Journal of Research High School

Vol. 2024 (3) 126 – 134
ISSN 2688-3651 [online]

Analyzing Ethical Guidelines for Reducing Racial Bias in Medical Machine Learning in Artificial Intelligence

Soha Budhani[1] *

[1]Pinewood School, Los Altos, CA, USA
*Corresponding Author: soha.budhani@gmail.com

Advisor: Dr. Rupinder Gill, robbie@rafay.co

Received December 24, 2023; Revised February 4, 2024; Accepted, February 19, 2024

Abstract

Over the past seven decades, artificial intelligence (AI), specifically AI based on machine learning (AI/ML), has evolved from a mere concept to a ubiquitous force permeating nearly every facet of modern society. Today, many medical devices leverage AI/ML-based technologies to recommend actions and treatments. Because AI/ML technologies inadvertently inherit discriminatory practices from their creators and from the data used to "train" them, bias has been found in a number of medical systems that rely on AI/ML technologies. This bias has directly resulted in the misalignment of treatment offered to people of different races for the same ailment, resulting in the exacerbation of existing inequalities between racial groups and the further marginalization of vulnerable communities including but not limited to women, low-income families, or persons identifying as LGBTQ+. This paper advocates for a comprehensive set of ethical guidelines for development and use of AI/ML-based systems in the medical space to alleviate the presence of racial bias. This paper further presents a systematic review of prevailing ethical frameworks aimed at mitigating racial bias in medical AI/ML domains while proposing avenues for continuous improvement in ethical practices.

Keywords: Artificial Intelligence, Medicine, Racial Bias, Machine Learning, Healthcare

1. Introduction

"Computing Machinery and Intelligence" is widely regarded as the seminal paper on AI/ML and has remained the bedrock for the discipline's broad applications. Authored by Alan Turing in 1950, the paper ushered concepts and mental frameworks about AI/ML that continue to be employed today. One notable topic of discussion in Turing's paper is the question of whether machines can "think" (Turing, 1950). The obvious conclusion is reached: a machine is not a living being, does not possess a brain, and therefore does not "think," at least not like a human does. Instead, it imitates, or tries to imitate as best it can, the manner in which humans think, termed the "imitation game" (Turing, 1950). Since the conceptualization of this "imitation game," AI/ML technologies have been created to imitate human behavior and, in many cases, surpass it.

Contemporary AI/ML technologies can efficiently initiate an almost infinite number of arduous, demonstrating vast and expeditious growth since Christopher Strachey's success in building the first working AI program in 1951: a program capable of playing a game of checkers (Bleakley, 2020; Martinez-Millana et al., 2022). In the 1990's, just forty years after Turing's paper was published, an IBM computer defeated a chess grandmaster in thirty-seven moves by implementing the concepts in "Computing Machinery and Intelligence" (Weber, 1996). Today machines adeptly play the "imitation game," emulating human skill with increasing efficiency.

AI/ML technologies mimic human behavior however a notable deficiency lies in the absence of explicit regulations to prevent imitations of, or assimilating, aberrant human behaviors, e.g. racial bias (Noseworthy et al., 2020). There appears to be a direct causality between the behavior of AI/ML technologies and the training data; the material it is exposed to and the integrity of their developers. Consequently, training AI models on the breadth of often unreliable material available on the Internet may lead to inadvertent embedding of biases and stereotypes against marginalized peoples. Recently, ChatGPT has emerged as a tremendously advanced large language model (LLM) based AI tool trained on a large portion of the Internet (De Angelis et al., 2023). An LLM made to facilitate almost any action possible involving textual input on a screen, the technology is one of the first of its kind with the capabilities of mimicking human text based on the information it is given, "answering" almost any human question. Despite these merits, ChatGPT outputs racist responses when asked about certain subjects, including its opinions on famous people of color such as African American celebrity boxer Muhammad Ali (Deshpande et al., 2023).

