Journal of Health Ethics

Volume 20 | Issue 1 Article 4

10-2024

Ethical Requirements for Achieving Fairness in Radiology Machine Learning: An Intersectionality and Social Embeddedness Approach

Dessislava S. Fessenko
Harvard Medical School, dessislava_fessenko@hms.harvard.edu

Recommended Citation

Fessenko, D. S. (2024). Ethical Requirements for Achieving Fairness in Radiology Machine Learning: An Intersectionality and Social Embeddedness Approach. *Journal of Health Ethics, 20*(1). http://dx.doi.org/10.18785/jhe.2001.04

Ethical Requirements for Achieving Fairness in Radiology Machine Learning: An Intersectionality and Social Embeddedness Approach

Cover Page Footnote

I gratefully acknowledge Prof. Vardit Ravitsky, President & CEO of The Hastings Center, Dr. Elizabeth Nilson, Director, Medical Ethics, at Lahey Hospital, and Dr. Judy Gichoya, associate professor at Emory University School of Medicine, for their feedback to and guidance on my capstone work on which this publication is based.

Ethical Requirements for Achieving Fairness in Radiology Machine Learning: An Intersectionality and Social Embeddedness Approach

Dessislava S. Fessenko
Harvard Medical School

ABSTRACT

Radiodiagnostics by machine-learning (ML) systems is often perceived as objective and fair. It may, however, exhibit bias towards certain patient sub-groups. The typical reasons for this are the selection of disease features for ML systems to screen, that ML systems learn from human clinical judgements, which are often biased, and that fairness in ML is often inappropriately conceptualized as "equality". ML systems with such parameters fail to accurately diagnose and address patients' actual health needs and how they depend on patients' social identities (i.e. intersectionality) and broader social conditions (i.e. embeddedness). This paper explores the ethical obligations to ensure fairness of ML systems precisely in light of patients' intersectionality and the social embeddedness of their health. The paper proposes a set of interventions to tackle these issues. It recommended a paradigm shift in the development of ML systems that enables them to screen both endogenous disease causes and the health effects of patients' relevant underlying (e.g. socioeconomic) circumstances. The paper proposes a framework of ethical requirements for instituting this shift and further ensuring fairness. The requirements center patients' intersectionality and the social embeddedness of their health most notably through (i) integrating in ML systems adequate measurable medical indicators of the health impact of patients' circumstances, (ii) ethically sourced, diverse, representative and correct patient data concerning relevant disease features and medical indicators, and (iii) iterative socially sensitive co-exploration and co-design of datasets and ML systems involving all relevant stakeholders.

Keywords: Machine Learning; Fairness; Justice; Ethical Requirements; Radiology

INTRODUCTION

Machine learning (ML) systems find an ever greater application in clinical care settings for healthcare monitoring, diagnostics and risk management (Bates & Zimlichman, 2015; Chen et al., 2024; Obermeyer et al., 2019). In radiology specifically, ML systems are used to assist with or augment clinicians' work in a variety of image acquisition, analysis, interpretation, diagnostics and decision support tasks (Hanneman et al., 2024; Yu et al., 2024). A key driver of this wider adoption appears to be the expectation -- or at least the perception—that ML systems could outperform clinicians in image interpretation and precisions (Pot et al., 2021; Satariano et al., 2023; Yu et al., 2024). The so-called "automation bias"—humans' proclivity to "overestimate the validity and the predictive power of the information produced by an automated system" (Pot et al., 2021, p. 7)—has endowed ML systems with an aura of objectivity and fairness (Pot et al., 2021). However, this aura may not be entirely justified. A major challenge to the fulfillment of these high hopes has proven to be the prevalence in ML systems of bias and the associated unfair outcomes (Gichoya et al., 2022, 2023; Hanneman et al., 2024; Pot et al., 2021). Empirical research has systematically documented the persistence in ML systems of bias towards and unfair treatment of historically underserved patient populations, including in radiology (Obermeyer et al., 2019; Gichoya et al., 2023; Seyyed-Kalantari et al., 2020, 2021; Pierson et al., 2021; Beheshtian et al., 2023; Bernhardt et al., 2022; Mukherjee et al., 2022). The inadequate healthcare access and treatment that ML bias specifically leads to has raised ethical concerns about the overall unfairness induced or perpetuated by ML systems.

This background triggers the question of how fairness in ML could be achieved. This paper analyzes the problem in further detail and proposes possible solutions. To this end, three sets of issues have been examined. The first set is technical and concerns the types of standard parameters of ML

models that habitually yield bias and unfairness. The second set of questions is conceptual and seeks to reconcile the essence and inherent purpose of fairness under the dominant philosophical theories in order to illuminate possible shared meaning and objectives that could serve as a common theoretical ground for the design of fairness in ML. The third set of questions relates to the specific ethical requirements that healthcare providers and ML developers should meet in order to realize the essence and objective of fairness when building or using their ML systems.

