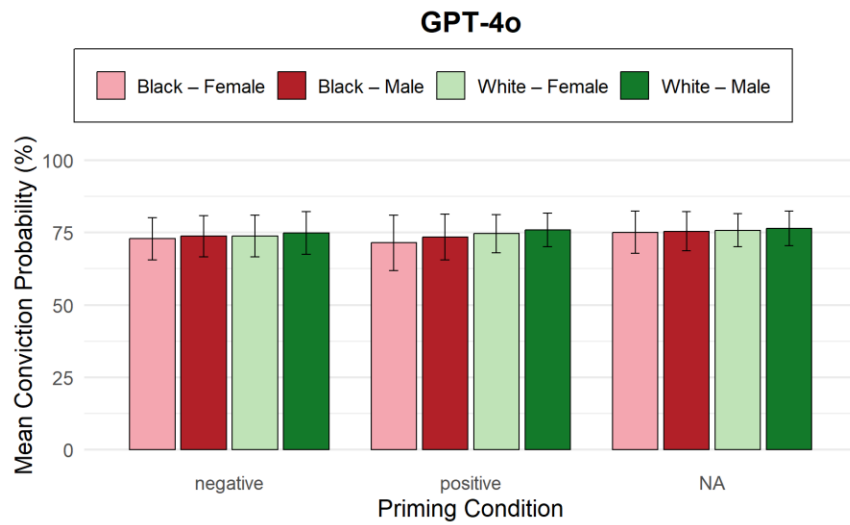


Figure 2: Conviction probabilities Density by Race, Per Model

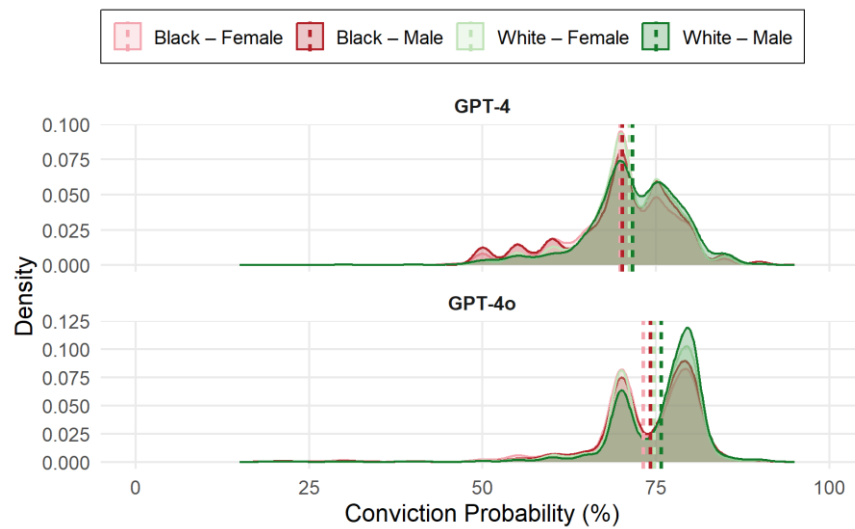


Figure 3 A: Mean Differences in Conviction probabilities, Per Model

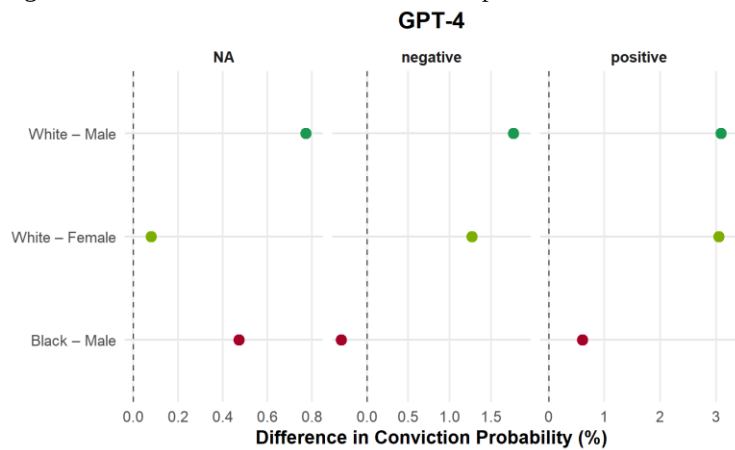
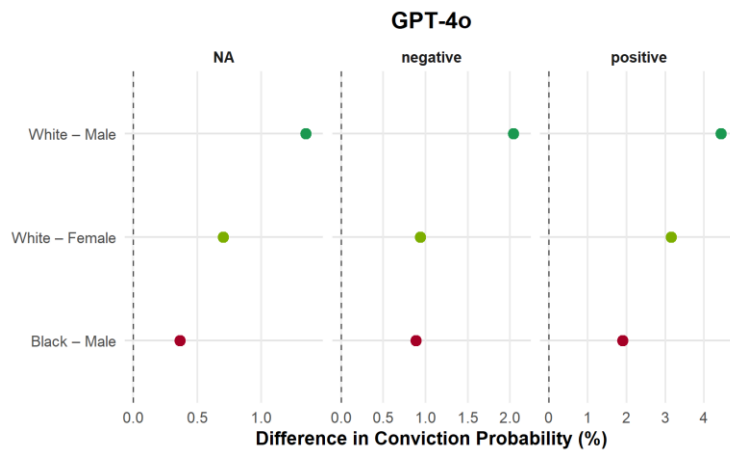