In the field of medicine, where AI/ML often assists with surgeries, interprets imaging, and predicts healthcare needs, enforcing ethical guidelines is critical to reduce racial biases and ensure fair treatment. Many examples of racial bias have been reported in medical systems, and others may exist that have gone unchecked. With the use of AI likely to accelerate with the rise of technologies such as Generative AI, the medical and AI communities must independently and jointly explore avenues to reduce racial bias in the foundational models used by the medical industry and in the medical systems built based on these models.

The objective of this paper is to set forth that medical AI/ML technologies must be given robust ethical guidelines to abide by in order to execute their duties effectively without developing and perpetuating existing racial biases and further embedding systematic racism into the ever-evolving field of AI/ML. Moreover, this paper hypothesizes that because we are aware of these implicit and explicit biases, it is essential to proactively establish such ethical guidelines, also known as guardrails, to rectify present harm and prevent future harm. By carrying out tasks that today require humans to complete, AI/ML-based technologies can accelerate processes and deliver services to orders-of-magnitude larger populations. This type of reach can truly bring life-saving diagnoses and treatments to patients everywhere. Nevertheless, without actively establishing and enforcing guardrails to reduce and ideally remove bias from these systems, medical treatment could vary by ethnicity, race, etc. over time which would be a huge step back in our collective march towards racial equality.

2. Materials and Methods

A systematic review conducted using PubMed, JSTOR, and Google Scholar identified papers that pertained to the subjects of medicine, law/ethics, general biases, artificial intelligence, machine learning (ML), racial bias, and gender bias. PRISMA guidelines were referenced. Keywords and phrases searched for included "artificial intelligence," "medicine," "healthcare," "bias," "racial bias," "racism," "machine learning," "engineering," "law," "medical AI," "gender bias," "law," "reinforcement learning," "unsupervised learning," "supervised learning," and "medicine in AI."

The search was confined to papers published from 2017 to 2023 and two hundred and fifty-two papers met the inclusion criteria. This paper primarily focuses on AI/ML technologies based on reinforcement learning (both unsupervised and supervised types) methodology, therefore papers that did not directly or indirectly cover reinforcement learning were discarded. Subsequently, papers that did not directly discuss racial bias in medical settings were discarded. Ultimately sixty-seven quality papers were incorporated to effectively address ethical concerns in medical AI/ML and derive relevant ethical guidelines based on this domain.

3. Representative Case Studies

Bias in medicine related to AI/ML models and algorithms have been recorded in multiple publications. These studies reinforce the need for awareness of racial disparities in the healthcare system. One instance of racial bias in medical algorithms was identified in the area of obstetrics, where vaginal birth after cesarean delivery procedures were consistently not being recommended to women of Black/African American and Hispanic/Latino descent but were being recommended for non-Hispanic White women (Vyas et al., 2019). This led to significantly more cesarean deliveries as opposed to vaginal deliveries for Black/African American and Hispanic/Latino women. Another instance occurred in 2019, when an algorithm used to decide which patients needed care contained implicit racial biases

(Obermeyer et al.) with respect to pain threshold. The algorithm found that black patients must present with a greater severity of symptoms than their white counterparts to receive a similar degree of care.

4. Results

In the ever-evolving landscape of AI/ML, the implementation of a comprehensive set of ethical guidelines is more important now than ever before, particularly due to the unconscious presence of racial bias in AI/ML technologies. Ethical guidelines and principles can provide a framework to prioritize inclusivity, transparency, and accountability: key ethical values (Char et al., 2018). After the study of numerous relevant scholarly papers, five distinct ethical guidelines for AI/ML in healthcare have been formulated.

4.1 Guideline 1: Ensuring Healthcare Providers Recognize Racial Influence in Healthcare Technologies

Racism remains a complex issue in medical AI/ML technologies. A patient, subject to their race, may have certain genetic predispositions, be prone to particular diseases, or face unique challenges (Jindal, 2023). Understanding and addressing these disparities is crucial for improving equity in medical care.