RESEARCH METHODOLOGY

Systematic literature review and analysis were conducted of:

- Relevant publications on the technical set of questions described above (Barocas et al., 2019; Binns, 2020, 2021; Cooper et al., 2021; Rajkomar et al., 2018; Pierson et al., 2021; Gichoya et al., 2023; Holzinger et al., 2019; Pot et al., 2021; Mittelstadt et al., 2023; Mukherjee et al., 2022; Seyyed-Kalantari et al., 2020, 2021; Bernhardt et al., 2022; Beheshtian et al., 2023);
- Relevant monographies and manuscripts regarding the essence and purposes of justice and fairness according to utilitarianism (Mill, 1864, 2008), republicanism and libertarianism (E. Anderson, 2015), liberal egalitarianism (Daniels, 2001; Gomberg, 2010; Hsieh & Department of Philosophy, Florida State University, 2005; Rawls, 1971), narrow, broad and strict conceptions of equality (Arneson, 2018; Barry, 2005; John, 1970; Mittelstadt et al., 2023; Rae, 1981; Roemer, 1995), relational egalitarianism (E. Anderson, 2010, 2012; E. S. Anderson, 1999; Voigt & Wester, 2015), capabilities theories (Powers, 2019; Sen, 1995), non-ideal and relational theories (Baylis et al., 2008; Bennett & Keyes, 2020; Hoffmann, 2019; Llewellyn & Downie, 2012; Sherwin & Feminist Health Care Ethics Research Network, 1998; Young, 2011; Young & Nussbaum, 2011), and communitarianism (MacIntyre, 2013; Sandel, 2010);
- Relevant publications regarding the ethical requirements for ensuring fairness in ML systems (Barocas et al., 2019; Benjamin, 2019; Bennett & Keyes, 2020; Birhane, 2021; Cavaliere et al., 2019; Dignum, 2022; Faden et al., 2013; Farmer, 2004; Gichoya et al., 2023; Hoffmann, 2019; London, 2019, 2022; Malanga et al., 2018; McDougall, 2019; Metcalf & Crawford, 2016; Mühlhoff, 2023; Ravitsky, 2024; Sauer et al., 2022; Sherwin & Feminist Health Care Ethics Research Network, 1998; Voigt, 2019; Voigt & Wester, 2015; Xiang, 2021; Zook et al., 2017).

FINDINGS

Technical Aspects

The relevant technical literature singles out the choice of three key model parameters of ML systems as lead causes for bias and unfairness. The first one is the selection of disease features that radiology ML systems monitor in order to predict outcomes. Typically, ML systems measure standard endogenous disease characteristics, and fail to track external aggravating factors (Cooper et al., 2021, p. 5; Gichoya et al., 2023; Holzinger et al., 2019; Pot et al., 2021, pp. 8–10). For example, standard methods for diagnosing osteoarthritis measure causes within the knee, such as radiological severity or structural damages, but do not track causes external to the knee, such as life stress leading to higher experienced pain (Pierson et al., 2021). The failure of ML systems to track all relevant disease factors, including external ones, results in underdiagnoses, and treatment and overall health disparities (Bernhardt et al., 2022; Pierson et al., 2021; Seyyed-Kalantari et al., 2020, 2021).

The second key model parameter of significance to fairness in radiology ML systems is the type of data that they learn from. This is often data that contains some sort of a clinical judgements, i.e. how a clinician has diagnosed or would diagnose or classify a disease characteristics or a condition (e.g. radiological severity of knee osteoarthritis). The approach of learning from physicians' clinical judgments practically imbues implicit human bias into the data and thus into the assessment of ML systems (Gichoya et al., 2023; Pierson et al., 2021; Pot et al., 2021).

The third key model parameter that determines the fairness of ML systems is the technical notion and the statistical measurements of fairness incorporated into the systems. Recent research has revealed that fairness in ML is often technically conceptualized and statistically measured as equal predictive accuracy across sub-groups of patients (Barocas et al., 2019, pp. 79–101; Binns, 2020, 2021; Cooper et al., 2021, p. 4; Rajkomar et al., 2018; Seyyed-Kalantari et al., 2020, 2021). This means that whether a system is fair is judged by how the prediction rates in one sub-group of patients (e.g. Whites) compares (e.g. equates or not) with the prediction rate in another sub-group (e.g. Black patients). In this way, the widely adopted technical notion and statistical measures of fairness in ML implement a particular conception of the principle of justice. This conception entails strict equality based on some sort of parity (most often statistical or demographic) among subgroups of patients (Binns, 2020, 2021; Cooper et al., 2021; Mittelstadt et al., 2023; Seyyed-Kalantari et al., 2020, 2021). These technical notion and statistical measures have been criticized as inadequate for attaining fairness for two main reasons. First, they fail to adequately account for the structural inhibitors of health (Binns, 2020, 2021; Cooper et al., 2021; Seyyed-Kalantari et al., 2020, 2021). Second, the equality-based fairness notion and statistical measures may result in deterioration of the predictive accuracy of an ML system—and thus underdiagnoses—with regard to certain subgroups of patients (Mittelstadt et al., 2023).

