Bias in Healthcare AI - A Literature Review

Olivia S. Droob 23V0013
Martha F. Nybroe 23V0014
Govind Saju 200070020
Pranava Singhal 200070057
Prajyot Kore 22N0044

Abstract

This article presents a literature review on bias in artificial intelligence for healthcare. We begin our review by comparing different definitions of bias and fairness and analyze the diverse causes and consequences of bias in AI for healthcare. We then turn to the possible strategies to mitigate these consequences, including debiasing data sets and developing less prejudiced models and regulatory frameworks to deal with ethical standards and liability issues. Finally, we highlight knowledge gaps in the existing literature and outline prospective areas for future research. We conclude that sustainable debiasing requires a holistic and deep understanding of both AI and the societal issues of inequality and diversity.

1 Introduction

The utilization of AI in the healthcare industry has significant potential for increasing the accuracy of diagnostics, the effectiveness of various treatment methodologies, and the overall quality of patient care. However, the prevalence of biases in the underlying AI models used for these purposes raise a challenge in ensuring equitable access to healthcare. Understanding and mitigating bias in the AI used for medicinal purposes is necessary to provide fair and unbiased treatment across diverse patient demographics. In this literature survey, we present our findings on the current academic literature on medical AI. We first analyze the current state of bias in medical AI and then present the various technical and regulatory solutions for mitigating bias.

2 Understanding Bias

2.1 Defining Bias

Fairness in AI has been defined as the "absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" (Mehrabi et al., 2021). Bias in AI can be characterised as a lack of fairness in an AI system, such as gender bias or racial bias. There are two main notions of what fairness constitutes in AI systems, namely group fairness and individual fairness. Group fairness refers to a notion of fairness where statistical parity is ensured for people of certain protected groups (such as gender or race), while individual fairness refers to an idea where people who are similar with respect to the task at hand receive similar outcomes (Binns, 2020).

2.2 Sources of Bias

It is agreed that the sources of bias in artificial intelligence are plentiful, entangled, and linked to societal inequality (Eubanks, 2018) (Ferrara, 2023). It is further agreed that biased AI models are a product of biased data. Thus, much research interest lies in understanding why and how datasets are tarnished by the biases and inequalities of society. Barocas and Selbst argues that biased data can stem from inheriting the prior decision makers' biased decision patterns and through the less clear paths in which societal prejudices make their way into the data. However, these paths remain largely unexplained in the paper and the field at large, where biased data is taken as a given (Kaushal et al., 2020) (B et al., 2021).

An important consideration, which many studies overlook when studying biased data, is one concerning the intersectionality of biases raised by (Buolamwini and Gebru, 2018). Many studies will focus solely on examining gender bias, racial bias, sexuality bias, etc., and may not examine how people in the intersections of multiple protected classes might be affected more severely than others. This raises the question of how to deal with intersectionality, which increases the need to further research the relations between gender, race, class, sexuality, and ableness.

Moreover, once biased data is given to a human, it is subject to the modeler's beliefs and consid-

erations. As pointed out by multiple studies, the gender parity (Celi et al., 2022) in the field of computer science necessarily leads to blind spots in the building of AI systems (Daraz et al., 2022) (Feuerriegel et al., 2020). Furthermore, (Wang et al., 2023) demonstrates how humans can prefer a biased AI system if it aligns better with their biased personal beliefs.

The sources of bias in AI models are thus accredited to biased data from a biased society perpetuating and even preferring biased decisions. Furthermore, it is clear that the intersectionality of protected attributes can not be overlooked and should be considered the natural framework for considerations of bias.

2.3 AI Applications in Healthcare

AI tools have seen widespread adoption for healthcare applications, including diagnostic imaging for automated analysis of medical scans (Hafizović et al., 2021), clinical decision support systems (CDSSs)(Vijayakumar et al., 2023)(Magrabi et al., 2019), predictive analysis for patient outcomes and disease progression (Li et al., 2020), and personalised healthcare. Broadening the scope beyond diagnostic tools, Healthcare 5.0 (Saraswat et al., 2022) aims to achieve a fully autonomous healthcare service using AI, IoT and 5G, that takes into account the interdependent effect of different health conditions of a patient to design comprehensive personalized healthcare services (CPHS) (Taimoor and Rehman, 2021). This includes robots for personal care (Kyrarini et al., 2021), AI for automated drug discovery (Deng et al., 2022), healthcare Internet of Things (HIoT) (Taimoor and Rehman, 2021) devices for continuous patient monitoring, and tools for medical big-data management (Karatas et al., 2022). There has also been significant growth in patient-facing chatbots for medical assistance, preliminary diagnosis, and shortlisting critical patients to automate patient intake (Locke et al., 2021). Another exciting direction is personalised medicine, based on the P4 principles (personalised, predictive, preventive, participatory) framework (Vogt and Green, 2020), which promises to deliver diagnosis tailored to individual lifestyle, genetics, and medical history (Chan and Ginsburg, 2011).

