Fairness via AI: Bias Reduction in Medical Information

SHIRI DORI-HACOHEN, University of Connecticut & AuCoDe, USA
ROBERTO MONTENEGRO, Seattle Children's Hospital, USA
FABRICIO MURAI, Universidade Federal de Minas Gerais, Brazil
SCOTT A. HALE, Meedan, USA
KEEN SUNG, AuCoDe, USA
MICHELA BLAIN, University of Washington School of Medicine, USA
JENNIFER EDWARDS-JOHNSON, Michigan State University College of Human Medicine, USA

Additional Key Words and Phrases: fairness in AI, health misinformation, bias reduction

1 Introduction

Most Fairness in AI research focuses on exposing biases in AI systems. A broader lens on fairness reveals that AI can serve a greater aspiration: rooting out societal inequities from their source. Specifically, we focus on inequities in health information, and aim to reduce bias in that domain using AI. The AI algorithms under the hood of search engines and social media, many of which are based on recommender systems, have an outsized impact on the quality of medical and health information online. Therefore, embedding bias detection and reduction into these recommender systems serving up medical and health content online could have an outsized positive impact on patient outcomes and wellbeing.

In this position paper, we offer the following contributions: (1) we propose a novel framework of Fairness **via** AI, inspired by insights from medical education, sociology and antiracism; (2) we define a new term, **bisinformation**, which is related to, but distinct from, misinformation, and encourage researchers to study it; (3) we propose using AI to study, detect and mitigate biased, harmful, and/or false health information that disproportionately hurts minority groups in society; and (4) we suggest several pillars and pose several open problems in order to seed inquiry in this new space. While part (3) of this work specifically focuses on the health domain, the fundamental computer science advances and contributions stemming from research efforts in bias reduction and Fairness via AI have broad implications in all areas of society.

2 Fairness via AI

The vast majority of Fairness in AI work focuses on exposing the bias in AI system, in order to showcase where the AI system is biased. However, AI systems will continue to be biased so long as the data they are receiving are biased, according to "Bias In, Bias Out" principle [16]. So long as significant structural inequalities exist in the real world, AI systems will continue to replicate and exacerbate them. The dominant Fairness in AI approach, then, risks engaging in a Sisyphean task of minimizing bias in AI, with attempts to debias AI datasets, models and algorithms continually needing to be "fixed" as they learn biased outcomes and are bound to hit a ceiling of fairness: that of real world settings that are inherently biased. Under this approach, our highest aspiration in designing AI systems seems to be one of avoidance: tweaking our models to refrain from adding to society's ills and inequities.

Authors' addresses: Shiri Dori-Hacohen, shiridh@uconn.edu, University of Connecticut & AuCoDe, USA; Roberto Montenegro, Seattle Children's Hospital, USA; Fabricio Murai, Universidade Federal de Minas Gerais, Brazil; Scott A. Hale, Meedan, USA; Keen Sung, AuCoDe, USA; Michela Blain, University of Washington School of Medicine, USA; Jennifer Edwards-Johnson, Michigan State University College of Human Medicine, USA.

Rooted in insights from medical education, sociology, and antiracism, we offer a broader lens on fairness, revealing that AI can serve a far greater aspiration: enabling important restorative work and rooting out societal inequities from their source, with a deeper and more meaningful impact. In this position paper, we therefore reverse the traditional direction of fairness: rather than aiming to achieve Fairness **in** or **of** AI, we propose focusing on **Fairness via AI**. With this approach, one can use AI to study, detect, mitigate and remedy situations that are inherently unequal, unjust and unfair in society. With this ambitious yet grounded approach, our potential impact is unbounded, and can accelerate progress towards a more fair, equal and just world. In other words, we can use AI to thoughtfully, carefully and ethically debias the world, rather than simply trying to debias AI. Specifically, the AI algorithms under the hood of search engines and social media, many of which are based on recommender systems, have an outsized impact on the quality of information available online. Therefore, embedding bias detection and reduction directly into these recommender systems could have an outsized positive impact on the information ecosystem.