Compulsory training for healthcare providers on the impact of systemic racism and its relevance in medical AI/ML software would mirror existing mandates for sexual harassment awareness, as an example. Racism "stemm[ing] from misperceptions and stereotypes" profoundly impacts both the treatment of patients of color as well as providers of color and the overall healthcare services rendered to these groups (Smedley et al., 2001). Educating healthcare workers on racism embedded in their own field sheds light on unconscious biases, fostering improved patient outcomes and ultimately decreasing both insurance costs and public health costs. This solution is cost-effective and actionable because similar programs to combat other issues in the workplace already exist.

4.2 Guideline 2: Integrating Ethics Classes into Computer Science Curricula

The integration of ethics classes in Computer Science (CS) curricula is becoming increasingly important. Such coursework fosters an understanding of the societal impact and responsibility intrinsic to the design and application of AI/ML technologies (Karoff, 2019). Harvard University's proactive implementation of classes on the intersection of ethics and AI is reflective of a growing trend amongst educational institutions (Karoff, 2019). More than simply acquiring knowledge, it is imperative for students to cultivate an ethical mindset that guides their conduct in this area (W3C, 2022). The importance here is not the degree of comprehension but a student's awareness of the need to integrate ethics into CS. As AI/ML technologies improve in ability so too will their potential for harm (Manyika, 2022). It is essential to teach future generations about these issues to ensure they possess the knowledge required to navigate and mitigate them effectively. By integrating ethics classes into CS curricula, students will be better prepared to develop technologies in a way that promotes transparency, accountability, and equity (Karoff, 2019; Lo Piano, 2020).

4.3 Guideline 3: Establishing Ethical Values for Medical AI/ML Technologies

The World Health Organization (WHO) has identified six fundamental principles for responsible deployment of AI/ML in healthcare: "(1) protect autonomy; (2) promote human well-being, human safety, and the public interest; (3) ensure transparency, explainability, and intelligibility; (4) foster responsibility and accountability; (5) ensure inclusiveness and equity; (6) promote AI that is responsive and sustainable" (WHO, 2023). These principles should serve as foundational ethical values adhered to by creators and maintainers of AI/ML-based healthcare technologies. Further, it is important for these creators and maintainers to recognize the relevance of ethics in their work and delineate a concise set of ethical principles and values that their healthcare technologies should embody (Corro, 2022). Ethical values as a foundation for developing AI/ML technologies can mitigate biases and discrimination, promote inclusivity and accessibility, and enhance trust and acceptance in AI/ML systems. Publishing the ethical values codified for, and coded into, each AI/ML-based healthcare technology can further enhance trust.

4.4 Guideline 4: Evaluating AI/ML Technologies in Medicine and the Use of HITL Approaches

The integration of ethical values in AI/ML technologies serves as a positive step toward rectifying racial disparities in the medical field. Nevertheless, there remains the risk that these technologies have and may continue to unintentionally perpetuate racial biases manifesting as discriminatory practices (Shanklin et al., 2022). Given the impossibility of predicting all scenarios of bias and its impact, strict oversight is necessary. Thus, maintaining vigilant oversight and regular evaluations of AI systems is crucial to prevent and rectify emergent biases (Silberg et al., 2019).

Neglecting to employ ethical principles increases the likelihood for dissemination of inaccurate information. Ethical guardrails are essential across disciplines: an illustrative example of how technology can help society or cause it harm is the "violence risk assessment" framework that many law enforcement agencies use to determine a criminal's risk factor (Chohlas-Wood, 2020). These predictions inform high-stakes judicial decisions, such as whether to incarcerate an individual before their trial (Chohlas-Wood, 2020). Without a framework to ensure no ethnicity or race is adversely impacted by such an assessment, all such frameworks are suspect (Shanklin et al., 2022).

