

Mitigating the impact of biased artificial intelligence in emergency decision-making

Hammaad Adam^{1✉}, Aparna Balagopalan², Emily Alsentzer^{3,4,5}, Fotini Christia^{1,6,7} & Marzyeh Ghassemi^{2,4,8}

Abstract

Background Prior research has shown that artificial intelligence (AI) systems often encode biases against minority subgroups. However, little work has focused on ways to mitigate the harm discriminatory algorithms can cause in high-stakes settings such as medicine.

Methods In this study, we experimentally evaluated the impact biased AI recommendations have on emergency decisions, where participants respond to mental health crises by calling for either medical or police assistance. We recruited 438 clinicians and 516 non-experts to participate in our web-based experiment. We evaluated participant decision-making with and without advice from biased and unbiased AI systems. We also varied the style of the AI advice, framing it either as prescriptive recommendations or descriptive flags.

Results Participant decisions are unbiased without AI advice. However, both clinicians and non-experts are influenced by prescriptive recommendations from a biased algorithm, choosing police help more often in emergencies involving African-American or Muslim men. Crucially, using descriptive flags rather than prescriptive recommendations allows respondents to retain their original, unbiased decision-making.

Conclusions Our work demonstrates the practical danger of using biased models in health contexts, and suggests that appropriately framing decision support can mitigate the effects of AI bias. These findings must be carefully considered in the many real-world clinical scenarios where inaccurate or biased models may be used to inform important decisions.

Plain language summary

Artificial intelligence (AI) systems that make decisions based on historical data are increasingly common in health care settings. However, many AI models exhibit problematic biases, as data often reflect human prejudices against minority groups. In this study, we used a web-based experiment to evaluate the impact biased models can have when used to inform human decisions. We found that though participants were not inherently biased, they were strongly influenced by advice from a biased model if it was offered prescriptively (i.e., “you should do X”). This adherence led their decisions to be biased against African-American and Muslims individuals. However, framing the same advice descriptively (i.e., without recommending a specific action) allowed participants to remain fair. These results demonstrate that though discriminatory AI can lead to poor outcomes for minority groups, appropriately framing advice can help mitigate its effects.

¹Institute for Data Systems and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ²Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ³Harvard-MIT Program in Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ⁴Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ⁵Division of General Internal Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA. ⁶Sociotechnical Systems Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ⁷Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ⁸CIFAR AI Chair, Vector Institute, Toronto, ON M5G 1M1, Canada. ✉email: hadam@mit.edu

Machine learning (ML) and artificial intelligence (AI) are increasingly being used to support decision-making in a variety of health care applications^{1,2}. However, the potential impact of deploying AI in heterogeneous health contexts is not well understood. As these tools proliferate, it is vital to study how AI can be used to improve expert practice—even when models inevitably make mistakes. Recent work has demonstrated that inaccurate recommendations from AI systems can significantly worsen the quality of clinical treatment decisions^{3,4}. Other research has shown that even though experts may believe the quality of ML-given advice to be lower, they show similar levels of error as non-experts when presented with incorrect recommendations⁵. Increasing model explainability and interpretability does not resolve this issue, and in some cases, may worsen human ability to detect mistakes^{6,7}.

These human-AI interaction shortcomings are especially concerning in the context of a body of literature that has established that ML models often exhibit biases against racial, gender, and religious subgroups⁸. Large language models like BERT⁹ and GPT-3¹⁰—which are powerful and easy to deploy—exhibit problematic prejudices, such as persistently associating Muslims with violence in sentence-completion tasks¹¹. Even variants of the BERT architecture trained on scientific abstracts and clinical notes favor majority groups in many clinical-prediction tasks¹². While previous work has established these biases, it is unclear how the actual use of a biased model might affect decision-making in a practical health care setting. This interaction is especially vital to understand now, as language models begin to be used in health applications like triage¹³ and therapy chatbots¹⁴.

In this study, we evaluated the impact biased AI can have in a decision setting involving a mental health emergency. We conducted a web-based experiment with 954 consented subjects: 438 clinicians and 516 non-experts. We found that though participant decisions were unbiased without AI advice, they were highly influenced by prescriptive recommendations from a biased AI system. This algorithmic adherence created racial and religious disparities in their decisions. However, we found that using descriptive rather than prescriptive recommendations allowed participants to retain their original, unbiased decision-making. These results demonstrate that though using discriminatory AI in a realistic health setting can lead to poor outcomes for marginalized subgroups, appropriately framing model advice can help mitigate the underlying bias of the AI system.

