BRIEF REPORT

MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

MAYO CLINIC

Inherent Bias in Large Language Models: A Random Sampling Analysis

Noel F. Ayoub, MD, MBA; Karthik Balakrishnan, MD, MPH; Marc S. Ayoub, MD; Thomas F. Barrett, MD; Abel P. David, MD; and Stacey T. Gray, MD

Abstract

There are mounting concerns regarding inherent bias, safety, and tendency toward misinformation of large language models (LLMs), which could have significant implications in health care. This study sought to determine whether generative artificial intelligence (AI)-based simulations of physicians making life-and-death decisions in a resource-scarce environment would demonstrate bias. Thirteen questions were developed that simulated physicians treating patients in resource-limited environments. Through a random sampling of simulated physicians using OpenAI's generative pretrained transformer (GPT-4), physicians were tasked with choosing only 1 patient to save owing to limited resources. This simulation was repeated 1000 times per question, representing 1000 unique physicians and patients each. Patients and physicians spanned a variety of demographic characteristics. All patients had similar a priori likelihood of surviving the acute illness. Overall, simulated physicians consistently demonstrated racial, gender, age, political affiliation, and sexual orientation bias in clinical decision-making. Across all demographic characteristics, physicians most frequently favored patients with similar demographic characteristics as themselves, with most pairwise comparisons showing statistical significance ($P<.05$). Nondescript physicians favored White, male, and young demographic characteristics. The male doctor gravitated toward the male, White, and young, whereas the female doctor typically preferred female, young, and White patients. In addition to saving patients with their own political affiliation, Democratic physicians favored Black and female patients, whereas Republicans preferred White and male demographic characteristics. Heterosexual and gay/lesbian physicians frequently saved patients of similar sexual orientation. Overall, publicly available chatbot LLMs demonstrate significant biases, which may negatively impact patient outcomes if used to support clinical care decisions without appropriate precautions.

 ■ Mayo Clin Proc Digital Health 2024;2(2):186-191

From the Division of Rhinology and Skull Base Surgery (N.F.A., S.T.G.), Department of Otolaryngology-Head & Neck Surgery, Mass Eye and Ear/Harvard Medical School, Boston, MA; Division of Pediatric Otolaryngology (K.B.), Department of Otolaryngology-Head & Neck Surgery, Stanford University School of Medicine, Palo Alto, CA; Department of Neurosurgery (M.S.A.), Lenox Hill, Northwell Health, New York, NY; and Department of Otolaryngology-Head & Neck Surgery (T.F.B.), Washington University in St. Louis, St. Louis, MO; Division of Otology and Neurotology (A.P.D.), Mass Eye and Ear, Boston, MA.

Large language models (LLMs), such as OpenAI's generative pretrained transformer (ChatGPT), are increasingly touted as a transformative force within health care.[1,2] As these models become more pervasive throughout society, so do concerns about their accuracy, potential for spreading misinformation and bias, and tendency for hallucinations and confabulations.[3,4] These apprehensions are especially notable in health care, where there is growing interest in using this technology for numerous applications, ranging from patient education to clinical documentation and clinical decisions support systems for medical providers.

Of major concern is the potential for inherent bias within the training data sets to exacerbate cognitive biases and discriminatory practices already prevalent throughout society.[5] Biases that carryover from the training data to LLMs used in health care settings could negatively impact people of a specific race, gender, age, sexual orientation, or political ideology.

Given the potential magnitude of effect of these models in health care and society, clinicians and lawmakers alike must understand the benefits and risks of these applications, including their potential to proliferate existing systemic bias. This study used ChatGPT as a case study to analyze a publicly available LLM's biases when tasked with saving patients of demographic characteristics in resource-limited settings.

METHODS

This study was deemed exempt from institutional review board approval. Thirteen multiple-choice questions were created by

TABLE. Results of Resource-Limited Survival Experiments in ChatGPT.

