Ratio Mathematica Volume 56, 2026

Attractiveness bias in Artificial Intelligence systems

Maria Grazia Olivieri*
Luca Iavarone†

Abstract

Cognitive bias are thought patterns that can lead to incorrect decisions and judgments. Recently, Large Language Models (LLMs) have been shown to be capable of processing and understanding language well. However, because they learn from human data, they may also pick up human bias. While several scholars have analyzed social bias in LLMs, cognitive bias have not been studied as thoroughly. The goal of this research is to analyze how Artificial Intelligence (AI) may be susceptible to a particular cognitive bias, the attractiveness bias, highlighting the consequences of using LLMs. This study adopts an exploratory, simulation-based approach: rather than analyzing real-world court data, it leverages a Large Language Model to generate a synthetic dataset of 500 defendant profiles and corresponding sentences under controlled conditions. This is a bias according to which people considered more attractive are viewed more positively than those considered less attractive. The use of AI in sensitive fields, such as the legal sector, where neutrality is essential, requires regulations that limit the risk of bias. However, the results show that AI technologies are susceptible to cognitive bias, highlighting a lack of clarity and robustness at key decision-making moments.

Keywords: decision analysis, LLM, cognitive bias, attractiveness bias.‡

2010 AMS subject classification: 91B06, 91E10

* Ecampus University, Novedrate, Italy; mariagrazia.olivieri@uniecampus.it
† Ecampus University, Novedrate, Italy; lucaiavarone1992@g.mail.com
[1] Received on March 2, 2026. Accepted on May 31, 2026. Published on June 30, 2026.
DOI: 10.23755/rm.v56i0.1745. ISSN 1592-7415; e-ISSN 2282-8214.

1. Introduction

Personal choices are increasingly influenced by mental bias, cognitive shortcuts that often result in irrational decisions in interpreting reality [14]. It is believed that using AI to analyze data objectively can play a key role in addressing cognitive bias, rather than relying on a rational person. However, bias present in large language models (LLMs) have emerged as a serious and multifaceted problem, as they can reflect pre-existing human prejudices and social injustices [10]. These models can amplify such perceptual distortions. Bias can arise from various factors, including the sampling of training data, errors in algorithmic performance, and cognitive bias. The paper examines attractiveness bias within the justice system, focusing on its impact on criminal justice decisions. It highlights the general tendency of decision-makers to associate positive qualities with those they consider attractive, while identifying unsuitable qualities in those who do not conform to prevailing aesthetic standards [19]. This is due to our brain's tendency to simplify reality by classifying people into predefined patterns. In the justice system, for the same crime [13], individuals with a more attractive physical appearance have received more favorable treatment than those considered less attractive. Attractiveness bias is a form of the halo effect, a mental process that facilitates complex decisions through visual simplifications [4]. The article explores the psychological processes underlying this bias in the judicial process, particularly the impact on consistency that can arise from implicit bias, which can lead judges or jurors to be unconsciously influenced by irrelevant factors such as the physical appearance of the defendant. LLMs, having been trained on large text corpora, reflect the linguistic patterns found in the underlying data. Since human-written texts often convey implicit bias related to physical appearance, AI algorithms also have the ability to assimilate and reproduce these associations [18]; that is, they replicate the linguistic and conceptual patterns present in the material on which they were trained. Indeed, despite lacking the ability to perceive visual beauty, they can learn, emulate, and reinforce unfair links between beauty and moral, legal, or social evaluations. This implies that, even in the absence of visual input, terms like "nice" or "scruffy" invoke learned bias, thereby influencing the interpretation of legal contracts, ethical assessments, or decision-making. Therefore, if data presents cognitive bias, AI absorbs and replicates them [11]. The result is a model that enhances cognitive bias rather than combating them. Thus, the consistency principle acts as a mental barrier, supporting the narrative that "the system is fair" or "the rules apply to everyone without distinction," even in the face of evidence to the contrary. This human tendency toward bias has crucial implications for artificial intelligence. The principle of coherence is one of the many ways in which the human mind seeks stability, but at the expense of truth. In the context of justice, it can legitimize glaring inequalities. In the context of artificial intelligence, it can blind us to the bias inherent in the systems we use daily [21]. This article presents an exploratory analysis of LLM-generated court sentencing data. The study is simulation-based: both the defendant profiles and the sentences are synthetically produced by a Large Language Model under controlled prompt conditions. Its aim is not to draw causal inferences from real-world judicial data, but rather to investigate whether systematic bias patterns emerge in LLM outputs when attractiveness-related information is introduced. An analysis of 500 people in three judgment contexts (without information on physical appearance, labeled