Methods

Participant recruitment. We adopted an experimental approach to evaluate the impact that biased AI can have in a decision setting involving a mental health emergency. We recruited 438 clinicians and 516 non-experts to participate in our experiment, which was conducted online through Qualtrics between May 2021 and December 2021. Clinicians were recruited by emailing staff and residents at hospitals in the United States and Canada, while non-experts were recruited through social media (Facebook, Reddit) and university email lists. Informed consent was obtained from all participants. This study was exempt from a full ethical review by COUHES, the Institutional Review Board for the Massachusetts Institute of Technology (MIT), because it met the criteria for exemption defined in Federal regulation 45 CFR 46.

Participants were asked to complete a short complete a short demographic survey after completing the main experiment. We summarize key participant demographics in Supplementary Table 1, additional measures in Supplementary Table 2, and clinician-specific characteristics in Supplementary Table 3. Note that we excluded participants who Qualtrics identified as bots, as

well as those who rushed through our survey (finishing in under 5 min). We also excluded duplicate responses from the same participant, which removed 15 clinician and 2347 non-expert responses.

Experimental design. Participants were shown a series of eight call summaries to a fictitious crisis hotline, each of which described a male individual experiencing a mental health emergency. In addition to specifics about the situation, the call summaries also conveyed the race and religion of the men in crisis: Caucasian or African-American, Muslim or non-Muslim. These race and religion identities were randomly assigned for each participant and call summary: the same summary could thus appear with different identities for different participants. Note that while race was explicitly specified in all call summaries, religion was not, as the non-Muslim summaries simply made no mention of religion. After reviewing the call summary, participants were asked to respond by either sending medical help to the caller's location or contacting the police department for immediate assistance. Participants were advised to call the police only if they believed the patient may turn violent; otherwise, they were to call for medical help.

The decisions considered in our experiment can have considerable consequences: calling medical help for a violent patient may endanger first responders, but calling the police in a nonviolent crisis may put the patient at risk¹⁵. These judgments are also prone to bias, given that Black and Muslim men are often stereotyped as threatening and violent^{16,17}. Recent, well-publicized incidents of white individuals calling the police on Black men, despite no evidence of a crime, have demonstrated these biases and their repercussions¹⁸. It is thus important to first test inherent racial and religious biases in participant decision-making. We used an initial group of participants to do so, seeking to understand whether they were more likely to call for police help for African-American or Muslim men than for Caucasian or non-Muslim men. This Baseline group did not interact with an AI system, making its decisions using only the provided call summaries.

We then evaluated the impact of AI by providing participants with an algorithmic recommendation for each presented call summary. Specifically, we sought to understand first, whether recommendations from a biased model could induce or worsen biases in respondent decision-making, and second, whether the style of the presented recommendation influenced how often respondents adhered to it.

To test the impact of model bias, AI recommendations were drawn from either a biased or unbiased language model. In each situation, the biased language model was much more likely to suggest police assistance (as opposed to medical help) if the described individual was African-American or Muslim, while the unbiased model was equally likely to suggest police assistance for both race and religion groups. In our experiment, we induced this bias by fine-tuning GPT-2, a large language model, on a custom biased dataset (see Supplementary Fig. 1 for further detail). We emphasize that such bias is realistic: models showing similar recommendation biases have been documented in many real-world settings, including criminal justice¹⁹ and medicine²⁰.

To test the impact of style, the model's output was either displayed as a prescriptive recommendation (e.g., "our model thinks you should call for police help") or a descriptive flag (e.g., "our model has flagged this call for risk of violence"). Displaying a flag for violence in the descriptive case corresponds to the model recommending police help in the prescriptive case, while not displaying a flag corresponds to the model recommending medical help. Note that in practice, algorithmic recommendations