Philosophical Aspects

In light of this criticism, we considered whether another conception of justice might be better suited to reconstruct and ultimately meet the demands of justice in practice. Yet, any conception of justice under the various dominant philosophical theories is considered to have its deficiencies and is hence amenable to objections (Jennings, 2014, p. 1774; Kymlicka, 2001, p. 1). Moreover, the technical approaches for operationalizing these conceptions exhibit various inherent tradeoffs (e.g. in accuracy, comprehensiveness, context-sensitivity) (Binns, 2020, 2021; Birhane, 2021; Cooper et al., 2021; Baumann et al., 2023). To try and overcome these constraints and conceptual tensions, it appears sensible to establish the ultimate essence and inherent purpose of justice under the dominant theories and rely on this core as a common theoretical ground for conceptualizing fairness and tackling all three causes above in ML. We therefore further sought to establish this common ground.

The essence and inherent purpose of justice under the major philosophical theories listed above appear to coalesce into a shared meaning and a common objective of respecting one's inherent moral worth. This entails giving due regard to one's self, needs and interests and enabling one's self-development, realization and thriving (Mill, 1864, pp. 42–56; Rawls, 1971, pp. 54–57, 78–81; E. S. Anderson, 1999, pp. 308–316; E. Anderson, 2012; Sen, 1995, pp. 39–42, 49; Sandel, 2010, pp. 221–242; Kymlicka, 2001, pp. 1–5; Scanlon, 2003; Young, 2011, p. 39; Allen, 2023, pp. 16–56). Such due regard in particular involves structuring and operating social constructs, systems and practices (such as laws, healthcare systems, healthcare delivery) to support human agency and enablement (Allen, 2023, pp. 32–33; Young & Nussbaum, 2011, pp. 43–74). In this way, justice aims to safeguard, uphold and enhance one's dignity, autonomy and wellbeing.

In a healthcare context, respect for patients' inherent moral worth means giving due consideration to, i.e. care for, their health needs in ways that enable patients' functioning and flourishing (Daniels, 2001; Farmer, 2004; Sherwin & Feminist Health Care Ethics Research Network, 1998; Baylis et al., 2008; Llewellyn & Downie, 2012, pp. 63–88; Voigt & Wester, 2015; Powers, 2019, pp. 3–15). To achieve this, healthcare systems and providers must accurately diagnose and cater to patients' actual needs based on patients' particular condition, values, life-long preferences and worldviews (as opposed to perceived needs based on healthcare providers' sole assessment) (Farmer, 2004; Shapiro & Morley, 2022; Sherwin & Feminist Health Care Ethics Research Network, 1998; Mukherjee et al., 2022). In particular, when devising their interventions, healthcare systems and providers should consider and address the impact of social determinants of health and other structural impediments to staying healthy

and seeking and receiving healthcare (Daniels, 2001; Farmer, 2004; Sherwin & Feminist Health Care Ethics Research Network, 1998; Baylis et al., 2008; Llewellyn & Downie, 2012; Voigt & Wester, 2015; Powers, 2019). Such structural impediments may include existing social dependencies due to social roles (e.g. primary caregiver) or power imbalances (e.g. due to socioeconomic disparities), pervasive discrimination based on group affiliation (e.g. sex, race), implicit human bias, etc. Disregarding the effects of such external factors on health reinforces their controlling influence and thus further inhibits patients' agency (Llewellyn & Downie, 2012; Sherwin & Feminist Health Care Ethics Research Network, 1998). Hence, for healthcare interventions to be fair, they need to adequately account for and address patients' actual health needs, including the health effects of the associated aggravating factors, such as patients' underlying personal, health and socioeconomic circumstances.

Applied Ethics Aspects

The bioethics literature and empirical research also highlight the impact of various contingencies, such as patients' comorbidities, social roles (e.g. caregiver, frontline worker) and social determinants of health (e.g. educational attainment, income), on patients' health and health needs (Daniels, 2001; Harvard T.H. Chan School of Public Health, 2024; Nguyen et al., 2020; Voigt & Wester, 2015; Yearby, 2020). Various scholarship underscores the significance of several interventions for ML fairness. First, ML systems must be designed, developed and operate in consideration of the underlying social and historical context (e.g. social determinants of health, structural disparities) in which they will be deployed (Barocas et al., 2019; Benjamin, 2019; Bennett & Keyes, 2020; Birhane, 2021; Dignum, 2022; Faden et al., 2013; Gichoya et al., 2023; Hoffmann, 2019; London, 2019, 2022; Malanga et al., 2018; McDougall, 2019; Metcalf & Crawford, 2016; Ravitsky, 2024; Sauer et al., 2022; Xiang, 2021; Zook et al., 2017). Furthermore and consequently, training and validation data must be representative of the respective patient population and underlying context (Barocas et al., 2019; Bennett & Keyes, 2020; Dignum, 2022; Faden et al., 2013; Gichoya et al., 2023; Hoffmann, 2019; London, 2019, 2022; Malanga et al., 2018; Metcalf & Crawford, 2016; Mühlhoff, 2023; Ravitsky, 2024; Sauer et al., 2022; Xiang, 2021; Zook et al., 2017). Other important instruments for overcoming bias and achieving fairness in ML include ensuring diversity, inclusion and deliberation throughout the processes of ML design and development (Barocas et al., 2019; Benjamin, 2019; Bennett & Keyes, 2020; Birhane, 2021; Dignum, 2022; Faden et al., 2013; Hoffmann, 2019; London, 2019, 2022; Malanga et al., 2018; McDougall, 2019; Metcalf & Crawford, 2016; Mühlhoff, 2023; Ravitsky, 2024; Sauer et al., 2022; Xiang, 2021). Technical robustness of and privacy preservation by ML systems contribute to its fairness by safeguarding the accuracy and reliable performance of the systems (Mühlhoff, 2021, 2023). These substantive, procedural and organizational guardrails need to be present and operate in conjunction in order to mitigate bias and unfair outcomes in ML.