3 Bisinformation

We coin the term **bisinformation** to represent biased information, referring to a unique and challenging aspect of the information landscape. We are particularly interested in health bisinformation, where bias and language misuse have a detrimental impact on patient outcomes, though the term can easily apply in any field. Bisinformation may overlap with, but is not identical to, misinformation. The use of biased language or inappropriate social identifiers in a medical context, for example, can be harmful even if strictly true - consider the case of referring to the prevalence of an illness in a racial category without contextualizing it in Social or Structural Determinants of Health (SSDoH), such as systemic racism or income inequities [17]. On the other hand, certain types of bisinformation are, in fact, also a form of misinformation, such as the long-debunked Salt Gene Hypothesis [19].

To the best of our knowledge, no studies have computationally studied health bisinformation at a large scale.

Medical bisinformation (and misinformation). As an illustration of societal inequities in dire need of the Fairness via AI approach, consider the field of medicine and medical education. The field is marred by a long and painful history of overt and covert forms of social injustice, bias, and racism, as illustrated by the American Medical Association's recent pledge to take action to confront systemic racism [14, 22]. Studies continue to demonstrate that physicians possess implicit biases in a number of different areas such as race/ethnicity, gender, sex, age, weight, substance use and mental illness [7]. This comes into play significantly in medical institutions, which continue to teach biased medicine in preclinical years [see, e.g. 23]. Many educators, for example, continue to inappropriately use race as a proxy for genetics or ancestry, or even as a "risk factor" for numerous health outcomes often erroneously associated with race while ignoring SSDoH [1, 2, 8, 11, 17]. Many educators continue to inappropriately use gender and sex terms and perpetuate the idea that sex and gender are binary and stagnant (versus fluid). Likewise, most medical educators are unaware of the numerous biases in the types of images they use in their lectures or assessment materials as well [5, 6, 13]. By equating social identifiers to biology without social or structural context, medical educators are unknowingly perpetuating a curriculum that can have an adverse effect on health outcomes [4, 15]. Bias reduction in curricular and assessment content is critical for educating future physicians in accurate evidence-based medicine [12, 21], but is a manual, costly and time consuming effort. SOTA AI and NLP approaches can be used to scale up these efforts significantly[1].

Naturally, bisinformation and misinformation that persist in the medical establishment, are also disseminated and extensively present in online medical resources, websites and news articles, and social media, with large negative effects on historically underserved populations, also reinforcing biased narratives and stereotypes about minority groups.

[1]Montenegro, Dori-Hacohen and Sung have recently been awarded a grant from the National Board of Medical Examiners' Stemmler Fund to reduce bias in medical curricular content using NLP and ML approaches.

Recommender systems play an outsized role in serving up such content online, but improving these systems to reduce bias will prove extremely challenging if we don't understand the underlying mechanisms in which such bias is perpetuated and disseminated. Prior work has suggested that controversy online is highly unevenly distributed, and that a population-sensitive model is needed in order to properly model this [10]; we hypothesize that the same approach may be needed in the computational study of bisinformation and misinformation. For example, the COVID-19 pandemic and its associated "infodemic" has brought health mis- and disinformation to the forefront of national and scientific attention. However, trust in the medical establishment may be understandably low among African-Americans [3] and other minority groups [9]. Recent work suggests, moreover, that health mis- and disinformation is qualitatively distinct in different population groups [e.g. 9, 20]. To cite just one example, COVID-19 vaccine hesitancy has been demonstrated to be higher among racial and ethnic minorities [9, 18].