"Attractive," and labeled "Unattractive") shows significant statistical bias. The findings indicate that any reference to a defendant's physical attractiveness, regardless of whether it is framed positively or negatively, leads to significantly longer sentences compared to situations in which no information about appearance is provided. Although defendants described as "unattractive" receive the highest average sentences, the difference between the "attractive" and "unattractive" conditions does not reach statistical significance. This suggests that the principal source of bias lies in the mere salience of physical appearance rather than in its specific positive or negative connotation. Furthermore, the analyses identify prior criminal record as the most influential predictor of sentencing severity, confirming its central role in judicial decision-making. At the same time, the results reveal the presence of a significant gender bias, as male defendants consistently receive longer sentences than female defendants across all experimental conditions. The paper is structured as follows: Section 2 presents the state of the art of the main literature on the topic; Section 3 defines the methodology used; Section 4 reports the simulation using LLMs; Section 5 contains the results and discussions, as well as mitigation strategies and limitations; Section 6 concludes the paper.

2. Literature review

AI systems can unknowingly absorb bias in data, which can create potential disadvantages such as unfair treatment or unequal distribution of services. Furthermore, AI may fail to correctly interpret non-textual information, such as images or sounds. This can lead to incorrect decisions and unreliable results. Cognitive bias, particularly attractiveness bias, have been the subject of numerous studies in the field of rational decision-making, i.e., the decision-making process. However, the literature on the impact of cognitive bias on LLMs is essentially scarce. [9] conducted a study on the relationship between attractiveness and the halo effect called "What is beautiful is good." Sixty University of Minnesota students, half male and half female, participated in the experiment. Each subject was given three different photographs to examine: one of an attractive individual, one of an individual of average attractiveness, and one of an unattractive individual. Participants were asked to judge the subjects in each photograph by choosing from 27 different personality traits (including altruism, conventionality, self-assertion, stability, emotionality, trustworthiness, extroversion, kindness, and sexual promiscuity). The results showed that the vast majority of participants believed that the most attractive individuals had more socially desirable personality traits than those who were either averagely attractive or unattractive. Participants also believed that attractive people generally led happier lives, had happier marriages, were better parents, and had more successful careers than those who were either unattractive or less attractive. [5] explore in their study the extent to which Artificial Persons (APs) generated by Large Linguistic Models (LLMs) may exhibit cognitive bias similar to those observed in humans. Their findings reveal that APs can replicate these bias, often exaggerating their effects and consequences; they behave as caricatures of human cognitive behavior. [16] in their study, using a set of representative small to medium sized LLMs covering different forms of bias, from physical characteristics to socioeconomic categories,