DISCUSSION

The choice of the three sets of key model parameters discussed above is problematic for ensuring fairness in radiology ML systems. On the one hand, these technical approaches fail to consider all relevant disease characteristics, including those caused by external factors. On the other hand, the approaches do not (sufficiently) consider how these characteristics and factors compound, intersect and depend on the patient's personal, other health and social conditions. As a result, ML systems exhibiting these deficiencies cannot accurately diagnose and address patients' actual health needs and their various drivers. These ML systems therefore lack context-sensitivity and appreciation of patients' intersectionality and of the social embeddedness of patients' health.

Intersectionality and social embeddedness, however, undergird fairness in fundamental ways. Patients' comorbidities and personal and socioeconomic circumstances influence and sometimes profoundly shape patients' health and health needs (Daniels, 2001; Farmer, 2004; Harvard T.H. Chan

School of Public Health, 2024; Voigt & Wester, 2015; Yearby, 2020). Addressing these needs is crucial for patients' functioning, agency and flourishing, which are the essence and ultimate shared goal of justice under the dominant moral theories, as explained above. To adequately diagnose and address patients' health needs and realize the essence and goal of justice, radiology ML systems must therefore incorporate and address not only the endogenous disease causes but also the health effects of patients' relevant personal, health and socioeconomic circumstances (e.g. comorbidities, health determinants). That means, ML systems must be designed, developed and used with patients' intersectionality and the social embeddedness of patients' health in mind. We propose a set of solutions to attain this.

First, we recommend a paradigm shift in the ways that fairness is conceptualized and designed in radiology ML systems. ML developers and users should not merely work from a notion of fairness as strict equality when building and deploying ML systems, and should not chiefly aim for equal predictive accuracy across subgroups of patients. Rather, ML systems should be designed and operate from the shared meaning and inherent purpose of justice as due consideration of patients' actual health needs. This means, ML developers and users should construct and calibrate systems to accurately screen and predict patients' actual health needs. In particular, ML systems should track (including based on self-reporting) all these disease features that adequately capture these needs, including features (e.g. pain, stress) caused by external factors stemming from the patient's circumstances (e.g. social insecurity) of relevance to the treated condition. Moreover, ML systems should be designed to predict patient needs, not human clinical judgements. In these ways, ML systems will integrate the relevant context and social embeddedness of patients' health and their intersectionality.

To institute this paradigm shift and further ensure fairness, we also propose a framework of ethical requirements aimed at more context-sensitivity and regard to patients' intersectionality and social situatedness. The requirements are listed in the table below and accompanied by practical examples for further clarification. This framework is intended both for direct practical application and in support of policy initiates. Healthcare providers and ML developers can consult the framework when designing and building radiology ML systems. Policymakers, standard-setting and industry organizations and internal policy teams at AI developers and users can resort to the framework when crafting relevant policies, standards, guidelines, codes of conduct and internal governance.

The proposed ethical requirements center context-sensitivity, intersectionality and social embeddedness through five main categories of requirements: (1) integrating in ML systems all relevant disease features and their adequate measurable medical indicators, (2) ethically sourced, diverse, representative and correct patient data concerning these features and indicators, (3) iterative socially sensitive co-exploration and co-design of datasets and ML systems involving all relevant stakeholders, including representatives of socially marginalized patient populations, (4) diversity and competence in ML teams, and (5) technically robust, privacy-preserving, safe and secure functioning of ML systems that safeguards their fair performance and output.

Ethical Requirement	**Practical Example**
Fairness as Due Regard	
Work from a notion of fairness as giving due consideration to patients' actual health needs to enable patients' functioning and flourishing.	ML developers and health providers should not conceptualize fairness as some sort of parity (e.g. predictive parity) but as due consideration of patients' actual health needs. Such due consideration should encompass all relevant disease features and their external aggravating factors. For example: Drawing on Pierson et al.'s study on osteoarthritis, for an ML system to be fair, it should account for both causes internal to the