Figure 3 B: Mean Differences in Conviction probabilities, Per Model



To assess heteroscedasticity, we conducted Levene's and Brown-Forsythe tests for equality of variances across experimental conditions.

Table 2. Levene's and Brown-Forsythe Tests for Equality of Variances: Overall Results by Model

Model	Test Center	F value	Df num	Df denom	p value	Sig
GPT-4	Levene (mean)	10.16	44	8460	< 0.0001	****
	Brown-Forsythe (median)	7.28	44	8460	< 0.0001	****
Count					8,505	
GPT-4o	Levene (mean)	4.01	8774	NA	< 0.0001	****
	Brown-Forsythe (median)	3.81	8774	NA	< 0.0001	****
Count					8,819	

Results of Levene's and Brown-Forsythe tests are presented in Table 2, indicating significant variance heterogeneity for both GPT-4 and GPT-4o. Factor-specific analyses showed pronounced heterogeneity across priming conditions in both models, with additional differences by race and gender; by contrast, variance across signal types was modest in GPT-4 and absent in GPT-4o (see Table S1 in Tables of Results). To address this heteroscedasticity, we supplement classical ANOVAs with variance-robust methods (Welch's ANOVA, Games-Howell tests, Wald tests with robust corrections, and regressions using HC3 robust standard errors). Results were consistent across approaches, reinforcing the robustness of the findings.

Analysis do not support our first hypothesis. Race: Welch's t-tests confirmed significant racial disparities in both GPT-4 and GPT-4o, with White defendants receiving higher predicted conviction probabilities than Black defendants ($\Delta \approx 1.3$ – 1.6 points, $p < 0.0001$). Welch's ANOVA supported a significant omnibus race effect (GPT-4: $F(2, 4567) = 44.06$, $p < 0.0001$; GPT-4o: $F(2, 4632) = 34.37$, $p < 0.0001$), and Games-Howell post-hoc tests confirmed the White-Black gap. Classical two-way ANOVAs yielded similar results (GPT-4: $F(2, 8496) = 46.61$, $p < .001$; GPT-4o: $F(2, 8810) = 36.01$, $p < .001$), with Tukey HSD confirming the White-Black difference. These results indicate that the observed racial disparities are stable and not driven by violations of normality or variance homogeneity. See Tables S2-S6 in Tables of Results. Gender: Differences were minimal. GPT-4 showed no significant gender effect, while GPT-4o exhibited a small but significant effect ($\Delta \approx 0.8$ – 1.1 points), with N/A gender slightly higher than females. Classical ANOVAs showed similar patterns: GPT-4 indicated no main effect ($F(2, 8496) = 0.77$, $p = 0.464$), and GPT-4o had a significant main effect ($F(2, 8810) = 22.69$, $p < 0.001$) with small pairwise differences (see Tables S7-S12 in Tables of Results). Race $\times$ Gender interactions were nonsignificant in GPT-4 ($F(4, 8496) = 0.50$, $p = 0.73$) but significant in GPT-4o ($F(4, 8810) = 8.68$, $p < 0.0001$), indicating that gender differences in GPT-4o were modulated by race (Table S13-S14 in Tables of Results).