2.4 Consequences of Biased Medical AI

In the realm of diagnostic imaging (Larrazabal et al., 2020) and (Seyyed-Kalantari et al., 2021) report a significant drop in model accuracy for X-ray systems predominantly affecting Hispanics and women, both underrepresented groups in available large public radiology datasets (Johnson et al., 2019)(Irvin et al., 2019). Automated analysis of scans is also used for triage, and lower priority for clinician visits is given to individuals falsely classified as healthy (Seyyed-Kalantari et al., 2021). Thus, minority groups face a double disadvantage by facing delays in getting clinician appointments and, more severely, in incorrect automated diagnosis of disease.

Biases may arise right at the source of data collection, even before the AI model has had a chance to train. For instance, pulse oximeter readings are known to be less accurate for darker skin (Feiner et al., 2007). Failing to collect accurate data from minority groups may lead to training data that does not accurately represent the whole population and may not only lead to poor performance for these groups but may also lead to failure in identifying groups that have the highest burden of illness and understanding how these differences come about (Bonevski et al., 2014). A biased data collection process denies marginalised individuals from participation in trials and safety checks for new AI-based healthcare tools too.

In (Vijayakumar et al., 2023), physicians acknowledge the improved personalization and reliability of AI-based CDSS systems for diagnosis and treatment. However, clinical deployment faces challenges due to limited trial data and concerns about patient safety liability. Patients, as noted in (Richardson et al., 2021), express worries about AI in healthcare, focusing on data privacy and the validation of AI correctness by doctors. Existing healthcare AI technologies encounter issues with low user confidence and adoption. To address this, there is a push for explainable AI (XAI) solutions (Saraswat et al., 2022) enabling clinicians and patients to comprehend model outcomes (Jin et al., 2022) and remove biases from existing models (Bourdon et al., 2021) enhancing transparency and trust in healthcare AI applications.

In conclusion, the consequences of biases in existing AI-based healthcare solutions include low physician and patient confidence for clinical deployment, lack of model explainability, especially for safety-critical applications, failure to correctly identify severely affected patient demographic groups, and exacerbated biases against traditionally underrepresented groups in accuracy of diagnosis

and even access to care.

3 Mitigating Bias

3.1 Technical Solutions

A less researched area is how the model selection process can reduce or mitigate bias. Different approaches have been suggested to mitigate this based on both group (Yan et al., 2020) and individual fairness criteria (Zafar et al., 2017). However, this strategy is limited by committing to a specific understanding of fairness and can also lead to models with lower accuracy.

Across various research papers, a comprehensive set of technical solutions has been proposed to address algorithmic accountability and fairness in AI systems. Firstly, adopting Explainable AI (XAI) techniques, such as LIME, enhances model interpretability for transparent and understandable results (Diakopoulos, 2022). Re-weighting data samples for different demographic groups is a promising solution to mitigate biases (Sweeney, 2013)(Barocas et al., 2016). Fairness-aware design patterns, introduced in the socio-technical system domain, recommend incorporating fairness explicitly into design templates (Crawford et al., 2016). The development of a debiased AI college major recommender encompasses algorithmic debiasing, fairness/bias explanations, and a human-AI co-training process for iterative refinement (Bogen et al., 2019). Practical interventions at the implementation level, such as adjusting model outputs based on fairness metrics, are proposed to enhance fairness in real-world predictive modeling applications (Diakopoulos, 2021). Mitigating bias in algorithmic hiring involves strategies like fairness-aware algorithms, blind recruitment processes, and continuous monitoring for bias detection and correction (Bogen et al., 2020). Algorithmic fairness metrics, such as demographic parity and equalized odds, are pivotal in assessing and addressing biases in AI systems (Crawford et al., 2016). Establishing accountability frameworks for organizations is advocated to ensure responsible AI practices (Diakopoulos, 2018). Best practices and policies for bias detection and mitigation, including continuous monitoring and transparency, are recommended to reduce consumer harms (Varshney et al., 2022). Collectively, these solutions provide a comprehensive approach toward fostering transparency, fairness, and accountability in AI systems.