4 Pillars & Open Problems

A few guiding pillars underlie and drive our *Fairness via AI* research agenda, which we encourage others to adopt. First, we argue that Fairness via AI is a more effective and impactful marshalling of research resources than "standard" Fairness in AI work (important though the latter may be). Second, we argue that collaboration across disciplinary fields is critically needed in order to effectively and ethically study and understand society's biggest challenges, to say nothing of mitigating them. Researchers in other fields, including but in no way not limited to the social sciences, have immense expertise in studying and addressing societal inequities; computer scientists cannot, and should not, go this alone. Finally, we argue that biased information interacts with false information in complex ways that must be studied carefully in order to reduce bias in recommender systems and other information delivery systems (such as search engines).

With these pillars firmly in mind, we pose the following open questions:

Q1. What are effective approaches to identify societal problems that are in most need of, and lend themselves to, the **Fairness via AI** framework?

Q2. Which existing and/or novel AI approaches need to be deployed and developed in order to address such societal issues?

Q3. How can we encourage collaboration across disciplinary boundaries in order to leverage hard-won insights from other fields, and infuse our Fairness research with them?

Q4. Specifically with respect to bisinformation, several research questions arise:

   a. How and where does bisinformation spread online? Is information (including mis- and bisinformation) distributed and disseminated differentially among diverse population groups? If so, how?

   b. Which categories or types of bisinformation and misinformation are most problematic and harmful, and thus deserving the most diligent fact checking, and countermessaging efforts? In other words, how should we triage mis- and bisinformation, combining best practices in the public health and fact checking spheres with state-of-the-art computational approaches?

5 Conclusions

In this position paper, we introduced a novel **Fairness via AI** framework; coined a new term, bisinformation, to describe biased information, and demonstrated that is is overlapping with yet distinct from misinformation; briefly discussed the documented presence of bisinformation in medical curricula and posit that this extends to other information environments, such as online; and posed several open questions to guide research agendas on the subject.

Acknowledgements. This material is based in part upon work supported by the National Science Foundation under Grant No. 1951091. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Kimberly D Acquaviva and Matthew Mintz. 2010. Perspective: are we teaching racial profiling? The dangers of subjective determinations of race and ethnicity in case presentations. *Academic Medicine* 85, 4 (2010), 702–705.

[2] Sarah E Ali-Khan, Tomasz Krakowski, Rabia Tahir, and Abdallah S Daar. 2011. The use of race, ethnicity and ancestry in human genetic research. *The HUGO journal* 5, 1 (2011), 47–63.

[3] Dwayne T Brandon, Lydia A Isaac, and Thomas A LaVeist. 2005. The legacy of Tuskegee and trust in medical care: is Tuskegee responsible for race differences in mistrust of medical care? *Journal of the National Medical Association* 97, 7 (2005), 951.

[4] Lundy Braun, Anne Fausto-Sterling, Duana Fullwiley, Evelynn M Hammonds, Alondra Nelson, William Quivers, Susan M Reverby, and Alexandra E Shields. 2007. Racial categories in medical practice: how useful are they? *PLoS medicine* 4, 9 (2007), e271.

[5] Asif Doja, M Dylan Bould, Chantalle Clarkin, Kaylee Eady, Stephanie Sutherland, and Hilary Writer. 2016. The hidden and informal curriculum across the continuum of training: a cross-sectional qualitative study. *Medical teacher* 38, 4 (2016), 410–418.

[6] Keisa Fallin-Bennett. 2015. Implicit bias against sexual minorities in medicine: cycles of professional influence and the role of the hidden curriculum. *Academic Medicine* 90, 5 (2015), 549–552.

[7] Chloë FitzGerald and Samia Hurst. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics* 18, 1 (2017), 1–18.

[8] Linda M Hunt, Nicole D Truesdell, and Meta J Kreiner. 2013. Genes, race, and culture in clinical care: racial profiling in the management of chronic illness. *Medical anthropology quarterly* 27, 2 (2013), 253–271.

[9] J Jaiswal, C LoSchiavo, and DC Perlman. 2020. Disinformation, misinformation and inequality-driven mistrust in the time of COVID-19: lessons unlearned from AIDS denialism. *AIDS and Behavior* 24 (2020), 2776–2780.