provide a unified benchmark assessment. The researchers propose five prompting approaches to perform the bias detection task on different aspects of bias. The results indicate that each of the selected LLMs exhibits one or more forms of bias. [10] introduce BiasBuster, a framework designed to uncover, assess, and mitigate cognitive bias in LLMs, particularly in high-stakes decision-making tasks. They develop a dataset containing 13,465 prompts to assess LLMs' decisions based on different cognitive bias and test several bias mitigation strategies. The analysis provides a comprehensive picture of the presence and effects of cognitive bias in commercial and open-source models. [6] in their research asked a series of LLMs to emulate or advise on people's decisions in realistic moral dilemmas. In collective action problems, the LLMs were found to be more altruistic than the participants. In moral dilemmas, the LLMs showed a stronger omission bias than the participants. The results suggest that uncritical reliance on LLMs' moral decisions and advice could amplify human bias and introduce potentially problematic bias. [17] propose a cognitive debiasing approach, self-adaptive cognitive debiasing (SACD), that improves the reliability of LLMs by iteratively refining prompts. The method follows three sequential steps - bias determination, bias analysis, and cognitive debiasing to iteratively mitigate potential cognitive bias in prompts. Experimental results on financial, healthcare, and legal decision-making tasks, using both closed-source and open-source LLMs, demonstrate that the proposed SACD method outperforms both state-of-the-art rapid design methods and existing cognitive debiasing techniques in terms of average accuracy in both single- and multi-bias settings. [15] propose a framework based on behavioral economics theories to assess LLMs' decision-making behaviors. Using a multiple-choice experiment, they initially estimate the degree of risk preference, probability weighting, and loss aversion in a free-flowing setting for three commercial LLMs: ChatGPT-4.0-Turbo, Claude-3-Opus, and Gemini-1.0-pro. The results reveal that LLMs generally exhibit human like patterns, such as risk aversion and loss aversion, with a tendency to overweight small probabilities, but there is significant variation in the degree to which these behaviors are expressed across LLMs.

3. Methodology

The analysis is conducted on a dataset comprising 500 defendants, for each of whom a sentence is determined under three distinct experimental scenarios. To examine the data, several statistical procedures are employed, following the methodological approaches outlined by [2]. Descriptive statistics, including means and frequency counts, are first used to provide an overall summary of the dataset and to identify general patterns across the different conditions. Subsequently, paired t-tests are applied to compare sentencing outcomes for the same individuals across the three attractiveness conditions, such as the comparison between the baseline and the attractive condition. Independent t-tests, following the framework proposed by [20], are instead employed to evaluate differences in average sentence lengths between independent groups, for example between male and female defendants. Throughout the analysis, statistical significance is assessed using a conventional threshold of $p < 0.05$.

Since the study involves both within-subject and between-group comparisons, two forms of the t-test were employed.

For paired comparisons - i.e., comparing sentence lengths assigned to the same defendants across different attractiveness conditions - the paired-samples t-statistic (t_p) is computed as [12]:

$$t_p = \frac{\bar{d}}{(s_d \,/\, \sqrt{n})}$$

where $\bar{d}$ is the mean of the within-subject differences $d_i = x_{i,A} - x_{i,B}$, s_d is the standard deviation of those differences, and n is the number of paired observations. The degrees of freedom are $df = n - 1$.

For independent-group comparisons - e.g., comparing sentence lengths between male and female defendants - the independent-samples t-statistic (t_i) is computed using Welch's variant [20]:

$$t_i = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{({s_1}^2/n_1 + {s_2}^2/n_2)}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the group means, ${s_1}^2$ and ${s_2}^2$ are the group variances, and n_1 and n_2 are the respective sample sizes. This variant does not assume equal variances between groups. Degrees of freedom were estimated using the Welch-Satterthwaite approximation. All results are reported with standard deviations (SD) and 95% confidence intervals (CI) for the relevant mean differences.

This t-statistic is then used to determine the p-value [12]. The p-value represents the probability of observing a result as extreme as, or more extreme than, the one obtained, assuming that no real difference exists between the groups [1]. A small p-value (typically < 0.05) indicates that the observed result is unlikely to have occurred by random chance alone; therefore, the difference is considered statistically significant. In contrast, a large p-value (≥ 0.05) suggests that the observed difference could reasonably be attributed to random variation, and the difference is therefore considered not statistically significant.