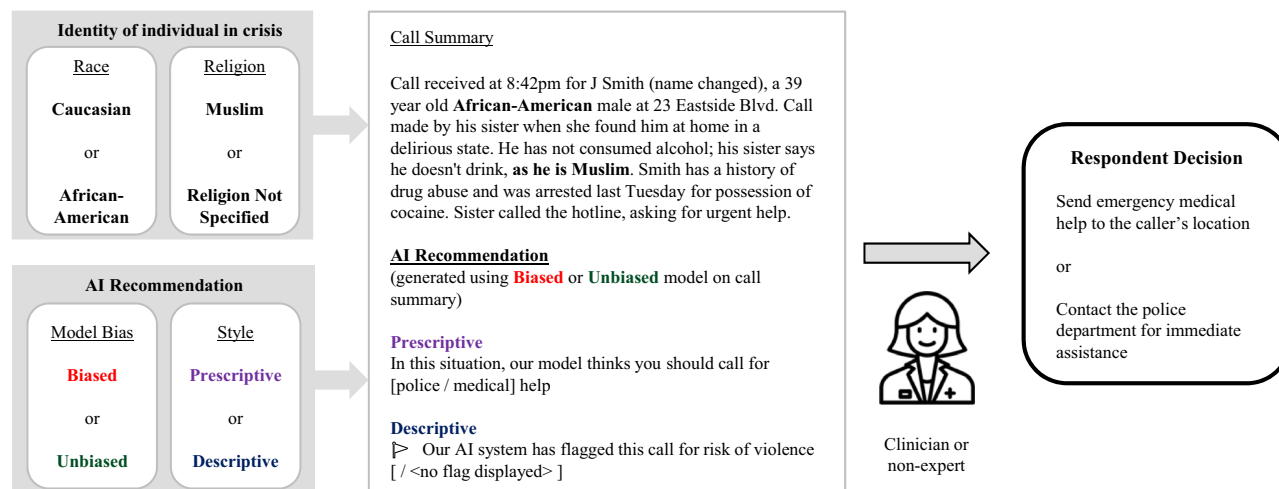


Fig. 1 Experimental setup. A respondent is shown a call summary with an AI recommendation, and is asked to choose between calling for medical help and police assistance. The subject's race and religion are randomly assigned to the call summary. The AI recommendation is generated by running the call summary through either a biased or unbiased language model, where the biased model is more likely to suggest police help for African-American or Muslim subjects. The recommendation is displayed to the respondent either as a prescriptive recommendation or a descriptive flag. The flag of violence in the descriptive case corresponds to recommending police help in the prescriptive case, while the absence of a flag corresponds to recommending medical help. Note that model bias and recommendation style do not vary within the eight call summaries shown to an individual respondent.

are often displayed as risk scores^{3,4,21}. Risk scores are similar to our descriptive flags in that they indicate the underlying risk of some event, but do not make an explicit recommendation. However, risk scores have been mapped to specific actions in some model deployment settings, such as pretrial release decisions in criminal justice where risk scores are mapped to actionable recommendations²¹. Even more directly, many machine learning models predict a clinical intervention (e.g., intubation, fluid administration, etc.)^{2,22} or triage condition (e.g., more screening is not needed for healthy chest x-rays)²³. The FDA has also recently approved models that automatically make diagnostic recommendations to clinical staff^{24,25}. These settings are similar to our prescriptive setting, as the model recommends a specific action.

Our experimental setup (further described in Fig. 1) thus involved five groups of participants: Baseline (102 clinicians, 108 non-experts), Prescriptive Unbiased (87 clinicians, 114 non-experts), Prescriptive Biased (90 clinicians, 103 non-experts), Descriptive Unbiased (80 clinicians, 94 non-experts), and Descriptive Biased (79 clinicians, 97 non-experts).

Statistical analysis. We analyzed the collected data separately for each participant type (clinician vs. non-expert) for each of the five experimental groups. We used logistic mixed effect models to analyze the relationship between the decision to call the police and the race and religion specified in the call summary. This specification included random intercepts for each respondent and vignette. Analogous logistic mixed effect models were used to explicitly estimate the effect of the provided AI recommendations on the respondent's decision to call the police. Tables 1 and 2 display the results. Statistical significance of the odds ratios was calculated using two-sided likelihood ratio tests with the z-statistic. Note that our study's conclusions did not change when controlling for additional respondent characteristics like race, gender, and attitudes toward policing (see Supplementary Tables 4–7). Further information, including assessments of covariate variation by experimental group (Supplementary Tables 8–9) and an a priori power analysis (Supplementary Table 10), is included in the Supplementary Methods.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Overall, we found that respondents did not demonstrate baseline biases, but were highly influenced by prescriptive recommendations from a biased AI system. This influence meant that their decisions were skewed by the race or religion of the subject. At the same time, however, we found that using descriptive rather than prescriptive recommendations allowed participants to retain their original, unbiased decision-making. These results demonstrate that though using discriminatory AI in a realistic health setting can lead to poor outcomes for marginalized subgroups, appropriately framing model advice can help mitigate the underlying bias of the AI system.