	knee (e.g. radiology severity of osteoarthritis and/or structural damages) and causes external to the knee (e.g. life stress that leads to higher experienced pain); Drawing on Seyyed-Kalantari et al's study, an ML system should account for delays in receiving treatment and how they might have exacerbated the health condition.
Adequate Target Function	
Design the machine learning system to predict patients' health needs, not clinical judgements.	AI developers should design a machine learning system to predict patients' actual needs based on all relevant disease features, e.g. biomarkers, pain levels, and their external aggravating factors, e.g. life stress, social determinants of health, delayed access to diagnosis and treatment, as per the examples above. ML systems should not be designed to predict from human clinical judgements, which are often biased, as Seyyed-Kalantari et al. note with respect to clinical notes and reports, manual image labeling, biased diagnosis by doctors of under-served subpopulations.
Relevant Target Variables	
Select as target variables relevant disease features that capture patients' actual health needs and all their aggravating (including external) factors. Integrate corresponding medical indicators into the machine learning system.	As target variables should be selected relevant disease features that reflect both endogenous and external factors and accurately and comprehensively reflect patients' actual health needs, such as: Drawing on Pierson et al.'s, radiology severity of osteoarthritis, structural damages in the knee and self-reported pain levels; Drawing on Seyyed-Kalantari et al, less regular / significantly delayed medical check-ups as an indicator of possible under-diagnosis/treatment and potential for (quick) aggravation of the patient's health condition. Corresponding medical indicators should be included in the measurement metrics of the system, e.g. appropriate pain scoring/measurement metrics.
Predictive Accuracy	
Optimize predictive accuracy. Do not compromise it in order to achieve predictive parity among sub-groups of patients.	AI developers should not compromise predictive accuracy in order to achieve equal predictive parity across sub-groups of patients (which e.g. Seyyed-Kalantari et al. refer to as a common fairness measure). AI developers should optimize predictive accuracy in order to establish patients' actual health needs as precisely as possible.
Reliable Training Data	
Ethically source diverse, representative and correct	AI developers and healthcare providers should secure training and

patient data concerning health needs and corresponding medical indicators.	validation datasets that are: ethically sourced, i.e. sourced patient data based on valid informed patient consent; diverse, i.e. reflecting the variety of sub-groups of patients and their intersectionality based on ethnicity, socioeconomic status, age, sex, gender, etc.; representative, i.e. illustrative, typical of the patient population, its sub-groups (based on race, socioeconomic status, age, sex. gender, etc.) and their typical living conditions; correct, i.e. accurate.
Co-Participation	
Involve all relevant stakeholders (incl. disadvantaged groups) in iterative co-exploration of health needs and disease features, and co-design of the machine learning system.	AI developers and healthcare providers should involve in the exploration and selection of target variables, target function and other relevant model parameters all relevant stakeholders, e.g. representatives of physicians who will be using the ML system, patients, including representatives of underserved sub-groups of patients (e.g. Black/Hispanic, female, of low socioeconomic status), payors, etc.
Diversity and Competence	
Ensure diversity of views, lived experiences and competences in the machine learning development team to better understand and address patients' health needs and circumstances.	AI developers should include in the ML team data scientists, computer engineers, social scientists, etc., with expertise, experience, and skill sets that would allow them to identify and better understand the relevance and peculiarity of the disease features and external factors at play. For example, data scientists and social scientists with background in/knowledge of the social determinants of health of relevance to the respective health condition.
Technical Robustness	
Ensure technically robust, privacy-preserving, safe and secure functioning of the machine learning system to safeguard its fair performance and output.	AI developers should take technical measures that would ensure robust performance of the ML systems and would avert interference with its operations in ways that may skew/bias the output of the ML system. For example, adequate calibration and debugging of the ML model, state of the art cyber-security measures, privacy-preserving techniques (e.g. federated learning, differential privacy).

Table 1. Framework of Ethical Requirements for Achieving Fairness by Integrating Patients' Intersectionality and Social Embeddedness of Their Health. Illustrated with Practical Examples as per Seyyed-Kalantari et al. (Seyyed-Kalantari et al., 2021) and Pierson et al. (Pierson et al., 2021).

LIMITATIONS AND FUTURE WORK

Our research and recommendations have three main limitations. We have not analyzed in detail the legal viability of the proposed conceptualization of fairness in light of the prevalent legal paradigm of fairness as equality. Although controversial, the U.S. legal doctrines of disparate treatment and disparate impact as forms of discrimination appear to possibly inhibit the pursuit of fairness in any forms other than strict equality (Barocas, Solon; Selbst, Andrew D, 2016; Wachter et al., 2021; Xiang, 2021). The European Union non-discrimination laws and court jurisprudence appear to offer more wiggle room in this regard (Wachter et al., 2021). These aspects, however, require further research and assessment.

Outside the scope of our research also remained the question of whether appropriate technical approaches and measures existed, or could be developed, for implementing the proposed conceptualization of fairness and the first three ethical requirements of the framework. The question requires further examination as it may impact the actual operationalization of these conceptualization and requirements.

Another potential constraint on the recommendations above is the determination of adequate medical indicators of relevant disease factors. Empirical research, for example, on osteoarthritis has demonstrated that such adequate medical indicators (e.g. self-reported pain levels) exist and can be successfully used in ML-assisted diagnostics (Pierson et al., 2021). However, recent studies highlight the complexity of pinpointing uniform yet comprehensive population descriptors in genetics and genomics research given the interplay of historic, social and biological factors that potentially drive genetic variations (Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research et al., 2023). We cannot exclude that similar considerations apply with regard to some disease features and corresponding medical indicators. However, we have not probed the matter further.