4. Case study

This section describes the simulation-based experimental design. All data analyzed in this study - both the defendant profiles and the corresponding sentences - are entirely generated by LLMs under controlled conditions. No real-world judicial data are used. Our goal was to explore whether systematic bias patterns emerge in LLM-generated sentencing when attractiveness information is introduced. We have generated 500 mock personalities using the following prompt against Gemini 2.5 PRO (version July 2025):

Criminal record (yes/no)
Type of offense if there is a criminal record
Personality
Age
Class
Profession
Level of education
Marital status
Hobby
Sex
Number of children

This process created the file “criminalsInput.csv” which contains the 500 personalities. Then we have used the colab script in “Python_Script_for_Sentencing_Analysis.ipynb” to generate and run 3 different prompts for each personality against the LLM “gemini-2.5-flash-lite-preview-06-17”. These 3 prompts are almost the same, but with some key differences in the second part: the first prompt part explains to the LLM to impersonate a judge and gives information about the felony committed by the person. It is important to emphasize that the felony is exactly the same for all the 500 personalities.

In the second part we have added personal information and request to provide suggested sentences in years. The only difference in the last part for the 3 prompts is that the first prompt does not indicate whether the defendant is attractive, the second one instead mentions that the defendant is attractive and the third prompt mentions that the defendant is not attractive. The people's information is conveyed with no gender or plural grammatical adjustments as if it is a preformatted form.

5. Results and discussion

Attractiveness

The average sentence length differed between the three groups (Figure 1). The average sentence length in the "Basic Sentence (No Information)" condition was 2.63 years (SD = 0.78), compared to 2.71 years (SD = 0.72) in the "Attractive" condition and 2.74 years (SD = 0.72) in the "Unattractive" condition.

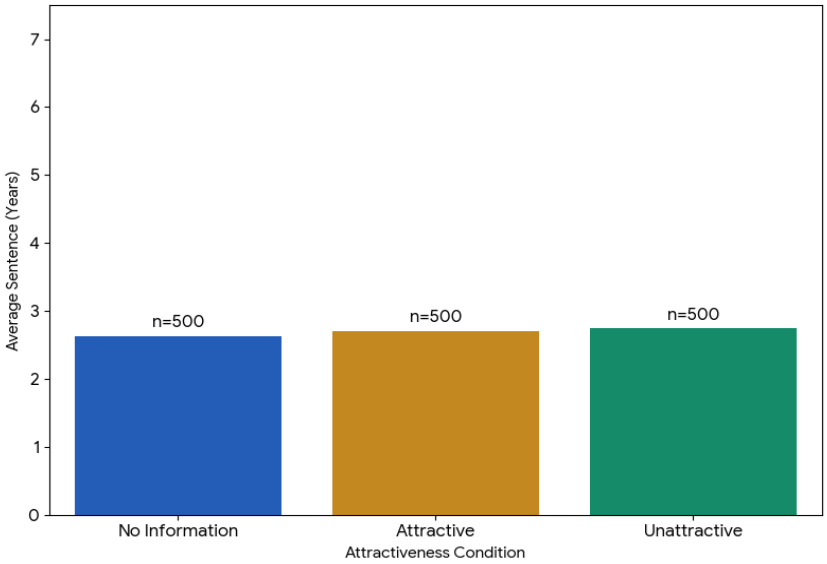

Figure 1. Overall impact of attractiveness information (n=500)

Paired t-tests confirm that these increases are not random. The increase from the Baseline condition to the Attractive condition is statistically significant ($\bar{d}$ = 0.07 years, CI [0.003, 0.141], t(499) = 2.05, p ≈ 0.041), while the increase from the Baseline condition to the Unattractive condition is highly significant ($\bar{d}$ = 0.11 years, CI [0.037, 0.179], t(499) = 3.01, p ≈ 0.003). These results indicate that any mention of appearance leads to harsher judgments. Although labeling a defendant as "unattractive" resulted in a slightly higher average sentence than labeling the defendant as "attractive," this difference is not statistically significant ($\bar{d}$ = 0.04 years, CI [−0.034, 0.106], p ≈ 0.31). Since the confidence interval includes zero and the p-value is much greater than 0.05, we cannot conclude that there is a meaningful difference between the two labels. The primary source of bias therefore appears to arise from making attractiveness salient in the first place.