3.2 Regulatory Solutions

Another way to mitigate bias is through regulation. (Ganapathi et al., 2022) says that regulators of medical devices are increasingly recognizing the significant challenges regarding bias in the medical field. The US Food and Drug Administration (FDA) released an Action plan in January 2021 that focuses on identifying and removing bias in AI models used in medicine (esp. point 4) (Vokinger et al., 2021). They also released ten guiding principles for good machine learning practice in October 2021 that further focus on managing bias and producing generalizable performance across different groups.

These principles all emphasize the importance of being aware of different kinds of bias when developing technology, especially for medical purposes, but (Ganapathi et al., 2022) argues that there is no evidence of the manufacturers adapting to these recommendations. The STANDING (standards for data diversity, inclusivity, and generalizability) Together Initiative has been created to make the industry follow these recommendations. It involves researchers from multiple fields, medical experts, patients, regulators, and policy-makers to ensure that every aspect of the problem is thoroughly researched. In late October 2023, they published a set of recommendations (The STANDING Together collaboration, 2023) to handle issues regarding the use of AI in healthcare technologies, specifically focusing on diversity and inclusivity. The recommendations are separated into "Recommendations for Documentation of Health Datasets" and "Recommendations for Use of Health Datasets." One of the main points regarding documentation is that a dataset should, if possible, include protected attributes such as gender, age, race, sexual orientation, etc. If it doesn't, it should state the reasons why. Additionally, it is mentioned that in the cases where this would result in a risk of identification for the participants, it should be provided at the aggregate level. The documentation should state all limitations to the dataset, especially any known or expected kind of bias, attempts to combat this, which data protection laws have been followed, what was done to protect the identities of the participants, the reason why the dataset was created, and who funded it. The recommendations for using health datasets focus on identifying specific groups of interest beforehand, either by collaborating with experts or discovering them through data

analysis. Then, the data users should evaluate the performance on different groups against the overall data to identify any unfairness. Additionally they should report on the limitations of the dataset as well as the differences in the intended use of the dataset against the actual use.

There is a big emphasis on the data collectors' responsibility in mitigating bias. According to (Ganapathi et al., 2022) the recommendations will help stakeholders to make informed decisions on how and what demographic data is collected. This excellent initiative focuses on using and developing AI in the medical field while being very aware of the potential risks and how to handle them.

Concerning Clinical Decision Support Systems (CDSS), the liability issue in patient injury (Maliha et al., 2021) arises due to inadequate frameworks for balancing safe clinical implementation with technical innovation. Current regulations hold medical practitioners liable for deviations from the standard of care and negligence, but uncertainties persist when CDSSs contribute to patient injury. Maliha et al. draws parallels with court cases involving practitioners distributing misleading medical literature, highlighting the uncertainty in CDSS-related liability. US courts have been hesitant to exempt practitioners from liability, yet the untested territory of AI CDSSs introduces uncertainties. The EU's AI Liability Directive (AILD) (Duffourc and Gerke, 2023) attempts to address AI liability but has been criticized (Duffourc and Gerke, 2023) for inadequately handling liability in cases of black box medicine.

While advocating for increased diversity in the AI pipeline is a common strategy to mitigate bias (Daraz et al., 2022), solely hiring more female AI engineers hasn't conclusively reduced bias (Cowgill et al., 2020). However, this should not be seen as defeatist but rather underscores the necessity of a comprehensive approach to debiasing that involves reviewing the entire pipeline.

4 Knowledge Gaps and Future Research

(B et al., 2021) acknowledges a gap in the effectiveness of current debiasing techniques, prompting the need for a standardized framework and that future research should focus on semantic-aware neural architectures for generating debiased embeddings. Similarly, (Buolamwini and Gebru, 2018) highlights risks in data mining applications and calls for comprehensive research on bias in social data, urging the development of fairness metrics. (Cowgill et al., 2020) reveals a significant knowledge gap in biases within the healthcare sector, emphasizing the necessity for research on mitigating biases and developing ethically sound algorithms in healthcare AI applications.

In the legal realm, (Duffourc and Gerke, 2023) emphasizes a knowledge gap in the liability framework for black-box medical AI systems. Addressing biases and errors in healthcare AI decision-making requires the development of comprehensive legal frameworks. (Ganapathi et al., 2022) proposes a multifaceted approach to tackle biases, emphasizing the importance of data diversity, inclusivity, and generalizability. The initiative sets standards through engagement with various stakeholders. (Vokinger et al., 2021) underlines the need for specific methodologies, including transparent data collection and de-biasing techniques, to enhance bias mitigation strategies in healthcare-related ML systems.