CONCLUSION

Radiology ML systems should operationalize a notion of fairness that centers patients' needs. To achieve this, ML systems must incorporate all relevant patients' circumstances and address their impact on patients' health needs. The processes of designing and developing ML systems should also become more inclusive and deliberative. These avenues towards fairness may have wider implications for the field of ML as a whole as it would need to open up and reconnect with the social realities in healthcare and more broadly. Possible future research directions transpiring from our research include the questions of: (a) how the dominant legal paradigm of fairness as equality (e.g. disparate treatment and disparate impact doctrines in the United States) may have to also evolve in order to respond to the public appeals and ML's further attempts for more social justice, (b) what adequate medical indicators of all relevant disease factors would constitute, and (c) if appropriate technical approaches and measures exist or can be developed that can realize the recommendations of this paper.

ACKNOWLEDGEMENTS

I gratefully acknowledge Prof. Vardit Ravitsky, President & CEO of The Hastings Center, Dr. Elizabeth Nilson, Director, Medical Ethics, at Lahey Hospital, and Dr. Judy Gichoya, associate professor at Emory University School of Medicine, for their feedback to and guidance on my capstone work on which this publication is based.

REFERENCES

Allen, D. (2023). *Justice by Means of Democracy*. University of Chicago Press. http://ebookcentral.proquest.com/lib/harvard-ebooks/detail.action?docID=7194666

Anderson, E. (2010). *The Imperative of Integration*. Princeton University Press. https://muse.jhu.edu/pub/267/monograph/book/36293

Anderson, E. (2012). Epistemic Justice as a Virtue of Social Institutions. *Social Epistemology*, *26*(2), 163–173. https://doi.org/10.1080/02691728.2011.652211

Anderson, E. (2015). EQUALITY AND FREEDOM IN THE WORKPLACE: RECOVERING REPUBLICAN INSIGHTS. *Social Philosophy and Policy*, *31*(2), 48–69. https://doi.org/10.1017/S0265052514000259

Anderson, E. S. (1999). What Is the Point of Equality? *Ethics*, *109*(2), 287–337. https://doi.org/10.1086/233897

Arneson, R. (2018). Four Conceptions of Equal Opportunity. *The Economic Journal*, *128*(612), F152–F173. https://doi.org/10.1111/ecoj.12531

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. https://fairmlbook.org/

Barocas, Solon; Selbst, Andrew D. (2016). *Big Data's Disparate Impact*. https://doi.org/10.15779/Z38BG31

Barry, B. (2005). *Why social justice matters*. Polity.

Bates, D. W., & Zimlichman, E. (2015). Finding patients before they crash: The next major opportunity to improve patient safety. *BMJ Quality & Safety*, *24*(1), 1–3. https://doi.org/10.1136/bmjqs-2014-003499

Baumann, J., Hertweck, C., Loi, M., & Heitz, C. (2023). *Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics* (arXiv:2206.02897). arXiv. https://doi.org/10.48550/arXiv.2206.02897

Baylis, F., Kenny, N. P., & Sherwin, S. (2008). A Relational Account of Public Health Ethics. *Public Health Ethics*, *1*(3), 196–209. https://doi.org/10.1093/phe/phn025

Beheshtian, E., Putman, K., Santomartino, S. M., Parekh, V. S., & Yi, P. H. (2023). Generalizability and Bias in a Deep Learning Pediatric Bone Age Prediction Model Using Hand Radiographs. *Radiology*, *306*(2), e220505. https://doi.org/10.1148/radiol.220505

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity.

Bennett, C. L., & Keyes, O. (2020). What is the point of fairness?: Disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing*, *125*, 1–1. https://doi.org/10.1145/3386296.3386301

Bernhardt, M., Jones, C., & Glocker, B. (2022). Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, *28*(6), 1157–1158. https://doi.org/10.1038/s41591-022-01846-8

Binns, R. (2020). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–524. https://doi.org/10.1145/3351095.3372864

Binns, R. (2021). *Fairness in Machine Learning: Lessons from Political Philosophy* (arXiv:1712.03586). arXiv. http://arxiv.org/abs/1712.03586

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Cavaliere, G., Devolder, K., & Giubilini, A. (2019). Regulating Genome Editing: For an Enlightened Democratic Governance. *Cambridge Quarterly of Healthcare Ethics*, *28*(1), 76–88. https://doi.org/10.1017/S0963180118000403

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, *30*(3), 850–862. https://doi.org/10.1038/s41591-024-02857-3

Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, Board on Health Sciences Policy, Committee on Population, Health and Medicine

Division, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine. (2023). *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field* (p. 26902). National Academies Press. https://doi.org/10.17226/26902

Cooper, A. F., Abrams, E., & NA, N. (2021). Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 46–54. https://doi.org/10.1145/3461702.3462519

Daniels, N. (2001). Justice, Health, and Healthcare. *American Journal of Bioethics*, *1*(2), 2–16. https://doi.org/10.1162/152651601300168834

Dignum, V. (2022). *Relational Artificial Intelligence* (arXiv:2202.07446). arXiv. http://arxiv.org/abs/2202.07446

Faden, R. R., Kass, N. E., Goodman, S. N., Pronovost, P., Tunis, S., & Beauchamp, T. L. (2013). An Ethics Framework for a Learning Health Care System: *A Departure from Traditional Research Ethics and Clinical Ethics*. *Hastings Center Report*, *43*(s1). https://doi.org/10.1002/hast.134