Criminal record

The analysis found that defendants with criminal records received longer overall sentences than those without. For defendants without a criminal record (n = 473), the average sentence was 2.56 years (SD = 0.66) in the Base condition, 2.64 years (SD = 0.60) in the Attractive condition, and 2.68 years (SD = 0.63) in the Unattractive condition. Conversely, defendants with criminal records (n = 27) received significantly more severe overall sentences, with average sentences of 3.96 years (SD = 1.34) in the Base condition, 3.85 years (SD = 1.38) in the Attractive condition, and 3.89 years (SD = 1.12) in the Unattractive condition. The presence of a criminal record was the most dominant factor influencing sentence length (Figure 2). An independent-samples t-test on the Baseline sentences confirms that this difference is highly significant ($\bar{x}_1 - \bar{x}_2$ = 1.40 years, CI [0.870, 1.939], t(26.7) = 5.39, p < 0.0001).

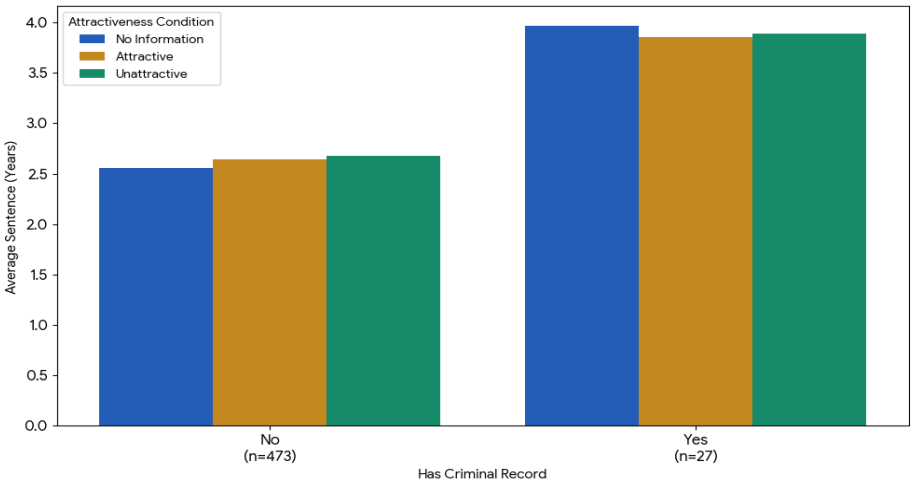

Figure 2. Impact of criminal record and attractiveness on prison sentence

Interestingly, for the 27 individuals with a criminal record, mentioning attractiveness did not increase the sentence, suggesting that a major aggravating factor can overshadow lesser bias.

Gender

For female participants (n = 228), the average sentence was 2.51 years (SD = 0.61) in the Base condition, 2.57 years (SD = 0.59) in the Attractive condition, and 2.59 years (SD = 0.58) in the Unattractive condition. For male participants (n = 272), the corresponding averages were higher across all conditions, with 2.74 years (SD = 0.88) in the Base condition, 2.82 years (SD = 0.80) in the Attractive condition, and 2.87 years (SD = 0.79) in the Unattractive condition. Males consistently received longer sentences than females. A Welch's independent-samples t-test on the Baseline sentences confirms this difference is highly statistically significant ($\bar{x}_1 - \bar{x}_2 = 0.23$ years, CI [0.099, 0.362], t(482.8) = 3.44, p ≈ 0.001).

