Perspective: Machine Learning for Health Should Consider Social Drivers of Health

Neha Srivathsa NEHASRIV@STANFORD.EDU
Department of Computer Science, Stanford University School of Engineering

Sherri Rose SHERRIROSE@STANFORD.EDU
Department of Health Policy, Stanford University School of Medicine

Abstract

Clinical machine learning (ML) algorithms can exacerbate a wide range of injustices across multiple domains and levels of society. These harms are often underemphasized and differentially distributed, with minoritized communities disproportionately experiencing the harms and not the benefits of health ML algorithms. By proposing a correspondence between prominent algorithmic harm and social drivers of health (SDOH) frameworks, we show that a range of algorithmic harms ultimately impact human health through SDOH factors, especially structural factors. This presents an inherent tension in the development of ML for health, where the harms of algorithms may lead to the worsening of health inequities. We recommend the consideration of SDOH throughout the pipeline of ML system development for examining algorithmic harms to health. Effectively considering SDOH necessitates developing competencies in structural analysis and community-engaged approaches. Accounting for SDOH could illuminate pathways toward equity-promoting algorithms, although we highlight that, in many cases, an equity-promoting algorithm may not exist. Thus, we also emphasize the need for interventions on the root causes of health inequities.

Keywords: social drivers of health, structural drivers of health, algorithmic harms, machine learning for health, structural analysis, community-engaged research, health inequities

1. Introduction

Clinical machine learning (ML) algorithms, although widely portrayed as offering significant potential for improving healthcare outcomes, are known to exacerbate a wide spectrum of injustices (Chen et al., 2021). These impact multiple domains and levels of society, ranging from systematic underdiagnosis for minoritized groups (Seyyed-Kalantari et al., 2021), to ruptures in institutional trustworthiness for populations who already have histories of medical marginalization (Nong and Platt, 2025), to environmental and ecological degradation (Brevini, 2020). Despite the existence of several frameworks detailing the multilevel contours of algorithmic harms (Suresh and Guttag, 2021; Shelby et al., 2023; Abercrombie et al., 2024; Domínguez Hernández et al., 2024), evaluations of clinical ML algorithms tend to focus on individual impact and responsibility, partially rooted in biomedical ethics (Vandemeulebroucke et al., 2022; Vandemeulebroucke, 2025). This leads to a prioritization of issues pertaining to algorithmic bias, fairness, safety, and data privacy, while other harms remain underemphasized (Murphy et al., 2021; van Uffelen et al., 2024).

In this perspective, we argue for the comprehensive consideration of social drivers of health (SDOH) in the conceptualization, creation, evaluation, and use of clinical algorithms. By proposing a correspondence between algorithmic harm and SDOH frameworks, we show how algorithmic harms can be understood as harms to factors that fundamentally impact health outcomes and inequities, especially upstream structural factors. Neglecting to examine the full range of harms caused by the health ML pipeline thus threatens to worsen health inequities, in contradiction to the purported goal of ML for health. We recommend an SDOH-conscious algorithmic praxis, modeled after a recent call for racism-conscious praxis in medicine (Khazanchi et al., 2023), to accurately consider the distribution of benefits versus harms of clinical algorithms, and to facilitate the development of equity-promoting algorithmic systems for health. Additionally, we highlight that interventions on the root causes of health inequities are essential and that, in many contexts, an equity-promoting algorithm may not exist.

2. Correspondence between social drivers and algorithmic harms

SDOH are the non-medical factors that impact health outcomes and risks, and they are known to be key drivers of health inequities (Solar and Irwin, 2010). They comprise diverse social domains, including economic stability, education, health care access, neighborhood and built environment, and social and community context (HHS, 2024). The *conceptual framework for action on the social determinants of health* (Solar and Irwin, 2010), developed by the World Health Organization, organizes SDOH into two levels along a spectrum of causal priority: structural and intermediary determinants. Structural determinants are upstream factors that generate social stratification and locate individuals within systems of power, including social, economic, and political forces. Intermediary determinants include community-level and individual-level factors that directly impact patients' health outcomes, including material, psychological, social, behavioral, and biological factors. The structural determinants, by virtue of being located causally upstream of and influencing the inequitable distribution of intermediary factors, are also known as social drivers of health inequities. They shape the allocation of intermediary factors, which then impact downstream health outcomes. This conceptual framework was developed in explicit acknowledgment of the diversity of domains that these factors comprise, encouraging an intersectoral approach to addressing health inequities (Solar and Irwin, 2010). Thus, the framework is appropriate for algorithmic researchers to consider in their work.