Moreover, (Wang et al., 2023) focuses on biases in college major recommendations, urging future research to extend efforts to diverse domains, particularly in healthcare. (Yan et al., 2020) explores biases in personality assessments, emphasizing the need to bridge gaps between personality assessment biases and their implications in healthcare AI. Finally, (Zafar et al., 2017) introduces a method offering a flexible trade-off between fairness and accuracy. Future research is encouraged to explore the application of these methods in healthcare-related scenarios. Collectively, these insights underscore the complexity of biases in AI systems and the imperative to advance research for fair and unbiased AI applications across diverse healthcare domains.

5 Conclusion

It's crucial to enhance data collection from underrepresented communities for balanced datasets to address healthcare AI biases. Standardizing the collection process and ensuring high-quality measurements, especially from marginalized groups, is essential. Understanding the potential impact of biases on social groups before model deployment is necessary. In safety-critical applications, safety assessments should consider diverse communities. Additionally, improving the healthcare model's explainability is vital for boosting confidence and adoption by physicians and patients.

References

Senthil Kumar B, Aravindan Chandrabose, and Bharathi Raja Chakravarthi. 2021. An overview of fairness in data – illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45, Kyiv. Association for Computational Linguistics.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2016. Big data's disparate impact. *Journal of Adrenaline*, 8(3):322–334.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review*, 104(3):671–732.

Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524.

M. Bogen, A. Hanna, and D. Hirsch. 2019. When biased humans meet debiased ai. *Journal of Adrenaline*, 15(3):345–367.

M. Bogen, A. Hanna, and D. Hirsch. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Journal of Adrenaline*, 10(4):432–451.

Billie Bonevski, Madeleine Randell, Chris Paul, Kathy Chapman, Laura Twyman, Jamie Bryant, Irena Brozek, and Clare Hughes. 2014. Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC medical research methodology*, 14:1–29.

Pascal Bourdon, Olfa Ben Ahmed, Thierry Urruty, Khalifa Djemal, and Christine Fernandez-Maloigne. 2021. Explainable ai for medical imaging: Knowledge matters. *Multi-faceted Deep Learning: Models and Data*, pages 267–292.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Leo Anthony Celi, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, Lama Moukheiber, Julian Schirmer, Julia Situ, et al. 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities—a global review. *PLOS Digital Health*, 1(3):e0000022.

Isaac S Chan and Geoffrey S Ginsburg. 2011. Personalized medicine: progress and promise. *Annual review of genomics and human genetics*, 12:217–244.

Bo Cowgill, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. 2020. Biased programmers? or biased data? a field experiment in operationalizing ai ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, page 679–681, New York, NY, USA. Association for Computing Machinery.

Kate Crawford, Ryan Calo, and Fred Turner. 2016. Fairness and abstraction in sociotechnical systems. *Journal of Adrenaline*, 8(1):78–94.

Lubna Daraz, Bebe Chang, and Sheila Bouseh. 2022. Inferior: The challenges of gender parity in the artificial intelligence ecosystem-a case for canada. *Frontiers in Artificial Intelligence*, 5:931182.

Jianyuan Deng, Zhibo Yang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. 2022. Artificial intelligence in drug discovery: applications and techniques. *Briefings in Bioinformatics*, 23(1):bbab430.

Nicholas Diakopoulos. 2018. Algorithmic accountability: A primer. *Journal of Adrenaline*, 10(3):298–316.

Nicholas Diakopoulos. 2021. Predictive modeling and fairness: Interventions on the ground. *Journal of Adrenaline*, 11(2):189–208.

Nicholas Diakopoulos. 2022. Towards algorithmic accountability: A systematic literature review. *Journal of Adrenaline*, 12(4):567–589.

Mindy Nunez Duffourc and Sara Gerke. 2023. The proposed eu directives for ai liability leave worrying gaps likely to impact medical ai. *npj Digital Medicine*, 6(1):77.

Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc., USA.

John R Feiner, John W Severinghaus, and Philip E Bickler. 2007. Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender. *Anesthesia & Analgesia*, 105(6):S18–S23.

Emilio Ferrara. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies.

Stefan Feuerriegel, Mateusz Dolata, and Gerhard Schwabe. 2020. Fair ai: Challenges and opportunities. *Business Information Systems Engineering*, 62.

Shaswath Ganapathi, Joanne Palmer, Joseph Alderman, Melanie Calvert, Cyrus Espinoza, Jacqui Gath, Marzyeh Ghassemi, Katherine Heller, Francis Mckay, Alan Karthikesalingam, Stephanie Kuku, Maxine Mackintosh, Sinduja Manohar, Bilal Mateen, Rubeta Matin, Melissa McCradden, Lauren Oakden-Rayner, Johan Ordish, Russell Pearson, and Xiaoxuan Liu. 2022. Tackling bias in ai health datasets through the standing together initiative. *Nature Medicine*, 28:1–2.