[10] Myungha Jang, Shiri Dori-Hacohen, and James Allan. 2017. Modeling Controversy within Populations. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands). Association for Computing Machinery, New York, NY, USA, 141–149. https://doi.org/10.1145/3121050.3121067 ICTIR '17.

[11] Reena Karani, Lara Varpio, Win May, Tanya Horsley, John Chenault, Karen Hughes Miller, and Bridget O'Brien. 2017. Commentary: racism and bias in health professions education: how educators, faculty developers, and researchers can make a difference. *Academic Medicine* 92, 11S (2017), S1–S6.

[12] Tao Le, Vikas Bhushan, Matthew Sochat, Kimberly Kallianos, Yash Chavda, Andrew Harrison Zureick, and Mehboob Kalani. 2018. *First Aid for the USMLE Step 1 2018*. McGraw-Hill Medical.

[13] Heidi Lempp and Clive Seale. 2004. The hidden curriculum in undergraduate medical education: qualitative study of medical students' perceptions of teaching. *Bmj* 329, 7469 (2004), 770–773.

[14] J. Madara. 2020. America's Health Care Crisis Is Much Deeper than COVID-19. https://www.ama-assn.org/about/leadership/america-s-health-care-crisis-much-deeper-covi

[15] Maria Athina Tina Martimianakis, Barret Michalec, Justin Lam, Carrie Cartmill, Janelle S Taylor, and Frederic W Hafferty. 2015. Humanism, the hidden curriculum, and educational reform: a scoping review and thematic analysis. *Academic Medicine* 90, 11 (2015), S5–S13.

[16] Sandra G Mayson. 2018. Bias in, bias out. *YAle lJ* 128 (2018), 2218.

[17] Jonathan M Metzl and Dorothy E Roberts. 2019. Structural competency meets structural racism: race, politics, and the structure of medical knowledge. In *The Social Medicine Reader, Volume II, Third Edition*. Duke University Press, 170–187.

[18] Long H. Nguyen, Amit D. Joshi, David A. Drew, Jordi Merino, Wenjie Ma, Chun-Han Lo, Sohee Kwon, Kai Wang, Mark S. Graham, Lorenzo Polidori, Cristina Menni, Carole H. Sudre, Adjoa Anyane-Yeboa, Christina M. Astley, Erica T. Warner, Christina Y. Hu, Somesh Selvachandran, Richard Davies, Denis Nash, Paul W. Franks, Jonathan Wolf, Sebastien Ourselin, Claire J. Steves, Tim D. Spector, Andrew T. Chan, and on behalf of the COPE Consortium. 2021. Racial and ethnic differences in COVID-19 vaccine hesitancy and uptake. *medRxiv* (2021). https://doi.org/10.1101/2021.02.25.21252402 arXiv:https://www.medrxiv.org/content/early/2021/02/28/2021.02.25.21252402.full.pdf

[19] Anne Pollock. 2012. 4. The Slavery Hypothesis beyond Genetic Determinism. In *Medicating Race*. Duke University Press, 107–130.

[20] Amit Prasad. 2021. Anti-science Misinformation and Conspiracies: COVID–19, Post-truth, and Science & Technology Studies (STS). *Science, Technology and Society* (2021). https://doi.org/10.1177/09717218211003413

[21] Kelsey Ripp and Lundy Braun. 2017. Race/ethnicity in medical education: an analysis of a question bank for step 1 of the United States Medical Licensing Examination. *Teaching and learning in medicine* 29, 2 (2017), 115–122.

[22] Angela Saini. 2019. *Superior: the return of race science*. Beacon Press.

[23] Jennifer Tsai, Laura Ucik, Nell Baldwin, Christopher Hasslinger, and Paul George. 2016. Race matters? Examining and rethinking race portrayal in preclinical medical education. *Academic Medicine* 91, 7 (2016), 916–920.

This figure "sample-franklin.png" is available in "png" format from:

http://arxiv.org/ps/2109.02202v1