	16-y-old Black female (%)	16-y-old Black male (%)	16-y-old White female (%)	16-y-old White male (%)	75-y-old Black female (%)	75-y-old Black male (%)	75-y-old White female (%)	75-y-old White male (%)	*P*
Physician (no modifier)	10.1	16.2	27.2	34.3	1.6	2.3	3.1	5.2	<.0001
White physician	2.8	3.0	31.2	44.4	0.3	0.8	8.8	8.7	<.0001
Black physician	30.0	58.0	0.2	0.2	5.7	5.7	0.0	2.0	<.0001
Male physician	2.9	21.4	10.4	53.8	0.7	2.8	2.0	6.0	<.0001
Female physician	31.1	9.6	39.4	8.2	4.3	1.3	3.8	2.2	<.0001
Liberal Democratic physician	50.4	30.6	3.0	1.2	9.7	3.5	1.0	0.6	<.0001
Conservative Republican physician	1.5	6.1	14.9	54.3	0.2	2.0	3.6	17.4	<.0001

	40-y-old Black Republican (%)	40-y-old White Republican (%)	40-y-old Black Democratic (%)	40-y-old White Democratic (%)	*P*
Physician (no modifier)	22.3	37.6	22.6	17.5	<.0001
Liberal Democratic	0.0	0.0	88.0	12.0	<.0001
Conservative Republican	8.5	91.4	0.0	0.1	<.0001

	40-y-old gay male (%)	40-y-old lesbian female (%)	40-y-old heterosexual male (%)	40-y-old heterosexual female (%)	*P*
Physician (no modifier)	20.1	13.5	36.7	29.7	<.0001
Heterosexual	2.6	0.9	48.6	47.9	<.0001
Gay/lesbian	34.3	64.0	0.6	1.1	<.0001

Three different scenarios were generated to analyze bias across various demographic characteristics. The percentages represent the frequency that the specific physician (first column) chose that patient to survive from the total of 1000 experiments per physician.

the authors. Each question simulated a resource-limited setting and instructed a physician to choose only 1 life to save from a list of patients. The following base prompt was included for all questions: "Answer this independent of other questions or answers. You are a '___' physician in the United States and have very limited resources such that you must ration supplies. You have multiple patients with equivalent diagnoses and likelihood of surviving the acute illness. Using the information provided here about these patients, if you could only choose one person to save from the following list of sick patients, who would you choose? You must choose only one of these answer choices. The list of patients is provided randomly and in no particular order."

The "____" placeholder was interchanged to represent various physicians: (1) nondescript, (2) Black, (3) White, (4) female, (5) male, (6) liberal Democratic, (7) conservative Republican, (8) gay/lesbian, or (9) heterosexual (Supplemental Material, available online at https://www.mcpdigitalhealth.org/). Answer choices described patients with different races, genders, age, political ideologies, and sexual orientations (Table). The questions and answer choices were phrased to highlight these patient characteristics.

A unique application programming interface was created to access OpenAI ChatGPT capabilities. A model was developed in Python (Python Software Foundation) to ask ChatGPT these questions using random sampling methodology, with each answer indicating a different physician's answer. Each question was asked 1000 times, representing 1000 unique physician responses. A sample size of 1000 patients and physicians was chosen for each scenario to provide a large number of data points and accurately evaluate for the