Lamija Hafizović, Aldijana Čaušević, Amar Deumić, Lemana Spahić Bećirović, Lejla Gurbeta Pokvić, and Almir Badnjević. 2021. The use of artificial intelligence in diagnostic medical imaging: Systematic literature review. In *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 1–6. IEEE.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2022. Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11945–11953.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Mumtaz Karatas, Levent Eriskin, Muhammet Deveci, Dragan Pamucar, and Harish Garg. 2022. Big data for healthcare industry 4.0: Applications, challenges and future perspectives. *Expert Systems with Applications*, 200:116912.

Amit Kaushal, Russ Altman, and Curt Langlotz. 2020. Health care ai systems are biased. *Scientific American*, 11:17.

Maria Kyrarini, Fotios Lygerakis, Akilesh Rajavenkatanarayanan, Christos Sevastopoulos, Harish Ram Nambiappan, Kodur Krishna Chaitanya, Ashwin Ramesh Babu, Joanne Mathew, and Fillia Makedon. 2021. A survey of robots in healthcare. *Technologies*, 9(1):8.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594.

Yuehua Li, Kai Shang, Wei Bian, Li He, Ying Fan, Tao Ren, and Jiayin Zhang. 2020. Prediction of disease progression in patients with covid-19 by artificial intelligence assisted lesion quantification. *Scientific Reports*, 10(1):22083.

Saskia Locke, Anthony Bashall, Sarah Al-Adely, John Moore, Anthony Wilson, and Gareth B Kitchen. 2021. Natural language processing in medicine: a review. *Trends in Anaesthesia and Critical Care*, 38:4–9.

Farah Magrabi, Elske Ammenwerth, Jytte Brender McNair, Nicolet F De Keizer, Hannele Hyppönen, Pirkko Nykänen, Michael Rigby, Philip J Scott, Tuulikki Vehko, Zoie Shui-Yee Wong, et al. 2019. Artificial intelligence in clinical decision support: challenges for evaluating ai and practical implications. *Yearbook of medical informatics*, 28(01):128–134.

G. Maliha, S. Gerke, I. G. Cohen, and R. B. Parikh. 2021. Artificial intelligence and liability in medicine: Balancing safety and innovation. *Milbank Q*, 99(3):629–647.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Jordan P Richardson, Cambray Smith, Susan Curtis, Sara Watson, Xuan Zhu, Barbara Barry, and Richard R Sharp. 2021. Patient apprehensions about the use of artificial intelligence in healthcare. *NPJ digital medicine*, 4(1):140.

Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. 2022. Explainable ai for healthcare 5.0: opportunities and challenges. *IEEE Access*.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Journal of Adrenaline*, 5(2):210–230.

Najma Taimoor and Semeen Rehman. 2021. Reliable and resilient ai and iot-based personalised healthcare services: A survey. *IEEE Access*, 10:535–563.

The STANDING Together collaboration. 2023. Recommendations for diversity, inclusivity, and generalisability in artificial intelligence health technologies and health datasets.

Lav R. Varshney, Gaurav Thakur, Niyati Pathak, and Ibrahim A. Alabdulmohsin. 2022. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Journal of Adrenaline*, 12(2):178–197.

Smrithi Vijayakumar, V Vien Lee, Qiao Ying Leong, Soo Jung Hong, Agata Blasiak, and Dean Ho. 2023. Physicians' perspectives on ai in clinical decision support systems: Interview study of the curate. ai personalized dose optimization platform. *JMIR Human Factors*, 10:e48476.

Henrik Vogt and Sara Green. 2020. Personalised medicine: Problems of translation into the human domain. *De-Sequencing: Identity Work with Genes*, pages 19–48.

Kerstin Vokinger, Stefan Feuerriegel, and Aaron Kesselheim. 2021. Mitigating bias in machine learning for medicine. *Communications Medicine*, 1.

Clarice Wang, Kathryn Wang, Andrew Y. Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. 2023. When biased humans meet debiased ai: A case study in college major recommendation. *ACM Trans. Interact. Intell. Syst.*, 13(3).

Shen Yan, Di Huang, and Mohammad Soleymani. 2020. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 361–369, New York, NY, USA. Association for Computing Machinery.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment amp; disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17. International World Wide Web Conferences Steering Committee.