Amidst these evaluations risk assessment instruments (RAIs) have been designed to predict the probability of a convict's recidivism (Cholas-Wood, 2020). These assessment tools consider a multitude of factors, including age and prior misconduct to generate distinct risk scores. These scores then inform a decision-making framework that creates recommendations for release conditions, where higher risk scores correspond to more stringent release conditions. It is important to note that judges have discretion to deviate from these recommendations if they perceive them to be excessively strict or lenient (Chohlas-Wood, 2020). Nevertheless, these risk assessments are imperfect because they can inadvertently incorporate racial biases to draw conclusions rather than focusing solely on the defendant's actual crimes. To address this issue, it is logical to develop a "racial bias risk assessment" specifically for RAIs, which can be applied by humans to evaluate RAI frameworks. By conducting routine assessments of this nature, it becomes possible to identify and rectify racial biases consistently, thereby mitigating a significant portion of racially-driven decisions from influencing outcomes.

This idea is present in the Human in the Loop (HITL) concept applied to AI/ML technologies, an approach that "places human beings in a supervisory role and is more relevant for healthcare purposes" (ICMR, 2023). The idea is simple: All new instructions for AI/ML-based models should be reviewed and adjusted by humans during an algorithm's training and evaluation stages to ensure that the model is not learning things it should not (Wu et al., 2022). With HITL incorporation, AI/ML can be particularly useful in medical technologies because it enables ongoing collaboration between humans and machines (Patel et al., 2019). To illustrate how HITL can address the bias issue, consider that an AI/ML ethicist could pose theoretical situations to test a medical system. The system in question could be asked to evaluate a diverse group of patients with a variety of symptoms and determine what treatments are most appropriate for each patient. When the same question is posited with patients of a different background, ethnicity and/or race, inherent biases will be exposed. The existence of racial bias risk assessments for AI/ML addresses several concerns but would necessitate consistent oversight over AI/ML applications. Implementing HITL in medical AI/ML technologies can be a valuable tool, allowing for continuous interaction between humans and machines and ensuring greater accuracy in diagnosis as well as treatment options (Patel et al., 2019).

4.5 Guideline 5: Open-Sourcing Models and Code Used in AI/ML Healthcare Technologies

Although medical AI/ML technologies undergo periodic evaluations to check for racial bias, it remains essential for patients and users to have transparent access to the datasets, information, code, and models behind the technologies (El-Kassas). One of the most crucial ethical guardrails for AI/ML is transparency. Any code that could potentially determine outcomes based on race should be openly accessible, allowing visibility into how the code was written and its functionality (Raths, 2022). Ideally, this would also include open-sourcing the data used to train these models. This guideline promotes transparency and accountability in AI/ML technologies, enabling users and third parties to audit and inspect the systems for security, privacy, bias, fairness, and other ethical concerns.

Open-source AI/ML code promotes transparency and accountability; popular code repositories are peer-reviewed by researchers and analysts, ideally resulting in a codebase presenting fewer ethical challenges (Felzmann et al., 2020). This approach will lead to the identification of potential ethical concerns, the development of more inclusive and representative technologies, and the creation of solutions that better serve the needs of diverse groups. Moreover, developers are more likely to adhere to ethical guidelines and standards as their work is subject to public scrutiny. For instance, in a situation where a healthcare AI/ML technology is used to assess eligibility for medical care, a decision with profound consequences on a patient's life, open-sourcing all underlying code safeguards patients from bias while ensuring these innovations do not cause inadvertent harm to patients seeking care. Inviting community and industry feedback will further allow for contributions and scrutiny from diverse perspectives.

4.6 Limitations

In analyzing these ethical guidelines, it is critical to note that definitively eliminating racism in medical AI/ML is an unattainable goal, though it can be mitigated as much as possible. Adequate studies have yet to be conducted to understand if and how racism can be effectively eradicated. AI/ML technologies are complex technologies which transcend mere ethical guidelines and cannot be effortlessly "fixed." Employing ethical guardrails serves as a viable approach to tackling this issue (Peng et al., 2010).

5. Discussion

The intersection of ethics and medicine is particularly significant, as AI/ML technologies are being utilized more extensively in healthcare. There should be no intentional or unintentional logic in the code that inherently maligns a particular race, for example, when making recommendations related to patient care. Unchecked racial bias in such technologies can lead to serious harm for the patient and even death. To properly create ethical guardrails for racial bias, it is essential to understand how bias, specifically racial bias, is processed in the brain. By understanding this process, one can understand how and why AI/ML technologies can exacerbate human biases, and how to implement ethical guidelines to address the root cause of the issue of perpetuated racial biases in AI/ML technologies.