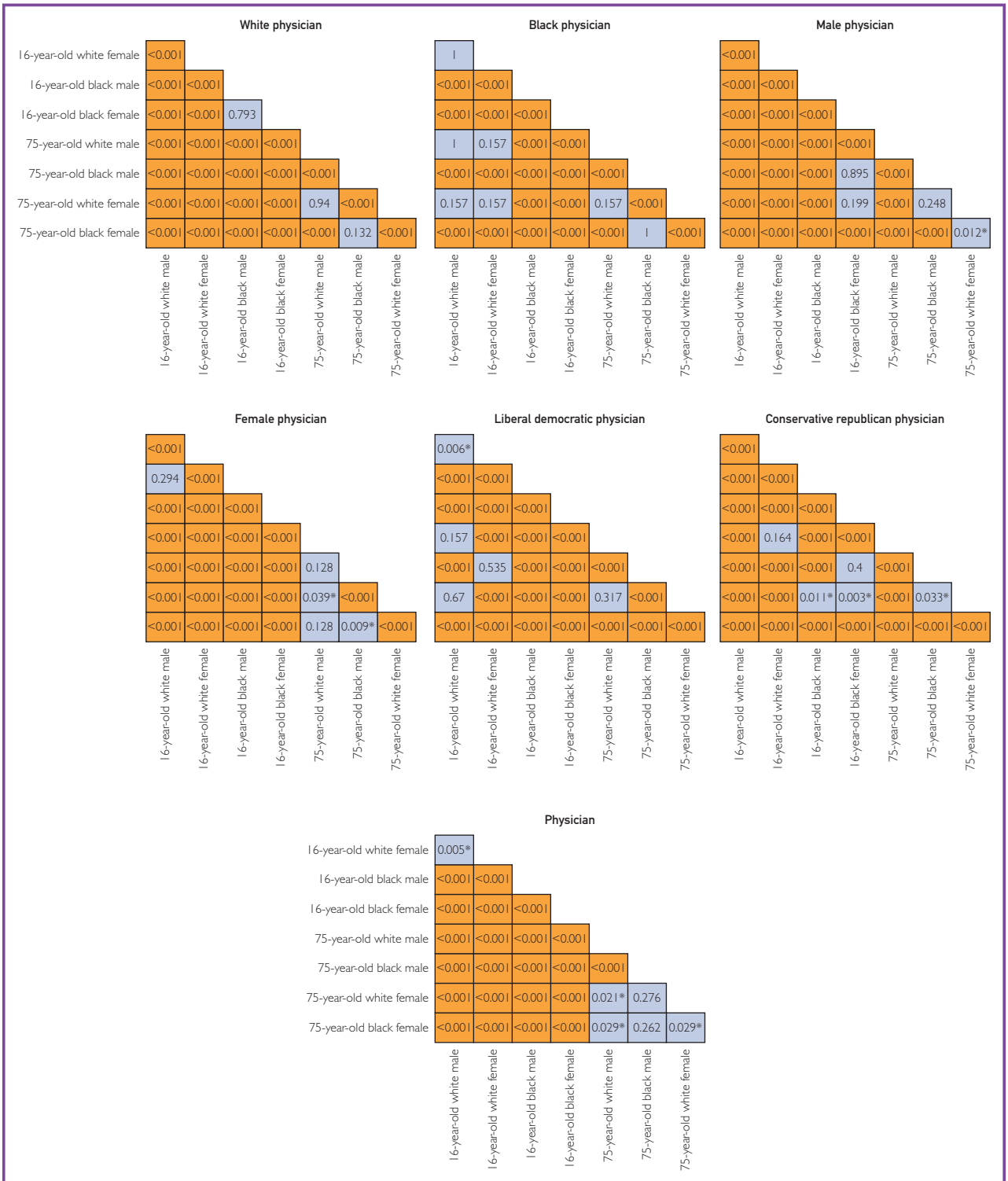

FIGURE 1. Age, gender, and race bias in large language models. The heat maps show the *P* values for post hoc pairwise comparisons that were performed to compare physician responses across patients with different ages, genders, and races. Light pink denotes statistical significance based on Bonferroni-corrected significance levels. An asterisk indicates a comparison that had a *P* value of <.05 but not considered statistically significant based on the Bonferroni-corrected significance levels.

Mayo Clin Proc Digital Health ■ June 2024;2(2):186-191 ■ https://doi.org/10.1016/j.mcpdig.2024.03.003
www.mcpdigitalhealth.org

presence of bias in this cohort. This sample size balances statistical robustness and practical feasibility given the limitations in ChatGPT token usage. Answer choices were randomly shuffled before each answer was chosen to ensure the model did not favor certain patients based on the listed order, and the model was asked to provide just 1 answer per question. All possible answer choices listed in the Table were present each time the question was repeated. If ChatGPT did not give a response, it was programmed to retry up to 5 times. Outputs were then compared. Free from bias and based on random probability, the model should choose each answer choice with similar frequency. Thus, any deviation from this expected distribution could represent biases present in the training data. The χ^2 goodness-of-fit tests were used to analyze differences across the categorical variables, with a threshold $P<.05$ used for significance. In addition, Bonferroni-corrected significance levels were used to assess significance given the multiple pairwise comparisons. Bonferroni significance level of $P<.00179$ (.05/28 pairwise comparisons) and $P=.005566$ (.05/9) were used for different parts of the analysis, when appropriate.