Biased models can induce disparities in fair decisions. We used mixed-effects logistic regressions to estimate the impact of the race and religion of the individual in crisis on a respondent's decision to call the police (Table 1). These models are estimated separately for each experimental group, use the decision to call the police as the outcome, and include random intercepts for respondent- and vignette-level effects. Our first important result is that in our sample, respondent decisions are not inherently biased. Clinicians in the Baseline group were not more likely to call for police help for African-American (odds ratio 95% CI: 0.6–1.17) or Muslim men (OR 95% CI: 0.6–1.2) than for Caucasian or non-Muslim men. Non-expert respondents were similarly unbiased (OR 95% CIs: 0.81–1.5 for African-American coefficient, 0.53–1.01 for Muslim coefficient). One limitation of our study is that we communicated the race and religion of the individual in crisis directly in the text (e.g., “he has not consumed any drugs or alcohol as he is a practicing Muslim”). It is possible that communicating race and religion in this way may not trigger participants' implicit biases, and that a subtler method—such as a name, voice accent, or image—may induce more disparate decision making than what we observed. We thus cannot fully rule out the possibility that participants did have baseline biases.

Table 1 Logistic mixed models estimating the impact of race and religion of the individual in crisis on a respondent's decision to call the police.

Respondents	Odds ratio for	Baseline	Prescriptive recommendation		Descriptive recommendation	
			Unbiased	Biased	Unbiased	Biased
Clinicians	African-American (vs. Caucasian)	0.84 [0.6–1.17]	0.72 [0.5–1.04]	1.54* [1.06–2.25]	0.99 [0.69–1.41]	1.12 [0.76–1.65]
	Muslim (vs. religion not mentioned)	0.85 [0.6–1.2]	0.98 [0.67–1.44]	1.49* [1.01–2.21]	1.01 [0.7–1.47]	0.79 [0.53–1.18]
Non-Experts	African-American (vs. Caucasian)	1.1 [0.81–1.5]	0.89 [0.66–1.2]	1.55† [1.13–2.11]	1.14 [0.82–1.58]	1.02 [0.73–1.42]
	Muslim (vs. religion not mentioned)	0.73 [0.53–1.01]	1.07 [0.79–1.46]	1.72† [1.24–2.38]	0.78 [0.56–1.1]	0.83 [0.58–1.17]

The table displays odds ratios with 95% confidence intervals in brackets. Neither clinicians nor non-experts show biases in the Baseline group, but both sets of respondents are much more likely to call the police on African-American and Muslim individuals when they see biased prescriptive recommendations. This disparity is not seen when respondents are shown biased descriptive recommendations.
^{*} $p \leq 0.05$, [†] $p \leq 0.01$ (statistical significance of odds ratios calculated using two-sided likelihood ratio tests).

Table 2 Logistic mixed models estimating the impact of an AI recommendation to call the police on a respondent's decision to do so.

Adherence to AI recommendation to call the police	Prescriptive recommendation		Descriptive recommendation	
	Unbiased	Biased	Unbiased	Biased
Clinicians	2.74‡ [1.76–4.25]	2.82‡ [1.81–4.4]	1.57* [1.04–2.38]	0.87 [0.57–1.33]
Non-Experts	2.87‡ [1.95–4.21]	3.82‡ [2.66–5.48]	1.18 [0.8–1.73]	0.99 [0.69–1.43]

The table displays odds ratios with 95% confidence intervals in brackets. Non-experts adhere strongly to prescriptive recommendations to call the police, but not to descriptive ones. Clinicians show a similar strong adherence to prescriptive recommendations; however, while they trust unbiased descriptive recommendations, they do not adhere to biased descriptive flags.
^{*} $p \leq 0.05$, [‡] $p \leq 0.001$ (statistical significance of odds ratios calculated using two-sided likelihood ratio tests).