Human behavior is inherently subject to biases, and AI/ML technologies inherit these biases from their creators. These inherited biases become further exacerbated if the data being used to train the underlying model(s) is not selected with care. Most medical AI/ML technologies in use today lack programmed ethical guidelines, which results in there being no means to regulate the biases and "opinions" that could influence the technologies' performance and "decisions." Given the direct connection between how the human brain works and how AI/ML systems are essentially designed to imitate human behavior, it is critical to understand neuroscience and the human brain in creating ethical guidelines medical AI/ML technologies that may result in a reduction of racial bias in these technologies. Ethical guidelines can serve as objective mirrors for parts of the brain, including the prefrontal cortex and hippocampus, acting as a guardrail against embedded racial bias within AI/ML.

Ethical guidelines are not limited to implementation in AI/ML technologies. If healthcare providers and medical professionals are unaware of the dangers of bias in medical technologies and the implications in the field, it will be impossible for them to take steps to respond to this issue. One way to educate healthcare providers would be to add a requisite provider course on racism in the healthcare system and its relevance to AI/ML to mandated training. Similar classes have been successfully implemented in other areas, such as sexual harassment prevention, indicating that provider training may serve as an effective means of raising these issues to healthcare providers. By understanding what to look for to prevent the presence of racial biases in medical AI/ML, hospitals may choose to implement more specific ethical guidelines relevant to their technologies. To ensure that a priori overworked healthcare professionals will take these classes and workshops seriously, it makes sense to include real-world case studies into the curriculum such as the ones discussed earlier in this paper, with the goal of ensuring that these professionals internalize the role of racial bias in the technologies they use on a daily basis.

By integrating ethics classes into CS curricula in schools and universities, students will also be exposed to the dangers of biases in CS, and the importance of combating this issue (Karoff, 2019). Though a course focused on racial bias would more than likely cover general information, it is essential for students to be aware of this issue as AI/ML technologies will only continue to improve and subsequent issues will arise as the technology becomes increasingly ubiquitous. Ethical guidelines are an essential aspect of combatting biases in medical AI/M and this problem must be understood by students from a young age (Karoff, 2019). Many business schools globally added ethics to their curriculum for future bankers and businesspeople after the global economic collapse of 2008 though the effect of this action remains unclear (Rasche et al., 2013). Similarly, exposing the next generation developers of AI/ML-based technologies to potential biases that can arise in their future creations and the ethical standards they should follow is a worthy exercise.

Another important ethical guideline is the act of establishing specific values for different medical AI/ML technologies. Depending on the action the technology is implementing, it may be able to learn biases in different ways. By integrating ethical values into the creation of these technologies, creators can combat learned racial bias. Values such as integrity and transparency are critical and must be front of mind for those who create these technologies that permeate and impact the masses, emphasizing the importance of the WHO's core principles. If we are to create technology that mimics the behavior of the human brain, we must allow it to mimic our positive qualities to the best of its ability as a technology as well. Further research is necessary on how to establish such values into the CS curricula or weed out those persons who may transgress or bring their inherent problematic worldviews to the table. This is not likely a solvable problem but is something we must be aware of and ready to counteract as evidence of such arises.

It is also essential to continuously assess medical AI/ML technologies to check for acquired racial biases. Through continuous assessments, hospitals can mitigate or prevent the usage of such technologies when necessary to correct the presence of biases and take steps to understand how they occurred and how to prevent similar occurrences in the future. This can be accomplished through HITL AI/ML, which would involve human and technology collaboration throughout all stages of the creation and implementation process (ICMR, 2023). An approach like this one can work to significantly reduce the biases learned by medical AI/ML, positively affecting the treatment received by patients of the technology.