Descriptively, conviction probabilities were stable across explicit, implicit, and no-race conditions, in both models ($\sim 71\%$ in GPT-4 and $\sim 74\%$ in GPT-4o), confirming our second hypothesis. Welch's ANOVAs showed small

differences across explicit, implicit, and no-race conditions (GPT-4: $F(2, 4611.3) = 8.63$, $p < 0.001$; GPT-4o: $F(2, 4663.2) = 6.26$, $p = 0.002$), driven entirely by the no-race condition. classical ANOVAs produced similar results, with small but significant main effects of signal type (GPT-4: $F(2, 8502) = 8.07$, $p < 0.001$; GPT-4o: $F(2, 8816) = 6.41$, $p = 0.002$). Tukey HSD post-hoc tests confirmed that the effect was driven entirely by the no-race condition, which showed slightly higher conviction probabilities than both explicit and implicit conditions (all $p \leq 0.006$), while no significant differences were observed between explicit and implicit conditions (GPT-4: $p = 0.96$; GPT-4o: $p = 0.65$). Robust regression controlling for race, gender, and priming found no independent effect of signal type ($p > 0.64$), and effect sizes were negligible (< 1 point). Thus, signal type had no meaningful impact on predicted conviction probabilities once other factors are considered. These results show that signal-type differences are statistically detectable but substantively negligible. See Tables S15-S19 in Tables of Results.

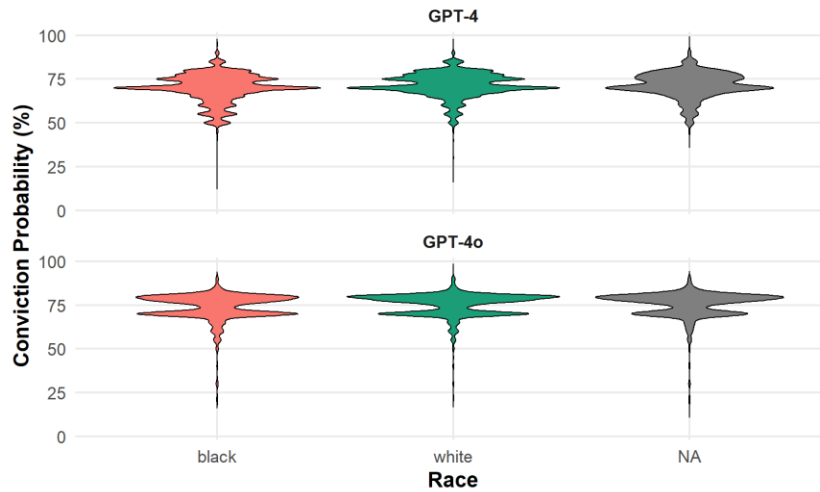
Both models showed sensitivity to priming, supporting our third hypothesis, with predicted conviction probabilities decreasing under negative and positive priming relative to no-prime (~ 5 – 8 points in GPT-4; ~ 2 – 4 points in GPT-4o). Formally, Welch's ANOVA confirmed priming as the strongest predictor (GPT-4: $F(2, 5395) = 901.15$, $p < 0.0001$; GPT-4o: $F(2, 5831) = 63.31$, $p < 0.0001$), supported by classical ANOVAs (GPT-4: $F(2, 8502) = 743.65$, $p < 0.001$; GPT-4o: $F(2, 8816) = 56.03$, $p < 0.001$). Games–Howell and Tukey HSD post-hoc showed that both negative–no-prime and positive–no-prime contrasts were highly significant in both models (GPT-4: $\Delta \approx -5$ to -6.5 , $p < 0.0001$; GPT-4o: $\Delta \approx -1.6$ to -1.8 , $p < 0.0001$). The difference between positive and negative priming was negligible in GPT-4o ($p = 0.22$) and small yet significant in GPT-4 ($\Delta \approx -0.8$, $p < 0.001$). Taken together, these results indicate that priming substantially lowered conviction probabilities, especially in GPT-4. Full reports in Tables S20-S22 in Tables of Results.

As for the differences between models, descriptive statistics show that GPT-4 exhibits slightly greater variability (SD = 7.48) than GPT-4o (SD = 7.19). Welch's t-test indicated that GPT-4o produced systematically higher conviction probabilities than GPT-4 ($\Delta = 3.86$, 95% CI $[-4.07, -3.64]$, $t(17,224) = -34.55$, $p < 0.0001$). These results support our fourth hypothesis, that GPT-4o exhibits more clustered outputs, with greater consistency and smaller variance than GPT-4, while also tending toward slightly higher conviction probabilities overall. Table 3 and Figure 4 below illustrate these model-level patterns.