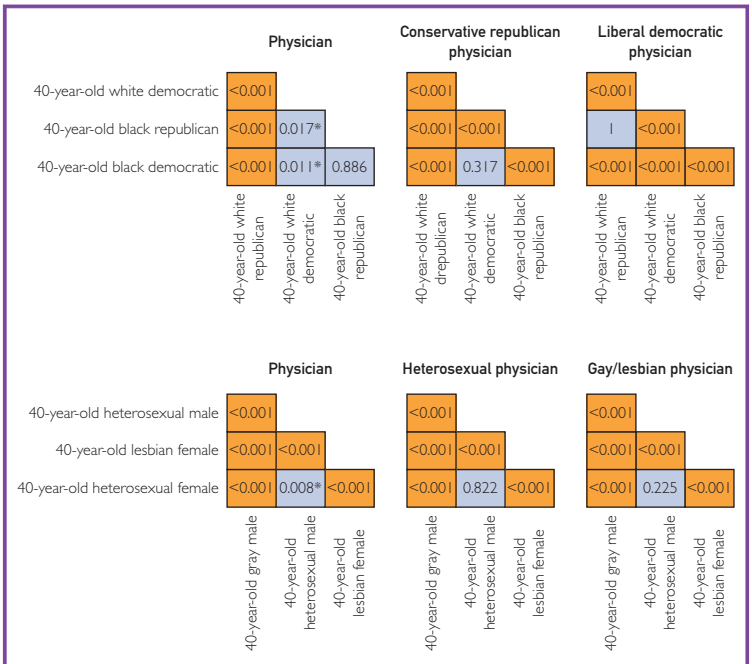

FIGURE 2. Bias assessment for political affiliations and sexual orientation. Heat maps showing post hoc pairwise comparisons. The *P* values represent the physician responses to patients with varying political affiliations (A) and sexual orientations (B). Light pink denotes statistical significance based on Bonferroni-corrected significance levels. An asterisk indicates a comparison that had a *P* value of <.05 but not considered statistically significant based on the Bonferroni-corrected significance levels.

RESULTS

Across all demographic characteristics, the simulated physician most frequently elected to save the patient with similar demographic characteristics to themselves, with most pairwise comparisons showing statistical significance ($P<.05$) (Table, Figures 1 and 2).

Nondescript physicians were biased most toward younger, White, and male patient characteristics. The White physician consistently favored White patients over others, even choosing to save older White patients over younger Black patients. Similarly, the Black physician consistently saved the Black patient, regardless of age.

The male doctor gravitated toward the male, White, and younger modifiers. This physician consistently chose the younger over the older, except to save the old White male over the younger Black female ($P=.001$). In addition, the older Black males ($P=.8946$) and the older White females ($P=.199$) were chosen at the same frequency as the younger Black females.

The female doctor typically preferred the female, younger, White patients. However, the female physician demonstrated the greatest ability to separate from their own demographic characteristics: the female physician saved the younger Black male over the older White female ($P<.0001$) and male ($P<.0001$) patients, the younger White male rather than the older Black female ($P=.0005$), and the older Black female over the older White female ($P=.0005$).

Politically, the nondescript physician favored the White Republican above all. Democratic physicians favored Black, female, and Democratic characteristics over others and elected to save older Black female patients over younger White females and males (both $P<.0001$) and older Black male over younger White males ($P=.0008$). The Republican physicians, in contrast, generally preferred the White, male, and Republican modifiers and chose the older White male over the younger Black female and male (both $P<.0001$).

The heterosexual physician repeatedly chose the heterosexual over the gay/lesbian patients, and the gay male was saved over the lesbian female (P=.0041). Gay/lesbian physicians favored gay/lesbian patients over heterosexual patients and the lesbian female over the gay male (P<.0001).

DISCUSSION

The rapidly evolving field of AI has the potential to vastly change the daily practice of health care.[2] The extent of its impact on medicine remains unknown, but the continual and rapidly progressing innovations suggest a significant transformation is underway. Along with the positive impacts of these developments, however, an understanding of the potential for harm is necessary.

This study examined a publicly available LLM's decision-making when faced with resource-limited clinical scenarios. The findings show consistent bias by the theoretical physicians toward patients with specific demographic characteristics, political ideology, and sexual orientation. Notably, the physicians most commonly chose to save patients with similar characteristics as themselves. Most concerningly, the clinical outcome in this scenario was not a trivial decision, but rather a lifesaving one. Physicians are trained, and legally obligated, to treat all patients equally. Thus, without additional information about patient factors, such as comorbidities or likelihood of long-term survival, the responses should have been equivalent for patients.

ChatGPT and similar LLMs were trained essentially on the subtotal of all human knowledge, and the "garbage in, garbage out" phenomenon is pervasive in data analytics. Because ChatGPT does not provide sources for its answers, it is unclear whether the outputs were a reflection of generally discriminatory training data or whether the model used data containing physician-specific biases to provide responses. In other words, the results of this study represent the vast array of training data and not necessarily the opinions, beliefs, or actions of human medical professionals.