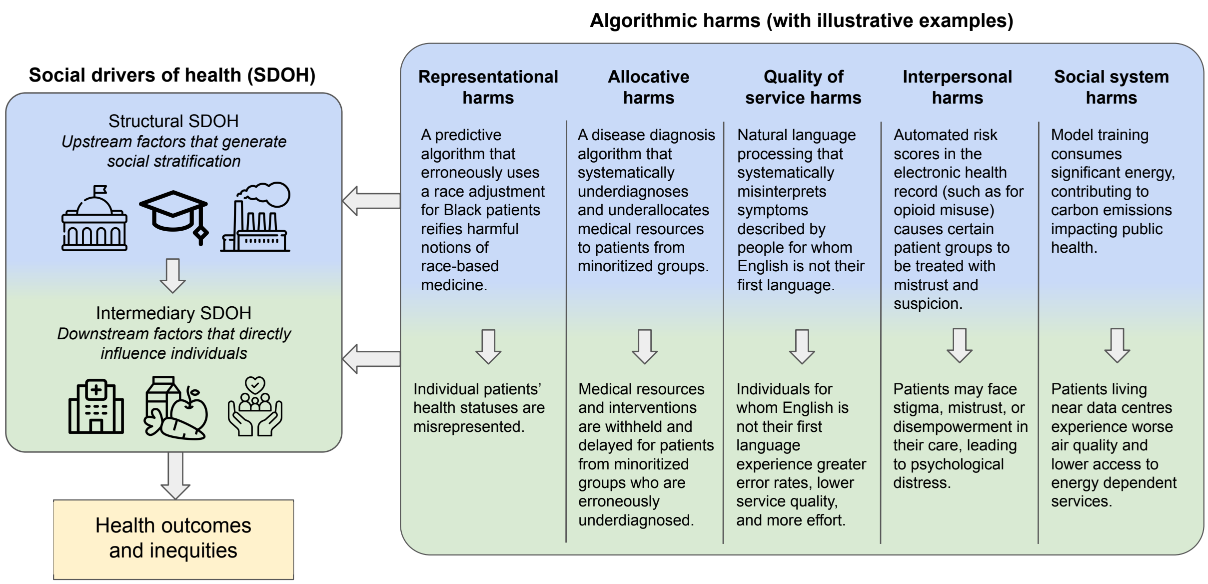
In recent years, a lineage of work from responsible artificial intelligence (AI), AI ethics, and related fields has explored the landscape of algorithmic harms (Suresh and Guttag, 2021; Shelby et al., 2023; Abercrombie et al., 2024; Domínguez Hernández et al., 2024). A key distinction was made between representational harms versus allocative harms (Barocas et al., 2017), where representational harms contain harms caused due to negatively stereotyping, erasing, or alienating social groups, while allocative harms comprise the inequitable (mis)allocation of resources and opportunities. Following this, and due to the existence of harms not accounted for in this binary, subsequent constructs expanded to add further categories. A prominent taxonomy for the sociotechnical harms of algorithmic systems delineates five categories of harms: representational, allocative, quality of service, interpersonal, and social system harms (Shelby et al., 2023). Quality of service harms involve the alienation and lower quality of services experienced by minoritized groups, including the additional labor they must perform to experience comparable conditions. Interpersonal harms entail impacts to individual and community relationships, including impacts to trust, trustworthiness, privacy, autonomy, agency, and empowerment. Social system harms impact broader societal processes and structures, including the environment and ecology, political and civic systems, and cultures. These five harm categories are not mutually exclusive, and this framework explicitly notes that they may occur at different levels of society, such as the macro-, meso-, or micro-level (Shelby et al., 2023).