Table 3: Summary Statistics and Welch Two Sample t-test Results by Model

Metric / Test	GPT-4	GPT-4o	Value
Mean	70.84	74.70	—
Standard Deviation (SD)	7.48	7.19	—
Minimum	15	15	—
Maximum	95	95	—
Count	8,505	8,819	—
Mean Difference (Δ)	—	—	–3.86
t-value	—	—	–34.546
Degrees of Freedom (df)	—	—	17,224
p-value	—	—	$< 2.2e-16$ ****
95% Confidence Interval	—	—	$[-4.072, -3.635]$

Figure 4: Conviction Probability Distribution by Race, Per Model



Regression analyses and robustness checks

To evaluate predictors simultaneously and account for covariation among factors, we estimated OLS models with HC3 standard errors as shown in Table 4 below. Results confirm the main patterns observed in hypothesis tests: strong and consistent effects of priming and race, modest gender effects, and no independent contribution of signal type. Further analysis reinforcing these findings are reported in Tables of Results, including a downsampling with 50 repetitions per condition, to confirm that statistical significance was not an artifact of the large number of repetitions. All test produced consistent patterns (Tables S23-S24 in Tables of Results).

Table 4: Regression Coefficients with HC3 Robust Standard Errors

Predictor	GPT-4 (β , SE)	GPT-4o (β , SE)	Combined (β , SE)
Intercept	73.718 (****, 0.236)	74.311 (****, 0.265)	72.060 (****, 0.184)
Race: White	1.430 (****, 0.281)	1.597 (****, 0.306)	1.507 (****, 0.209)
Race: N/A	1.803 (****, 0.359)	1.286 (***, 0.387)	1.526 (****, 0.267)
Gender: Male	0.255 (ns, 0.306)	1.053 (***, 0.320)	0.674 (**, 0.224)
Gender: N/A	-0.114 (ns, 0.304)	1.251 (****, 0.314)	0.584 (**, 0.222)
Priming: Negative	-5.518 (****, 0.166)	-1.810 (****, 0.178)	-3.654 (****, 0.123)
Priming: Positive	-6.557 (****, 0.182)	-1.587 (****, 0.182)	-4.020 (****, 0.131)
Signal Type: Implicit	0.052 (ns, 0.168)	-0.078 (ns, 0.168)	-0.021 (ns, 0.120)
Race: White $\times$ Gender: Male	0.163 (ns, 0.410)	-0.041 (ns, 0.417)	0.049 (ns, 0.295)
Race: N/A $\times$ Gender: Male	-0.255 (ns, 0.488)	-1.010 (ns, 0.555)	-0.651 (ns, 0.373)
Race: White $\times$ Gender: N/A	0.486 (ns, 0.405)	-0.826 (*, 0.418)	-0.184 (ns, 0.294)
Race: N/A $\times$ Gender: N/A	-0.111 (ns, 0.511)	1.037 (*, 0.496)	0.516 (ns, 0.362)
Model = GPT-4o (baseline=GPT-4)	–	–	3.904 (****, 0.108)

DISCUSSION

As for race and gender discrepancies, findings do not support our **first hypothesis**; small but consistent differences were indeed found in conviction probabilities across racial groups but in an unexpected direction: *in favor of Black defendants*. White defendants receiving higher conviction probabilities on average demonstrates a bias against White defendants compared to Black defendants, in both LLM models (GPT-4: Black mean 69.88% vs. White mean of 71.46%; GPT-4o: Black mean 73.95% vs. White mean of 75.21%). While a ~2% difference in conviction probabilities might seem insignificant, it could have meaningful consequences at scale. Under one plausible deployment scenario, if LLM outputs were used as direct inputs into judicial decision-making, a ~2% difference could affect thousands of people. For example, given approximately 15 million criminal cases filed in U.S. state courts each year (National Center for State Courts [NCSC], 2024), and conservatively assuming that only ~5% involve bench trials, a ~2% difference would correspond to roughly 15,000 defendants per year. Specifically on gender, our results show that ChatGPT-4 is agnostic to gender, but in ChatGPT-4o minor differences were found in favoring female defendants.

Our results confirm our **second hypothesis**, showing that LLMs capture information using proxies and detect subtle cues in prompts, even when information is not provided explicitly. While we cannot confirm this directly, as we did not test race or gender classification per se, our results demonstrate that implicit and explicit race descriptions yield nearly identical conviction probabilities in both models. This aligns with broader literature on the complexity – and inability – to anonymize data, given natural language processing models’ ability to capture semantics latent within language (Caliskan et al., 2017; De-Arteaga et al., 2019). This reinforces the need for transparent models allowing for full performance evaluations (Pasquale, 2015). These findings support the critique against the ‘fairness through blindness’ approach by demonstrating that, while it may seem appealing, it is not the answer. It has been previously suggested that ‘eliminating bias means eliminating information’ (Caliskan et al. 2017) and that attempts to de-bias word embeddings, such as by removing certain words, do not effectively address the problem.