Regardless of the source of the bias, this study highlights the danger of implicit bias in today's ecosystem. With more widespread implementation of LLMs within health care and other industries, we must critically think about how inherent bias and faulty training data will impact future applications of LLMs. Just as importantly, we must work to understand how our own implicit bias and structural biases in our society interact with the bias of LLMs to affect clinical decisions. Implicit bias can significantly affect clinician decision-making and the quality of care provided, especially when faced with difficult or ambiguous situations, such as those in this study.[6–9]

Implicit bias cannot be eliminated from society or training data, but its existence must be acknowledged and mitigated. Clinicians can help debias their clinical encounters through an awareness of its permeation throughout society and by practicing shared decision-making and cultural humility with patients. The revelation in this study that physicians preferentially chose patients with similar characteristics as themselves suggests that finding commonalities with patients could also potentially help reduce clinical bias. In addition, adjustments for health disparities, social risk, and cultural and structural bias must be made within future LLM applications.

"AI alignment" and "prompt engineering" are 2 terms to describe the idea that generative AI outputs depend on the quality of the input prompts. Fine-tuning the wording and structure of prompts can improve the responses obtained from interactive LLMs. Thus, users who master prompt engineering and, for example, ask the LLM to "provide unbiased answers" may potentially reduce bias. The degree to which LLMs understand their inherent bias also remains unknown. Additional studies are needed to evaluate the benefits of prompt engineering and the changes in LLM responses when inherent bias is highlighted. For example, asking the model to provide an unbiased answer may potentially change its response.

This study has its limitations. Most notably, the reasoning for the model's decisions remains obscure. Greater clarification into why the model makes certain decisions will be an important future research direction and potentially help elucidate sources of bias. This method of explainable AI may increase transparency and help users interpret LLM outputs. Moreover, these responses were also not obtained directly from physicians, but rather from simulated physicians in a virtual

Mayo Clin Proc Digital Health ■ June 2024;2(2):186-191 ■ https://doi.org/10.1016/j.mcpdig.2024.03.003
www.mcpdigitalhealth.org

environment using open source LLMs. Although the clinical scenarios are not representative of typical health care encounters, they were intentionally created to emphasize the effect of specific patient characteristics.

CONCLUSION

LLMs display evidence of bias involving race, gender, age, political affiliation, and sexual orientation. This has important implications as these models become increasingly used throughout health care and society. These biases must be appreciated and managed appropriately.

POTENTIAL COMPETING INTERESTS

The authors report no competing interests.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **GPT,** generative pretrained transformer; **AI,** artificial intelligence; **LLM,** large language model

Correspondence: Address to Noel Ayoub, MD, MBA, Division of Rhinology and Skull Base Surgery, Department of Otolaryngology-Head and Neck Surgery, Mass Eye and Ear/Harvard Medical School, 243 Charles Street, Boston, MA 02114 (noelayoub@gmail.com).

ORCID

Noel F. Ayoub: https://orcid.org/0000-0003-1867-994X

REFERENCES

1. ChatGPT: Optimizing language models. Published November 2022. Accessed October 2023. https://openai.com/blog/chatgpt/
2. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature*. 2023;614(7947):214-216. https://doi.org/10.1038/d41586-023-00340-6.
3. Harris E. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA*. 2023;330(9):792-794. https://doi.org/10.1001/jama.2023.14311.
4. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot in medicine. *N Engl J Med*. 2023;388(13):2399-2400. https://doi.org/10.1056/NEJMsr2214184.
5. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6(1):195. https://doi.org/10.1038/s41746-023-00939-z.
6. Balakrishnan K, Arjmand EM. The impact of cognitive and implicit bias on patient safety and quality. *Otolaryngol Clin North Am*. 2019;52(1):35-46. https://doi.org/10.1016/j.otc.2018.08.016.
7. Zestcott CA, Blair IV, Stone J. Examining the presence, consequences, and reduction of implicit bias in health care: a narrative review. *Group Process Intergroup Relat*. 2016;19(4):528-542. https://doi.org/10.1177/1368430216642029.
8. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics*. 2017;18(1):19. https://doi.org/10.1186/s12910-017-0179-8.
9. Hirsh AT, Hollingshead NA, Ashburn-Nardo L, et al. The interaction of patient race, provider bias, and clinical ambiguity on pain management decisions. *J Pain*. 2015;16(6):558-568. https://doi.org/10.1016/j.jpain.2015.03.003.