Considering harms within the five sociotechnical algorithmic harm categories for health ML algorithms, we observe that each of the five harm categories fundamentally represents harms to human health via SDOH, especially at the structural level. Figure 1 captures examples of this correspondence, where algorithms that generate harms in each of the five harm categories negatively impact structural SDOH. These harms then generate impacts to intermediary SDOH, which cascade down to impact individual and community health outcomes and inequities. Thus, health ML systems introduce novel structural inequities impacting health into our social systems.

Additionally, the harms versus benefits of health ML algorithms are differentially distributed (Dauvergne, 2022). The communities experiencing certain harms are dissimilar from those experiencing benefits, and they may be already disproportionately affected by SDOH-related health inequities. Further, communities that bear certain harms, especially those located in the Global South—who disproportionately experience resource extraction of water, energy, and minerals, as well as labor exploitation (Crawford, 2021; Dauvergne, 2022; Fiske et al., 2025)—may be fundamentally unable to reap any benefits of health ML due to the material realities of their healthcare infrastructures (McCoy et al., 2025). For these communities, existing problems may be worsened by attempting to impose health ML.

Conceptualizing algorithmic harms as harms to health via SDOH illuminates the inherent contra-

Figure 1: Correspondence between social drivers of health (SDOH) (Solar and Irwin, 2010) and sociotechnical algorithmic harms (Shelby et al., 2023) frameworks. Each of the five categories of algorithmic harms impacts structural SDOH, which then impact downstream intermediary SDOH, followed by health outcomes and inequities.



diction in building health ML algorithms, which are aimed at improving health but may worsen health and especially health inequities. Previously, scholars have named the systematic underemphasis on the environmental harms of ML (Brevini, 2020; Crawford, 2021), calling out the tension between developing ML to improve human health while the underacknowledged climate and environmental impacts of health ML harm human health (Katirai, 2024; Fiske et al., 2025). We extend this argument by claiming that this is true not just for the climate and environmental harms of ML, but for many of the harms of ML across the five sociotechnical harm categories that ultimately impact SDOH, as shown in Figure 1.

3. Algorithmic praxis informed by social drivers

Praxis entails the process of circularly engaging in reflection based on theory, followed by putting theory into action, toward generating social and structural change (Freire, 1996). The conceptual framework of SDOH provides opportunities for identifying underemphasized sociotechnical harms and enabling the development of socially conscious and equity-promoting algorithms through considering the root causes of inequities. It thus provides an avenue for engaging in SDOH-informed algorithmic praxis. This means we must remain aware of how social considerations may be co-opted toward fairness-washing or ethics-washing initiatives, where superficial considerations of ethics or fairness replace substantive engagement (Schultz et al., 2025). In the context of SDOH, ethics- and fairness-washing could entail superficially including SDOH variables in health algorithms without analysis of structural causes. We must orient toward genuine reflection and engagement.

Just as health algorithms and technologies can materialize inequity—by codifying, reproducing, and obscuring oppressive social structures (Benjamin, 2019)—they can also be intentionally created to materialize health equity (Khazanchi et al., 2023). Until 2021, the clinical equation that was used to estimate kidney function and diagnose chronic kidney disease, called the eGFR equation, included a race-adjustment based on binarized race (Black versus not Black). This equation materialized racism and inequity through systematically estimating Black patients as being less sick than they were, delaying access to care, and reinforcing harmful essentialist ideas of race as biological (Eneanya et al., 2019; Vyas et al., 2020), thus causing allocative and representational harms. Although the current equation does not include race as a covariate, disparities persist (Cusick et al., 2024).