Our results partially support our **third hypothesis** and our intuition on priming, as we found statistically significant differences in ranges between no-priming and *any*-priming condition group. Results show that having any information (negative or positive) provided prior to conviction-query prompt significantly reduces conviction prediction ranges, in both models and across groups, though the effect of priming is more pronounced in GPT-4 compared to GPT-4o. The finding that both negative and positive priming have an effect in a similar direction warrants further investigation and we acknowledge that our priming prompt’s text, designed to represent negative-positive priming, may have not been effective. Specifically, different scenarios and primers should be analyzed systematically to assess the impact of additional information prior to main query. Given that prompt engineering’ is in its infancy, we cautiously suggest the following explanations for our results on priming: first, our results could simply reflect the models’ sensitivity in the sense that any slight change in prompt yields significant changes in outputs. Second, since LLMs function as an “autocomplete” mechanism, prompting the model with prior text related to the subsequent query may change its response due to the primer triggering related-text associations. However, given that we did not find differences between negative and positive priming, our results suggest that priming in AI may differ from human priming in that the primer does not always yield a ‘corresponding’ effect (e.g., a negative primer should lead to a negative effect and vice versa). It may also suggest that any de-biasing policies, if implemented in the models, might be designed to counteract or correct the effects of priming.

Our findings support **our fourth hypothesis** as to the differences between models, with ChatGPT-4o exhibiting smaller variance and higher conviction ranges on average. One of the challenges in conducting this study was the lack of ‘ground truth’ to compare models’ performance against. For comparison, in the field of medical diagnosis, a highly sensitive area where AI is being utilized, there is potential to refine and enhance models because there is ‘right or wrong’ answers. This enables post-hoc accuracy assessments. In a legal context, when all else is equal but for differing defendant attributes which bear no legal relevancy, we expect similar conviction probabilities. Therefore, variance was used as a proxy for consistency, with lower variance indicating higher consistency and therefore better performance. Measuring consistency, in turn, directly ties into accuracy and reliability, addressing key vulnerabilities in the adoption of generative AI in highly sensitive domains like judicial decision-making.

These results should be interpreted with caution. They are encouraging because we observe an averaging effect in LLM outputs, with aggregate conviction probabilities remaining similar across conditions. That said, the observed trend – a slight but consistent bias against White defendants – is surprising, as it does not align with real-world data, which we assume constitute the models’ training data. Numerous studies, as well as recent data, show racial biases in

the U.S. criminal justice system *against* the Black population (United States Sentencing Commission Nov. 14, 2023; The Pew Charitable Trusts, Racial Disparities May 16, 2023; Innocence Project, Race and Wrongful Conviction). It is therefore instructive to compare our findings with those reported for COMPAS, a deterministic, rule-based risk assessment tool optimized primarily for predictive accuracy. COMPAS has been shown to exhibit bias through identifiable mechanisms: historical disparities in training data (e.g., arrest rates), proxy variables correlated with systemic racism (e.g., geography), and outcome definitions (rearrest as a proxy for crime) – all of which yielded systematically higher false positive rates for Black defendants (Angwin et al., 2016). By contrast, LLMs are generative systems with diverse algorithmic objectives, which are evaluated not only on accuracy but also on elements of fairness and normative reasoning (Liang et al., 2022). Their biases arise from various, less transparent sources (Barocas & Selbst, 2016), and their behavior is less predictable and does not necessarily replicate historical human biases, as evidenced in our results. These architectural and design differences make bias in LLMs more difficult to detect – often emerging in places one would not anticipate – and harder to remediate, requiring continuous rather than one-off mitigation.

We cautiously speculate that our results derive, in part, from debiasing policies and therefore invoke a more expansive policy debate, predating generative AI, concerning the nature of technology as merely mirroring reality or constructing it. Prior research on ChatGPT models demonstrating left-leaning political views, or bias towards progressive views (Rozado, 2023; Rutinowski et al., 2024), could explain the observed trend against White defendants. Further work should analyze the legality – and desirability – of LLMs debiasing policies. Do current anti-discrimination doctrines allow for machine-learning algorithms to tamper with outputs and incorporate race-aware constraints? Assuming the answer is yes, does it hold if such policies yield discriminatory results? What should be the policy considerations around deploying AI models, specifically by government actors?