Open-sourcing the models, technologies, code, and other information used in medical AI/ML technologies promotes transparency between the patients and healthcare providers. Doing so also pressures creators of medical technologies to exercise caution with respect to the information they allow their technologies to consume, and increases scrutiny of those responsible for identifying and mitigating racial biases. Additionally, it is possible for others to find biases or issues in the open-source information and help to improve these technologies for the better. Through the implementation of these ethical guidelines, medical AI/ML technologies can be made safer for all and improve patient outcomes.

As we put forth these recommendations, we recognize that popular opinion is aligned with the wholesale adoption of AI/ML-based technologies in medical systems. With claims being made about AI-based systems having better bedside manner compared to humans, the adoption of AI/ML-based technologies will likely only accelerate (Lenharo, 2024). As a result, it is even more important for patients who may be directly impacted by bias in these systems to be hyperaware. Identifying racial bias is more difficult when perpetuated by a machine versus a human. Machines have no race or ethnicity, and patients may not realize how their creators' biases or choice of training data may impact their lives. It makes sense for patients to inquire whether any care-related decision was made by a human or a system, and if it is the latter, request a second opinion from a human until further research is carried out in this space. At a minimum, asking for second opinions will put pressure on the medical community to take the issue of unchecked biases in medical systems more seriously.

6. Conclusion

Simply turning a blind eye to the impact of race in healthcare will not resolve the issue of racial bias in AI/ML-based medical systems. The remarkable growth of AI/ML technologies is evident in their ability to, in many instances, outperform human capabilities. The performance of AI/ML technologies is contingent on the ethical guidelines woven into their code and development; a lack thereof can have detrimental consequences on the populace. Medical applications of AI/ML demand a set of values to abide by, ethical guidelines, to ensure equitable treatment for patients of all backgrounds (Busuioc, 2020). Though there lacks a single solution to the issue of racism in these technologies, ethical guidelines can play a crucial role in reducing racism in AI/ML technologies. Implementing broad, values-based guardrails in the medical field will establish a standard for AI/ML technologies' safety, efficacy, and inclusivity. The diverse subjects of this paper allow for an interdisciplinary approach to this issue, facilitating the development of a truly comprehensive set of ethical guidelines for reducing racial bias in medical AI/ML technologies. Ethical guidelines can help elevate the standards of safety of these technologies. In the future, it is imperative that professionals consider the implementation of ethical guidelines in medicine on a broad scale. However, this is not an individual effort; healthcare professionals, engineers, and policymakers will need to work collaboratively to ensure that overarching goals of equity and justice are aligned. This collective effort will amplify social trust for AI/ML technologies, paving the way for increased and effective application in healthcare domains.

References

Angwin, J., et al. (2016, May 23). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bleakley, C. (2020). *Poems That Solve Puzzles: The history and science of algorithms*. Oxford University Press. https://doi.org/10.1093/oso/9780198853732.002.0004

Busuioc, M. (2020). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, *81*(5), 825-836. https://doi.org/10.1111/puar.13293

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *The New England Journal of Medicine*, *378*(11), 981. https://doi.org/10.1056/NEJMp1714229

Chohlas-Wood, A. (2020, June 19). Understanding risk assessment instruments in criminal justice. *The Brookings Institution*. https://www.brookings.edu/articles/understanding-risk-assessment-instruments-in-criminal-justice/

Corro, E. (2022). *Why your organization needs a set of ethical principles for AI.* Office of the CTO Blog. https://octo.vmware.com/why-your-organization-needs-ethical-principles-for-ai/

De Angelis, L., et al. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*, 1166120. https://doi.org/10.3389/fpubh.2023.1166120

Deshpande, A., et al. (2023). *Toxicity in chatgpt: Analyzing persona-assigned language models*. https://doi.org/10.48550/arXiv.2304.05335

El-Kassas, S. (n.d.). *On the merits of the open source model.* https://www.wipo.int/edocs/mdocs/mdocs/en/isipd_05/isipd_05_www_103981.pdf

Ethical guidelines for application of artificial intelligence in biomedical research and heathcare. ICMR. (2023). https://main.icmr.nic.in/sites/default/files/upload_documents/Ethical_Guidelines_AI_Healthcare_2023.pdf