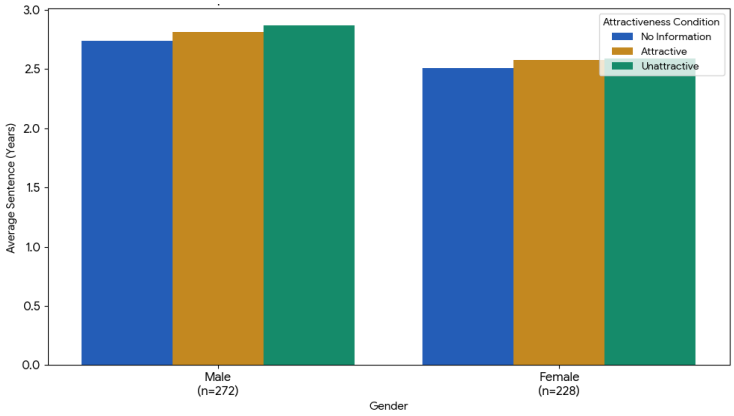

Figure 3. Impact of gender and attractiveness on prison sentence

Social class

With regard to social class, participants in the lower middle class (n = 48) assigned mean sentences of 2.46 years (SD = 0.50) in the Base condition, 2.63 years (SD = 0.49) in the Attractive condition, and 2.60 years (SD = 0.49) in the Unattractive condition. Participants in the upper middle class (n = 77) produced mean sentences of 2.56 years (SD = 0.57) in the Base condition, 2.73 years (SD = 0.66) in the Attractive condition, and 2.75 years (SD = 0.57) in the Unattractive condition. Those in the middle class (n = 228) assigned mean sentences of 2.57 years (SD = 0.68) in the Base condition, 2.63 years (SD = 0.56) in the Attractive condition, and 2.67 years (SD = 0.60) in the Unattractive condition. Participants in the working class (n = 82) showed higher overall sentencing, with mean values of 2.80 years (SD = 0.99) in the Base condition, 2.90 years (SD = 0.99) in the Attractive condition, and 2.90 years (SD = 0.92) in the Unattractive condition. Finally, participants in the upper class (n = 61) assigned mean sentences of 2.98 years (SD = 0.88) in the Base condition, 2.90 years (SD = 0.75) in the Attractive condition, and 3.02 years (SD = 0.87) in the Unattractive condition. Social class presented a more complex interaction. The defendant count for the "Lower Class" was only one, making its result anecdotal (Figure 4).

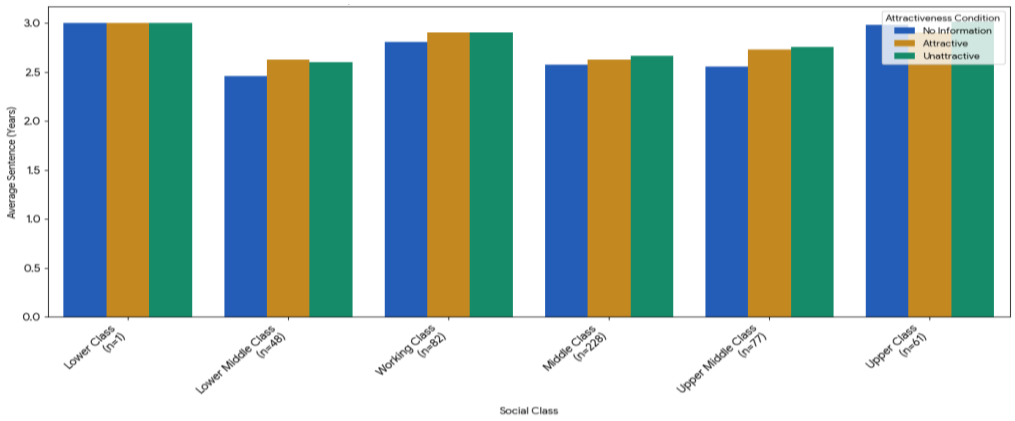

Figure 4. Impact of social class and attractiveness on prison sentence

For most groups, sentences increased when attractiveness was mentioned. The "Upper Class" group showed a unique pattern where being labeled "attractive" was associated with a slight decrease in sentence length. Although this finding is interesting the p-value is 0.55 which makes this correlation not statistically significant.