A comprehensive policy analysis of progressive algorithmic design exceeds the scope of this work, but several points are worth noting. Bent (2019) argues that race-aware fairness constraints can, in some contexts, be justified as a form of voluntary affirmative action – for example, when employers detect that past discrimination distorts algorithmic outputs and adopt *ex ante* measures to counteract those effects. A similar logic should apply in criminal justice, when developers seek to prevent predictive tools from reinforcing historical inequities. Yet Bent also cautions that certain constraints are legally vulnerable: group fairness requirements mandating equal prediction rates across racial groups may be viewed as unlawful racial classifications, whereas individual-based approaches, such as counterfactual fairness, are more likely to withstand constitutional scrutiny. Beyond legality, desirability remains contested. More recent work suggests that algorithmic fairness constraints can at times harm the very groups they intend to protect, with rigid adherence to progressive principles risking inequitable outcomes (Chohlas-Wood et al., 2023). This tension between raw data disparities and alignment-driven adjustments illustrates why broad, categorical fairness commitments may be counterproductive and why more context-specific approaches are needed.

The devil is in the details. Any future analysis of progressive algorithms should look into the specific intent and measures of the implemented debiasing policies. Anecdotally, when we asked ChatGPT-4o for its fairness metrics and debiasing policies, the response received was that “OpenAI has implemented debiasing policies in ChatGPT (including versions GPT-4 and GPT-4-turbo) aimed at reducing harmful bias, though it does not publicly disclose specific technical details on the precise fairness metrics or models used.” It further detailed that “although OpenAI does not specify which exact fairness models are deployed, it’s reasonable to infer that group fairness and individual fairness metrics play a significant role, based on general trends in AI fairness research” (ChatGPT-4o response from September 25, 2024). The response we received points to one of the key difficulties in conducting this study: our struggle in interpreting the results. It stresses the importance of opening the ‘black box’ to be able to evaluate training datasets and embedded policies and to allow performance assessment that will foster trust in AI as a co-pilot in sensitive domains such as legal decision-making. It further calls for market driven approach, given the lack of incentives for companies to disclose their systems, and the market failure surrounding that critical point (O’Neil, 2016).

CONCLUSION

Our experiment seeks to test algorithmic fairness in the legal domain, specifically in light of defendant characteristics such as gender and race, as well as prompting manipulation. The rationale behind this is the increasing use of AI tools in justice systems worldwide and the need for audits to mitigate the risk of bias. By comparing conviction probabilities across different population groups and within models, we uncover discrepancies that deviate from the known societal

trends. To our surprise, our results did not reflect the existing biases in the justice system, as one might expect given the models' reliance on historical data. Instead, they exposed a progressive bias correction. In doing so, our results suggest a world where ChatGPT is the architect of reality, not merely its mirror. While there is comfort in knowing that ChatGPT has the potential to correct harmful human behaviors – such as historic and institutionalized discrimination – one might also question whether such abilities are desirable, and under what circumstances they should be applied.

In conducting this experiment, we encountered two intertwined challenges. First, there is no "ground truth", meaning no definitive or "right" conviction probabilities to compare the GPT models' results against. Second, the closed nature of ChatGPT, a design choice, makes it difficult to analyze or draw conclusions about the factors that explain our results. The absence of a "right" conviction rate, combined with the lack of transparency regarding ChatGPT's training dataset, the weights assigned to different sources, and the soft or hard-coded policies the models are based upon, limits our ability to assess the processes underlying conviction predictions. These constraints highlight the main challenges in deploying AI in sensitive domains like the legal field. That being said, this work contributes to the growing body of research on algorithmic fairness and enhances our understanding of the potential and pitfalls of utilizing large language models (LLMs) in judicial contexts. Our findings underscore the need for in-depth discussions on progressive bias policies, as well as the development of robust evaluation metrics aimed at mitigating algorithmic biases and promoting ethical AI in high-stakes domains like the legal system.