Farmer, P. (2004). Rethinking Medical Ethics: A View From Below. *Developing World Bioethics*, *4*(1), 17–41. https://doi.org/10.1111/j.1471-8731.2004.00065.x

Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L.-C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.-C., Kuo, P.-C., Lungren, M. P., Palmer, L. J., Price, B. J., Purkayastha, S., Pyrros, A. T., Oakden-Rayner, L., Okechukwu, C., Seyyed-Kalantari, L., … Zhang, H. (2022). AI recognition of patient race in medical imaging: A modelling study. *The Lancet Digital Health*, *4*(6), e406–e414. https://doi.org/10.1016/S2589-7500(22)00063-2

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., Seyyed-Kalantari, L., Trivedi, H., & Purkayastha, S. (2023). AI pitfalls and what not to do: Mitigating bias in AI. *The British Journal of Radiology*, *96*(1150), 20230023. https://doi.org/10.1259/bjr.20230023

Gomberg, P. (2010). Dilemmas of Rawlsian Opportunity. *Canadian Journal of Philosophy*, *40*(1), 1–24. https://doi.org/10.1353/cjp.0.0085

Hanneman, K., Playford, D., Dey, D., Van Assen, M., Mastrodicasa, D., Cook, T. S., Gichoya, J. W., Williamson, E. E., Rubin, G. D., & on behalf of the American Heart Association Council on Cardiovascular Radiology and Intervention; and Council on Lifelong Congenital Heart Disease and Heart Health in the Young. (2024). Value Creation Through Artificial Intelligence and Cardiovascular Imaging: A Scientific Statement From the American Heart Association. *Circulation*, *149*(6). https://doi.org/10.1161/CIR.0000000000001202

Harvard T.H. Chan School of Public Health (Director). (2024, February 27). *How air pollution impacts our brains*. https://www.youtube.com/watch?v=PUZw-jg5SiE

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Holzinger, A., Haibe-Kains, B., & Jurisica, I. (2019). Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging*, *46*(13), 2722–2730. https://doi.org/10.1007/s00259-019-04382-9

Hsieh, N. & Department of Philosophy, Florida State University. (2005). Rawlsian Justice and Workplace Republicanism: *Social Theory and Practice*, *31*(1), 115–142. https://doi.org/10.5840/soctheorpract20053116

Jennings, B. (Ed.). (2014). *Bioethics* (Fourth edition). Macmillan Reference USA, a part of Gale, Cengage Learning.

John, H. S. (1970). Equality of Opportunity, and Beyond. In *Power, Authority, Justice & Rights* (1st ed., pp. 135–153). Routledge. https://doi.org/10.4324/9781315127170-11

Kymlicka, W. (2001). Contemporary Political Philosophy: An Introduction. In *Contemporary Political Philosophy*. Oxford University Press. https://www.oxfordpoliticstrove.com/display/10.1093/hepl/9780198782742.001.0001/hepl-9780198782742

Llewellyn, J. J., & Downie, J. G. (Eds.). (2012). *Being relational: Reflections on relational theory and health law*. UBC Press.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973

London, A. J. (2022). Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care? *Cell Reports Medicine*, *3*(5), 100622. https://doi.org/10.1016/j.xcrm.2022.100622

MacIntyre, A. (2013). *After Virtue* (Reprint edition). Bloomsbury Academic.

Malanga, S. E., Loe, J. D., Robertson, C. T., & Ramos, K. S. (2018). Who's Left Out of Big Data?: How Big Data Collection, Analysis, and Use Neglect Populations Most in Need of Medical and Public Health Research and Interventions. In I. G. Cohen, H. F. Lynch, E. Vayena, & U. Gasser (Eds.), *Big Data, Health Law, and Bioethics* (1st ed., pp. 98–111). Cambridge University Press. https://doi.org/10.1017/9781108147972.010

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, *3*(1), 205395171665021. https://doi.org/10.1177/2053951716650211

Mill, J. S. (1864). *On liberty*. London : Longman, Green, Longman, Roberts & Green. http://archive.org/details/onlibertyooinmill

Mill, J. S. (2008). *The Subjection of Women*. https://www.gutenberg.org/ebooks/27083

Mittelstadt, B., Wachter, S., & Russell, C. (2023). *The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default* (arXiv:2302.02404). arXiv. https://doi.org/10.48550/arXiv.2302.02404

Mühlhoff, R. (2021). Predictive privacy: Towards an applied ethics of data analytics. *Ethics and Information Technology*, *23*(4), 675–690. https://doi.org/10.1007/s10676-021-09606-x

Mühlhoff, R. (2023). Predictive privacy: Collective data protection in the context of artificial intelligence and big data. *Big Data & Society*, *10*(1), 205395172311668. https://doi.org/10.1177/20539517231166886

Mukherjee, P., Shen, T. C., Liu, J., Mathai, T., Shafaat, O., & Summers, R. M. (2022). Confounding factors need to be accounted for in assessing bias by machine learning algorithms. *Nature Medicine*, *28*(6), 1159–1160. https://doi.org/10.1038/s41591-022-01847-7

Nguyen, L. H., Drew, D. A., Joshi, A. D., Guo, C.-G., Ma, W., Mehta, R. S., Sikavi, D. R., Lo, C.-H., Kwon, S., Song, M., Mucci, L. A., Stampfer, M. J., Willett, W. C., Eliassen, A. H., Hart, J. E., Chavarro, J. E., Rich-Edwards, J. W., Davies, R., Capdevila, J., … Chan, A. T. (2020). Risk of COVID-19 among frontline healthcare workers and the general community: A prospective cohort study. *medRxiv*, 2020.04.29.20084111. https://doi.org/10.1101/2020.04.29.20084111

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*.

Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S., & Obermeyer, Z. (2021). An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, *27*(1), 136–140. https://doi.org/10.1038/s41591-020-01192-7

Pot, M., Kieusseyan, N., & Prainsack, B. (2021). Not all biases are bad: Equitable and inequitable biases in machine learning and radiology. *Insights into Imaging*, *12*(1), 13. https://doi.org/10.1186/s13244-020-00955-7

Powers, M. (2019). *Structural injustice: Power, advantage, and human rights*. Oxford University Press.
Rae, D. W. (1981). *Equalities*. Harvard University Press.
Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, *169*(12), 866. https://doi.org/10.7326/M18-1990
Ravitsky, V. (2024). *2024 Rothenberger Speaker Series*. Zoom. https://umaryland.zoom.us/rec/play/YeEA9OC4MNQaOvqQd8r4cwAuBxi5sj5A4sn0YiFvdzr-4BDp2x1rLumbb7tz6WePs24OCiou72JznfoH.IgaSnw7N_Nsh3njj
Rawls, J. (1971). *A theory of justice*. Belknap Press of Harvard University Press.
Roemer, J. E. (1995). Equality and Responsibility. *Boston Review*. https://www.bostonreview.net/forum/equality-and-responsibility/
Sandel, M. J. (2010). *Justice: What's the right thing to do?* (First paperback edition.). Farrar, Straus and Giroux.
Satariano, A., Metz, C., & Times, A. S. F. T. N. Y. (2023, March 5). Using A.I. to Detect Breast Cancer That Doctors Miss. *The New York Times*. https://www.nytimes.com/2023/03/05/technology/artificial-intelligence-breast-cancer-detection.html
Sauer, C. M., Chen, L.-C., Hyland, S. L., Girbes, A., Elbers, P., & Celi, L. A. (2022). Leveraging electronic health records for data science: Common pitfalls and how to avoid them. *The Lancet Digital Health*, *4*(12), e893–e898. https://doi.org/10.1016/S2589-7500(22)00154-6
Scanlon, T. M. (2003). The diversity of objections to inequality. In *Cambridge University Press eBooks* (pp. 202–218). Cambridge University Press. https://doi.org/10.1017/CBO9780511615153.012
Sen, A. (1995). *Inequality Reexamined*. Oxford University Press. https://doi.org/10.1093/0198289286.001.0001
Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y., & Ghassemi, M. (2020). *CheXclusion: Fairness gaps in deep chest X-ray classifiers* (arXiv:2003.00827). arXiv. http://arxiv.org/abs/2003.00827
Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, *27*(12), 2176–2182. https://doi.org/10.1038/s41591-021-01595-0
Shapiro, D., & Morley, G. (2022). Evaluating Decision-Making Capacity: When a False Belief about Ventilators Is the Reason for Refusal of Life-Sustaining Treatment. *The Journal of Clinical Ethics*, *33*(1), 50–57. https://doi.org/10.1086/JCE2022331050
Sherwin, S. & Feminist Health Care Ethics Research Network. (1998). *The politics of women's health: Exploring agency and autonomy*. Temple University Press.
Voigt, K. (2019). Social Justice, Equality and Primary Care: (How) Can 'Big Data' Help? *Philosophy & Technology*, *32*(1), 57–68. https://doi.org/10.1007/s13347-017-0270-6
Voigt, K., & Wester, G. (2015). RELATIONAL EQUALITY AND HEALTH. *Social Philosophy and Policy*, *31*(2), 204–229. https://doi.org/10.1017/S0265052514000326
Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3792772
Xiang, A. (2021). *Reconciling Legal and Technical Approaches to Algorithmic Bias* (SSRN Scholarly Paper 3650635). https://papers.ssrn.com/abstract=3650635
Yearby, R. (2020). Structural Racism and Health Disparities: Reconfiguring the Social Determinants of Health Framework to Include the Root Cause. *Journal of Law, Medicine & Ethics*, *48*(3), 518–526. https://doi.org/10.1177/1073110520958876

Young, I. M. (2011). *Justice and the politics of difference* (Paperback reissue / with a new foreword by Danielle Allen.). Princeton University Press.

Young, I. M., & Nussbaum, M. (2011). *Responsibility for Justice*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.001.0001

Yu, F., Moehring, A., Banerjee, O., Salz, T., Agarwal, N., & Rajpurkar, P. (2024). Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine*, 1–13. https://doi.org/10.1038/s41591-024-02850-w

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology*, *13*(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399