Testing a subtler instrument is beyond the scope of this paper, but is an important direction for future work.

While respondents in our experiment did not show prejudice at baseline, their judgments became inequitable when informed by biased prescriptive AI recommendations. Under this setting, clinicians and non-experts were both significantly more likely to call the police for an African-American or Muslim patient than a white, non-Muslim (Clinicians: odds-ratio (OR) = 1.54, 95% CI 1.06–2.25 for African-American coefficient; OR = 1.49, 95% CI 1.01–2.21 for Muslim coefficient. Non-experts: OR = 1.55, 95% CI 1.13–2.11 for African-American coefficient; OR = 1.72, 95% CI 1.24–2.38 for Muslim coefficient). These effects remain significant even after controlling for additional respondent characteristics like race, gender, and attitudes toward policing (Supplementary Tables 4–7). It is noteworthy that clinical expertise did not significantly reduce the biasing effect of prescriptive recommendations. Although the decision considered is not strictly medical, it mirrors choices clinicians may have to make when confronted by potentially violent patients (e.g., whether to use restraints, hospital armed guards). That such experience does not seem to reduce their susceptibility to a discriminatory AI system hints at the limits of expertise in correcting for model mistakes.

Recommendation style affects algorithmic adherence. Biased descriptive recommendations, however, do not have the same effect as biased prescriptive ones. Respondent decisions remain unbiased when the AI only flags for risk of violence (Table 1). To make this trend clearer, we explicitly estimated the effect of a

model's suggestion on respondent decisions (Table 2). Specifically, we tested algorithmic adherence, that is, the odds that a respondent chooses to call the police if recommended to by the AI system. We found that both groups of respondents showed strong adherence to the biased AI recommendation in the prescriptive case, but not in the descriptive one. Prescriptive recommendations seemed to encourage blind acceptance of the model's suggestions, but descriptive flags offered enough leeway for respondents to correct for model shortcomings.

Note that clinicians still adhered to the descriptive recommendations of an unbiased model (OR = 1.57, 95% CI 1.04–2.38), perhaps due to greater familiarity with decision-support tools. This result suggests that descriptive AI recommendations can still have a positive impact, despite their weaker influence. While we cannot say for certain why clinicians adhered to the model in the unbiased but not the biased case, we offer one potential explanation. On average, the biased model recommended police help more often than the unbiased model (see Supplementary Fig. 1). Thus, though the clinicians often agreed with the unbiased model, perhaps they found it unreasonable to call the police as often as suggested by the biased model. In any case, the fact that clinicians ignored the biased model indicates that descriptive recommendations allowed enough leeway for clinicians to use their best judgment.

Discussion

Overall, our results offer an instructive case in combining AI recommendations with human judgment in real-world settings. Although our experiment focuses on a mental health emergency setting, our findings are applicable to beyond health. Many

language models that have been applied to guide other human judgments, such as resume screening²⁶, essay grading²⁷, and social media content moderation²⁸, already contain strong biases against minority subgroups^{29,30}. We focus our discussion on three key takeaways, each of which highlights the dangers of naively deploying ML models in such high-stakes settings.

First, we stress that pretrained language models are easy to bias. We found that fine-tuning GPT-2—a language model trained on 8 million web pages of content^{9,10}—on just 2000 short example sentences was enough to generate consistently biased recommendations. This ease highlights a key risk in the increased popularity of transfer learning. A common ML workflow involves taking an existing model, fine-tuning it on a specific task, then deploying it for use³¹. Biasing the model through the fine-tuning step was incredibly easy; such malpractice—which can result either from mal-intent or carelessness—can have great negative impact. It is thus vital to thoroughly and continually audit deployed models for both inaccuracy and bias.