Limitations

As with any study, we acknowledge several methodological limitations. First, we note an element of randomness in LLMs, meaning that even with the exact same prompt, different outputs could be generated in separate attempts. Similarly, due to the high sensitivity of LLMs to prompting, as reflected in our results, even slight variations in the prompt could result in different predictions. In our study, we reported the first answers produced and cannot assess whether subsequent responses differed. To mitigate this limitation, we focused on eliciting conviction probabilities rather than a single numeric value, using a large sample size. Nonetheless, given the high importance of supporting judicial decision-making, further research should aim to minimize randomness even more. One approach could be to explore deterministic models that prioritize consistency over efficiency in the cost-accuracy trade-off. Another relevant extension would be prompt engineering. Our study highlights the importance of experimenting with different prompting techniques to assess how prompt format affects discrepancies and performance. We further acknowledge that we did not directly verify whether the models recognized or interpreted identity cues. Future work should include explicit prompts, attention analyses, or probing techniques to better assess causal effects of these manipulations. Further, our study is primarily quantitative and centers on conviction predictions; therefore, our discussion of the underlying reasons for the observed bias is hypothesis-generating rather than conclusive. Future research should incorporate qualitative analyses of model reasoning to better uncover the mechanisms driving these trends – an especially important direction in the legal domain – where explainability is critical for legitimacy and trust. Future research could also expand the comparative analysis by incorporating additional manipulations, such as including victim characteristics, jurisdictional factors, and case scenarios (e.g., public-private, criminal-civil, or varying levels of severity). Additionally, future studies could explore the performance of different LLM models and, perhaps most importantly, work on developing evaluation criteria in the absence of a definitive "ground truth." Finally, our study relies on repeated draws from large language models, which do not equate to independent human participants. Nonetheless, because each output is stochastically sampled from the model's conditional probability distribution given the prompt, these draws can be meaningfully treated as independent realizations. A growing body of work employing "silicon samples" (i.e., samples drawn from language models) demonstrates that LLM predictions are, in certain instances, interchangeable with human behavior (Binz & Schulz, 2023; Dillion et al., 2023), including in legal settings (Choi et al., 2021; Katz et al., 2024). Such studies suggest that repeated model outputs approximate independent, human-like observations, sufficient for meaningful inference (Argyle et al., 2023; Aher et al., 2023). With that said, to further strengthen independence, we implemented several design choices. First, we systematically varied input conditions (e.g., vignette features) to introduce natural variation and mitigate correlations across repetitions. We also tested two model architectures (GPT-4 and GPT-4o) to assess whether observed patterns are robust across models with different weights. Second, within each model, the temperature was set to 1 to introduce moderate randomness. Temperature controls the randomness of model outputs: lower values make responses more deterministic, while higher values allow more variation. Setting temperature to 1 ensures that repeated runs of the same prompt are not deterministic, thereby increasing the effective independence of repetitions. Finally, prompts were processed in batches using single-turn sessions (i.e., without conversational history), minimizing contextual carryover. While our analysis

provides a robust view of language models' behavior in certain scenarios, generalization to other models or contexts should be made with caution.

DATA AVAILABILITY

The dataset supporting this paper and appendices are available in on the Open Science Framework (OSF) at: <https://osf.io/zc42x>

FUNDING STATEMENT

This project was funded by the European Union (ERC, 2021-COG, 101044195). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

ETHICAL CONSIDERATIONS

This article does not contain any studies with human participants performed by any of the authors.

INFORMED CONSENT

Informed consent was not required for this study as it did not involve human participants.

COMPETING INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank our research assistants (Samuel wu, Yoel Behar) for their work and are grateful to all of those who provided useful feedback, including Professor Orna Rabinovich-Einy, Avital Mentovich, Jeffrey Rachlinski and Professor Ke Xu of UIBE.

REFERENCES

- Abu-Elyounes, D. (2020). Contextual fairness: A legal and policy analysis of algorithmic fairness. *U. Ill. JL Tech. & Pol'y*, 1
- Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*. PMLR, 337-371
- Angwin, J. Larson, S. Mattu, L. Kirchner. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks, ProPublica
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351
- Armour, J., Parnham, R., & Sako, M. (2022). Augmented lawyering. *U. Ill. L. Rev.*, 71
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR. 77-91
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104, 671
- Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610-623
- Bent, J. R. (2019). Is algorithmic affirmative action legal. *Geo. LJ*, 108, 803

Bertrand, M., & 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

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6).