Ethical principles for web machine learning. W3C. (2022, November 29). https://www.w3.org/TR/webmachinelearning-ethics/

Felzmann, H., et al. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333-3361. https://doi.org/10.1007/s11948-020-00276-4

Jindal, A. (2022). Misguided artificial intelligence: How racial bias is built into clinical models. *Brown Hospital Medicine*, *2*(1). https://doi.org/10.56305/001c.38021

Karoff, P. (2019, January 28). Harvard works to embed ethics in Computer Science curriculum. *Harvard Gazette*. https://news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/

Lenharo, M. (2024). Google AI has better bedside manner than human doctors—and makes better diagnoses. *Nature*, *625*(7996), 643-644. https://www.nature.com/articles/d41586-024-00099-4

Lo Piano, S. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, *7*(1), 1-7. doi:10.1057/s41599-020-0501-9

Manyika, J. (2022, May 1). Getting AI right: Introductory notes on AI & society. *Daedalus* (2022) 151 (2): 5–27. https://doi.org/10.1162/daed_e_01897

Martinez-Millana, A., et al. (2022). Artificial intelligence and its impact on the domains of universal health coverage, health emergencies and health promotion: An overview of systematic reviews. *International Journal of Medical Informatics*, *166*, 104855. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9551134/

Noseworthy, P. A., et al. (2020). Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. *Circulation: Arrhythmia and Electrophysiology*. https://doi.org/10.1161/CIRCEP.119.007988

Obermeyer, Z., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453. doi: 10.1126/science.aax2342

Patel, B., et al. (2019). Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine*, *2*(1), 111. https://doi.org/10.1038/s41746-019-0189-7

Peng, Y., Zhang, Y., & Wang, L. (2010). Artificial intelligence in biomedical engineering and informatics: an introduction and review. *Artificial Intelligence in Medicine*, *48*(2-3), 71–73. https://doi.org/10.1016/j.artmed.2009.07.007

Rasche, A., Gilbert, D. U., & Schedel, I. (2013). Cross-Disciplinary Ethics Education in MBA Programs: Rhetoric or Reality? *Academy of Management Learning & Education*, *12*(1), 71–85. http://www.jstor.org/stable/23412393

Raths, D. (2022, October 7). Open-source network for AI gaining momentum. *Healthcare Innovation*. https://www.hcinnovationgroup.com/imaging/artificial-intelligence/article/21283155/open-source-network-for-ai-gaining-momentum

Shanklin, R., et al. (2022). Ethical redress of racial inequities in AI: Lessons from decoupling Machine Learning from optimization in medical appointment scheduling. *Philosophy & Technology*, *35*(4), 96. https://doi.org/10.1007/s13347-022-00590-8

Silberg, J. & Manyika, J. (2019). Tackling bias in AI (and in humans). *McKinsey Global Institute*, *1*(6). https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

Smedley, B. D., Stith, A. Y., & Nelson, A. R. (2003). Racial disparities in health care: highlights from focus group findings. *Unequal treatment: Confronting racial and ethnic disparities in health care*. National Academies Press (US). doi: 10.17226/12875

Turing, A.M. (1950). Computing machinery and intelligence. *Mind* (59), 433-560. https://redirect.cs.umbc.edu/courses/471/papers/turing.

Vyas, D. A., et al. (2019). Challenging the use of race in the vaginal birth after cesarean section calculator. *Women's Health Issues: Official publication of the Jacobs Institute of Women's Health*, *29*(3), 201–204. https://doi.org/10.1016/j.whi.2019.04.007

Weber, B. (1996, February 11). In Upset, Computer Beats Chess Champion. *The New York Times*. https://www.nytimes.com/1996/02/11/us/in-upset-computer-beats-chess-champion.html

World Health Organization. (2023). *WHO calls for safe and ethical AI for health*. https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health

Wu, X., et al. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364-381. https://doi.org/10.1016/j.future.2022.05.014