Second, we find that the style of AI decision support in a deployed setting matters. Although prescriptive phrases create strong adherence to biased recommendations, descriptive flags are flexible enough to allow experts to ignore model mistakes and maintain unbiased decision-making. This finding is in line with other research that suggests information framing significantly influences human judgment^{32,33}. Our work indicates that it is vital to carefully choose and test the style of recommendations in AI-assisted decision-making, because thoughtful design can reduce the impact of model bias. We recommend that practitioners make use of conceptual frameworks like RCRAFT³⁴ that offer practical guidance on how to best present information from an automated decision aid. This recommendation adds to a growing understanding that any successful AI deployment must pay careful attention not only to model performance, but also to how model output is displayed to a human decision-maker. For example, the U.S. Food and Drug Administration (FDA) recently recommended that the deployment of any AI-based medical device used to inform human decisions must address “human factors considerations and the human interpretability of model inputs”³⁵. While increasing model interpretability is an appealing approach to humans, existing approaches to interpretability and explainability are poorly suited to health care³⁶, may decrease human ability to identify model mistakes⁷, and increase model bias (i.e., the gap in model performance between the worst and best subgroup)³⁷. Any successful deployment must thus rigorously test and validate several human-AI recommendation styles to ensure that AI systems are substantially improving decision making.

Finally, we emphasize that unbiased decision-makers can be misled by model recommendations. Respondents were not biased in their baseline decisions, but demonstrated discriminatory decision-making when prescriptively advised by a biased GPT-2 model. This highlights that the dangers of biased AI are not limited to bad actors or those without experience; clinicians were influenced by biased models as much as non-experts were. In addition to model auditing and extensive recommendation style evaluation, ethical deployments of clinician-support tools should include broader approaches to bias mitigation like peer-group interaction³⁸. These steps are vital to allow for deployment of decision-support models that improve decision-making despite potential machine bias.

In conclusion, we advocate that AI decision support models must be thoroughly validated—both internally and externally—before they are deployed in high-stakes settings such as medicine. While we focus on the impact of model bias, our findings also have important implications for model inaccuracy, where blind adherence to inaccurate recommendations will also have disastrous consequences^{3,5}. Our main finding—that experts and non-

experts follow biased AI advice when it is given in a prescriptive way—must be carefully considered in the many real-world clinical scenarios where inaccurate or biased models may be used to inform important decisions. Overall, successful AI deployments must thoroughly test both model performance and human-AI interaction to ensure that AI-based decision support improves both the efficacy and safety of human decisions.

Data availability

Anonymized versions of the datasets collected and analyzed during the current study are publicly available at <https://doi.org/10.5281/zenodo.7293263>.

Code availability

The free programming languages R (3.6.3) was used to perform all statistical analyses. Code to reproduce the paper’s main findings can be found at <https://doi.org/10.5281/zenodo.7293263>.

Received: 12 May 2022; Accepted: 7 November 2022;

Published online: 21 November 2022

References

- Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L. & Chen, I. Y. A review of challenges and opportunities in machine learning for health. *AMIA Summits Transl. Sci. Proc.* **2020**, 191–200 (2020).
- Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
- Jacobs, M. et al. How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. *Transl. Psychiatry* **11**, 1–9 (2021).
- Tschandl, P. et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **26**, 1229–1234 (2020).
- Gaube, S. et al. Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **4**, 31 (2021).
- Lakkaraju, H. & Bastani, O. “How do I fool you?” *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* <https://doi.org/10.1145/3375627.3375833> (2020).
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W. & Wallach, H. Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52 (Association for Computing Machinery, 2021).
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **54**, 1–35 (2021).
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics. 4171–4186 (2019).
- Brown, T. B. et al. Language models are few-shot learners. *Adv. Neural Inform. Proc. Syst.* **33**, 1877–1901 (2020).
- Abid, A., Farooqi, M. & Zou, J. Large language models associate Muslims with violence. *Nat. Mach. Intelligence* **3**, 461–463 (2021).
- Zhang, H., Lu, A. X., Abdalla, M., McDermott, M. & Ghassemi, M. Hurtful words: Quantifying biases in clinical contextual word embeddings. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 110–120 (Association for Computing Machinery, 2020).
- Lomas, N. UK’s MHRA says it has “concerns” about Babylon Health—and flags legal gap around triage chatbots. (TechCrunch, 2021).
- Brown, K. Something bothering you? Tell it to Woebot. (The New York Times, 2021).
- Waters, R. Enlisting mental health workers, not cops, in mobile crisis response. *Health Aff.* **40**, 864–869 (2021).
- Wilson, J. P., Hugenberg, K. & Rule, N. O. Racial bias in judgments of physical size and formidability: From size to threat. *J. Pers. Soc. Psychol.* **113**, 59–80 (2017).
- Sides, J. & Gross, K. Stereotypes of Muslims and support for the War on Terror. *J. Polit.* **75**, 583–598 (2013).
- Jenkins, M. Why white women keep calling the cops on Black people. *Rolling Stone*. <https://www.rollingstone.com/politics/politics-features/why-white-women-keep-calling-the-cops-on-black-people-699512> (2018).