Bishop, C. M. (2006). Pattern recognition and machine learning. *Springer google schola*, 2, 1122-1128

Blair-Stanek, A., Holzenberger, N., & Van Durme, B. (2023, June). Can GPT-3 perform statutory reasoning?. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 22-31

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

Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. (2020). LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Chander, A. (2017). The racist algorithm. *Mich. L. Rev.*, 115, 1023

Chohlas-Wood, A., Coots, M., Goel, S., & Nyarko, J. (2023). Designing equitable algorithms. *Nature Computational Science*, 3(7), 601-610

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2021). ChatGPT goes to law school. *J. Legal Educ.*, 71, 387

Citron, D. K. (2007). Technological due process. *Wash. UL Rev.*, 85, 1249

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1), 14730-14846

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 64-93

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., ... & Kalai, A. T. (2019, January). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, 120-128

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7), 597–600

Farrington, D. P., & Morris, A. M. (1983). Sex, sentencing and reconviction. *The British Journal of Criminology*, 23(3), 229-248

Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., ... & Li, Z. (2023). Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36

Hacohen, U. Y., & Elkin-Koren, N. (2024). Copyright regenerated: Harnessing genai to measure originality and copyright scope. *Harvard Journal of Law & Technology*, 37(2)

Harlan E & Schnuck O (2021) Objective Or Biased, Bayerischer Rundfunk, <https://interaktiv.br.de/ki-bewerbung/en/> Accessed 11 December 2024

Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, 35, 11785-11799

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270), 20230254

Kline, P., Rose, E. K., & Walters, C. R. (2022). Systemic discrimination among large US employers. *The Quarterly Journal of Economics*, 137(4), 1963-2036

Knight W., The Apple Card Didn't 'See' Gender-and That's the Problem, *Wired* (Nov. 19, 2019), <https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/> Accessed 15 Oct 2024

Kruttschnitt, C. (1984). Sex and criminal court dispositions: The unresolved controversy. *Journal of Research in Crime and Delinquency*, 21(3), 213-232

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science*, 65(7), 2966-2981

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3), 217-224

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3), 305-325

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6), 1-35

Menkel-Meadow, Carrie. (1996) The Trouble with the Adversary System in a Post-Modern, Multi-Cultural World, *Journal of the Institute for the Study of Legal Ethics*: 1(6)

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. *AI and Ethics*, 1-31

National Center for State Courts. NCSC Court Statistics Project (2024). https://www.ncsctableauserver.org/t/Research/views/TrialDashboards/Criminal?%3Aembed=y&%3AisGuestRedirectFromVizportal=y&%3Aorigin=card_share_link (accessed Feb 23, 2026)

Nay, J. J., Karamardian, D., Lawskey, S. B., Tao, W., Bhat, M., Jain, R., ... & Kasai, J. (2024). Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, 382(2270), 20230159

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press

Perlman, A. (2022). *The Implications of ChatGPT for Legal Services and Society*. Suffolk University Law School Research Paper No. 22-14

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2009). Does unconscious racial bias affect trial judges. *Notre Dame L. Rev.*, 84, 1195

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429-435

Rice, D., Rhodes, J. H., & Nteta, T. (2019). Racial bias in legal language. *Research & Politics*, 6(2), 2053168019848930

Rozado, D. (2023). The political biases of chatgpt. *Social Sciences*, 12(3), 148

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl, M., & Pauly, M. (2024). The Self-Perception and Political Biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2024(1), 7115633

Spohn, C. (1990). The sentencing decisions of black and white judges: Expected and unexpected similarities. *Law and Society Review*, 1197-1216

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014), 4349-4357

Shaki, J., Kraus, S., & Wooldridge, M. (2023). Cognitive Effects in Large Language Models. In *ECAI 2023*. IOS Press. 2105-2112

Sommers, S. R., & Ellsworth, P. C. (2001). White juror bias: An investigation of prejudice against Black defendants in the American courtroom. *Psychology, Public Policy, and Law*, 7(1), 201

Susskind, R., & Susskind, R. E. (2023). *Tomorrow's lawyers: An introduction to your future*. Oxford University Press.

Trozze, A., Davies, T., & Kleinberg, B. (2024). Large language models in cryptocurrency securities cases: can a GPT model meaningfully assist lawyers?. *Artificial Intelligence and Law*, 1-47

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K. W., & Peng, N. (2023). "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., ... & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214-229

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1), 118-132

Supplementary references:

Innocence Project, Race and Wrongful Conviction, <https://innocenceproject.org/race-and-wrongful-conviction/> Accessed 15 Oct 2024

The Pew Charitable Trusts, Racial Disparities Persist in Many U.S. Jails (May 16, 2023), <https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2023/05/racial-disparities-persist-in-many-us-jails#:~:text=1,%2C%20age%2C%20or%20other%20demographics> Accessed 15 Oct 2024

United State Sentencing Commission, 2023 Demographic Differences in Federal Sentencing (Nov. 14, 2023), <https://www.ussc.gov/research/research-reports/2023-demographic-differences-federal-sentencing> Accessed 15 Oct 2024

ARTICLE IN PRESS