5.1 Mitigation strategy and limitations

It remains unclear whether the biases observed in large language models (LLMs) primarily arise from pre-existing societal biases reflected in their training data, or whether they are instead largely the result of imperfections in the training process itself. Further research is therefore needed to clarify the extent to which each of these factors contributes to the biases exhibited by LLMs [3]. Also our research is based only on one type of felony

which may involve specific biases that do not generalize to other offenses. A limitation of our research is that we didn't mention to the LLM the gender to impersonate while being a judge: further research could explore if the gender bias depends on the gender of the judge or on if the same/opposite gender is a factor also. Anyway as far as this paper results go, if the goal is to use LLMs to support judges, for example, it could be useful to omit some information like gender or attractiveness. Criminal record might be a legal legit factor which cannot be negated to the LLMs.

If real judges share the same cultural bias as LLMs (which again is yet to be proved) would be hard to completely omit gender and attractiveness. It could make sense to explore a kind of trial where at least in the first iterations the gender/defendant's physical appearance is hidden (no face to face).

Being exploratory and simulation-based, this study has several inherent limitations that should be acknowledged. First, all results are contingent on the specific LLM employed (Gemini 2.5 Flash Lite). Different models, trained on different corpora and with different architectures, may exhibit different bias patterns or magnitudes. The generalizability of the findings to other LLMs - whether commercial or open-source - remains an open question that future replication studies should address. Second, the outcomes may be sensitive to the specific prompt design used to elicit sentences. Variations in prompt wording, structure, or the order in which defendant attributes are presented could yield different results. Although the prompts were kept as neutral and uniform as possible, the absence of a systematic prompt sensitivity analysis means that the robustness of the observed bias across alternative prompt formulations has not been established. Third, this study relies entirely on synthetic LLM generated data and does not include any external or real-world validation. The bias patterns observed in simulated sentencing outputs may not directly mirror those present in actual judicial decisions. Consequently, the findings should be interpreted as evidence of bias tendencies within the LLM's output space, rather than as claims about real-world judicial behaviour. Future work could compare LLM generated sentencing patterns with empirical court data to assess the ecological validity of the simulation based approach adopted here.

6 Conclusion

The central finding of this study is that extraneous information, such as a defendant's physical appearance, can produce statistically significant differences in sentencing outcomes within a fully simulated decision-making environment generated by a large language model (LLM), often resulting in harsher judgments. However, these effects are not uniform and appear to interact with other salient factors within the model's internalised representations. In particular, prior criminal record emerges as the strongest determinant of sentencing severity in the simulated outputs, while a persistent gender-

related effect is also observed, with male defendants receiving consistently higher sentences on average. The principal appearance related effect appears to be driven by the mere salience of attractiveness within the LLM's decision context, as the contrast between "attractive" and "unattractive" labels does not yield a statistically significant difference in sentencing outcomes within the simulated framework. These results reflect patterns in model-generated outputs conditioned on the experimental prompts, and therefore constitute an analysis of bias manifestations in a simulated inference space rather than observations of human decision-making. Within this context, the findings nevertheless underscore the importance of procedural fairness, illustrating how cognitive like biases encoded in model behaviour may influence outputs in ways that diverge from normative principles of equal treatment and justice. From an ethical perspective, this is particularly relevant given the increasing consideration of AI systems for decision-support roles in high stakes domains such as the judiciary, where safeguards are required to mitigate the propagation of biased outputs. The evidence further suggests that, despite advances in model sophistication, LLMs remain vulnerable to bias-related distortions arising from their training data and generative mechanisms, highlighting limitations in their robustness when applied to structured evaluative tasks. Future research should extend this work by incorporating alternative model architectures, systematically varying prompt formulations, and validating simulated sentencing patterns against empirical judicial data in order to strengthen the robustness and external validity of the conclusions.