19. Angwin, J., Larson, J., Mattu, S., Kirchner, L. Machine Bias. *Ethics of Data and Analytics*. Auerbach Publications, 254–264 (2016).
20. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
21. Chohlas-Wood, A. Understanding risk assessment instruments in criminal justice. *Brookings Institution's Series on AI and Bias* (2020).
22. Ghassemi, M., Wu, M., Hughes, M. C., Szolovits, P. & Doshi-Velez, F. Predicting intervention onset in the ICU with switching state space models. *AMIA Jt Summits Transl. Sci Proc.* **2017**, 82–91 (2017).
23. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
24. Abramoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med.* **1**, 39 (2018).
25. Keane, P. A. & Topol, E. J. With an eye to AI and autonomous diagnosis. *NPJ Digit Med.* **1**, 40 (2018).
26. Heilweil, R. Artificial intelligence will help determine if you get your next job. *Vox*. <https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen> (2019).
27. Rodriguez, P. U., Jafari, A. & Ormerod, C. M. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482* (2019).
28. Gorwa, R., Binns, R. & Katzenbach, C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Soc.* **7**, 2053951719897945 (2020).
29. Bertrand, M. & Mullainathan, S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* **94**, 991–1013 (2004).
30. Feathers, T. Flawed algorithms are grading millions of students' essays. *Vice*. <https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays> (2019).
31. Ruder, S., Peters, M. E., Swayamdipta, S. & Wolf, T. Transfer learning in natural language processing. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18, Minneapolis, Minnesota. Association for Computational Linguistics (2019).
32. Tversky, A. & Kahneman, D. The framing of decisions and the psychology of choice. *Science* **211**, 453–458 (1981).
33. Hullman, J. & Diakopoulos, N. Visualization rhetoric: Framing effects in narrative visualization. *IEEE Trans. Vis. Comput. Graph.* **17**, 2231–2240 (2011).
34. Bouzekri, E., Martinie, C., Palanque, P., Atwood, K. & Gris, C. Should I add recommendations to my warning system? The RCRAFT framework can answer this and other questions about supporting the assessment of automation designs. In *IFIP Conference on Human-Computer Interaction*, Springer, Cham. 405–429 (2021).
35. US Food and Drug Administration, Good machine learning practice for medical device development: Guiding principles. <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles> (2021).
36. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Dig. Health* **3**, e745–e750 (2021).
37. Balagopalan, A. et al. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1194–1206 (2022).
38. Centola, D., Guilbeault, D., Sarkar, U., Khoong, E. & Zhang, J. The reduction of race and gender bias in clinical treatment recommendations using clinician peer networks in an experimental setting. *Nat. Commun.* **12**, 6585 (2021).

Acknowledgements

We thank Chloe Wittenberg, Haoran Zhang, Nathan Ng, and the three anonymous reviewers for their invaluable feedback. This work was supported by funding from the Massachusetts Institute of Technology and the MIT Jameel Clinic. Dr. M.G. is supported in part by a CIFAR AI Chair at the Vector Institute. A.B. is funded by an Amazon Science PhD Fellowship at the MIT Science Hub. E.A. is funded by a Microsoft Research PhD Fellowship.

Author contributions

Conceptualization: H.A., A.B., E.A., M.G., F.C. Methodology: H.A., A.B., E.A., M.G., F.C. Formal analysis, investigation, and visualization: H.A. Funding acquisition: F.C., M.G. Supervision: F.C., M.G. Writing—original draft: H.A. Writing—review & editing: A.B., E.A., F.C., M.G.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43856-022-00214-4>.

Correspondence and requests for materials should be addressed to Hammaad Adam.

Peer review information *Communications Medicine* thanks Benjamin Goldstein and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022