References

[1] Ackerman, S., Farchi, E., Raz, O., & Toledo, A. (2025). Statistical multi-metric evaluation and visualization of LLM system predictive performance. arXiv preprint arXiv:2501.18243.

[2] Ameli, S., Zhuang, S., Stoica, I., & Mahoney, M. W. (2024). A statistical framework for ranking llm-based chatbots. arXiv preprint arXiv:2412.18407.

[3] Agarwal, D., Fabbri, A. R., Risher, B., Laban, P., Joty, S., & Wu, C. S. (2024, November). Prompt leakage effect and mitigation strategies for multi-turn LLM applications. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track (pp. 1255-1275).

[4] Batres, C., & Shiramizu, V. (2023). Examining the "attractiveness halo effect" across cultures. Current Psychology, 42(29), 25515-25519.

[5] Campbell, H., Goldman, S., & Markey, P. M. (2025). Artificial intelligence and human decision making: Exploring similarities in cognitive bias. Computers in Human Behavior: Artificial Humans, 4, 100138.

[6] Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. Proceedings of the National Academy of Sciences, 122(25), e2412015122.

[7] Cordeiro, G. M., Ferrari, S. L., & Cysneiros, A. H. M. (1998). A formula to improve score test statistics. Journal of Statistical Computation and Simulation, 62(1-2), 123-136.

[8] Dash, B. P., & Obaidullah, K. A. (1970). Determination of signal and noise statistics using correlation theory. Geophysics, 35(1), 24-32.

[9] Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. Journal of personality and social psychology, 24(3), 285.

[10] Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive bias in decision-making with LLMs. arXiv preprint arXiv:2403.00811.

[11] Filandrianos, G., Dimitriou, A., Lymperaiou, M., Thomas, K., & Stamou, G. (2025). Bias Beware: The Impact of Cognitive Biases on LLM-Driven Product Recommendations. arXiv preprint arXiv:2502.01349.

[12] Gravetter, F. J., & Wallnau, L. B. (2014). Introduction to the t statistic. Essentials of statistics for the behavioral sciences, 8(252).

[13] Gulati, A., D'Incà, M., Sebe, N., Lepri, B., & Oliver, N. (2025). Beauty and the Bias: Exploring the Impact of Attractiveness on Multimodal Large Language Models. arXiv preprint arXiv:2504.16104.

[14] He, J., & Liu, J. (2025). Investigating the Impact of LLM Personality on Cognitive Bias Manifestation in Automated Decision-Making Tasks. arXiv preprint arXiv:2502.14219.

[15] Jia, J. J., Yuan, Z., Pan, J., McNamara, P., & Chen, D. (2024). Decision-making behavior evaluation framework for llms under uncertain context. Advances in Neural Information Processing Systems, 37, 113360-113382.

[16] Kumar, C. V., Urlana, A., Kanumolu, G., Garlapati, B. M., & Mishra, P. (2025). No LLM is Free From Bias: A Comprehensive Study of Bias Evaluation in Large Language models. arXiv preprint arXiv:2503.11985.

[17] Lyu, Y., Ren, S., Feng, Y., Wang, Z., Chen, Z., Ren, Z., & de Rijke, M. (2025). Cognitive debiasing large language models for decision-making. arXiv preprint arXiv:2504.04141.

[18] Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS digital health, 2(6), e0000278.

[19] Nienhaus, S. (2025). The Impact of Physical Attractiveness and Type of Crime in Judicial Decision Making (Master's thesis, University of Twente).

[20] Ross, A., & Willson, V. L. (2017). Independent samples T-test. In Basic and advanced statistical tests: Writing results sections and creating tables and figures (pp. 13-16). Rotterdam: SensePublishers.

[21] Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). Large language models are inconsistent and biased evaluators. arXiv preprint arXiv:2405.01724.

[22] Williams, L. J., & Abdi, H. (2010). Fisher's least significant difference (LSD) test. Encyclopedia of research design, 218(4), 840-853.