One potential approach toward materializing equity could involve investigating the complex web of intersecting social and structural factors that cause existing inequities, through considering SDOH, followed by redressing these inequities by developing algorithms that reallocate resources to those facing inequities. For instance, community-engaged or theoretical approaches could be used to construct a causal graph representing how social and structural variables impact health outcomes. Variables from this graph could then be incorporated as covariates in algorithms that inform resource allocation, to prioritize interventions for populations that may be at higher risk due to adverse SDOH, and prior work provides guidance on such inclusion of SDOH as covariates in predictive clinical algorithms (Foryciarz et al., 2025). Other examples in health applications have demonstrated integrating SDOH, algorithmic design, and equity-promoting resource allocation. For instance, a study from cardiac care piloted a digital clinical decision support tool that was created to actively recommend specialist care for Black and Latinx patients, who have been historically excluded, even when accounting for other social determinants (Wispelwey et al., 2022). Another study demonstrated the feasibility of equity-oriented extensions to Medicare risk adjustment algorithms for resource redistribution toward minoritized groups, given structural and social disparities (Reitsma et al., 2025). The SDOH framework thus provides a valuable starting point toward identifying structural and social causes of inequities, to then begin to redress them.

Effectively identifying root causes of inequities to develop equity-promoting algorithms requires that algorithmic researchers develop competencies in systems thinking and structural analysis, and form collaborative relationships with scholars and community practitioners who are experts in these areas. Frameworks developed for structural analysis in clinical practice and medical education could provide valuable tools for health ML researchers to investigate structural causes and impacts in diverse domains (Hassan and Bui, 2022). Importantly, equity-promoting clinical algorithms often cannot act upon the root causes of inequities, even as they can play a crucial role in equitable allocation of healthcare resources at a structural level. We emphasize that an algorithm may not be part of the solution and that an "equity-promoting" clinical algorithm may not exist. In these cases, we must remain cognizant of the risk of normalizing health inequities caused by SDOH, which could reinforce existing societal inequities (McCradden et al., 2020). While we can and should use SDOH to inform equitable resource allocation, as we recommend in this perspective, we must simultaneously call attention to the importance of structural interventions acting upon the root cause SDOH factors themselves.

In conjunction with structural analysis, community-engaged approaches, as introduced above, provide an avenue toward elucidating the social and structural factors that impact health outcomes (Parker et al., 2025). Social factors impacting certain inequities may be difficult to ascertain, and those directly experiencing these inequities are uniquely positioned to possess knowledge of these factors. Community-engaged approaches can additionally provide a pathway for communities to develop greater autonomy and influence in the systems that impact them (Asabor et al., 2024). Community-based system dynamics is one approach, aimed at modeling non-linear dynamics of complex social systems through co-creation of causal diagrams and quantitative simulation modeling, with directly impacted community members who have lived knowledge and experience (Hovmand, 2014). The use of community-based system dynamics in ML has previously been proposed (Martin Jr et al., 2020; Prabhakaran and Martin Jr, 2020), and this methodology has been used to identify causes of racial bias in clinical algorithms (Kuhlberg et al., 2020). The systems thinking approach within this methodology may be well suited to understanding the structural and social causes of health inequities, as well as the multi-domain and multi-level impacts of health ML algorithms, which may be otherwise challenging to model and measure.

Curriculum that introduces systems thinking, structural analysis, and community-engaged research could be included in applied ML courses. This would serve to introduce scholars to these areas and encourage deeper study as well as potentially lead to the formation of partnerships between early career ML scholars and experts in these approaches. Researchers affiliated with institutions that have entities specialized in developing effective and responsible community partnerships, such as offices of community engagement, can seek collaborations with these groups to center community-engaged approaches in their work. Additionally, substantial prior literature provides guidance on building community partnerships, developing trust and trustworthiness with community and patient partners, and conducting participatory or community-engaged research (Minkler and Wallerstein, 2011; Hovmand, 2014; Donia and Shaw, 2021; Birhane et al., 2022; Cohn, 2025).

Along with capturing communities' knowledge of the causes of structural inequities and algorithmic consequences, community-based system dynamics and other community-engaged approaches can raise the critical consciousness of community members (Decker and Wendel, 2023). Critical consciousness refers to the capacity to identify and take action against social and political oppression (Freire, 1996). Community-based system dynamics can create the conditions for communities to recognize the structural causes of their material realities, reframing their perspectives about the reasons for their individual experiences (Hovmand, 2014). When minoritized communities co-create models, they move from being objects of representation to subjects who claim their power to influence systems that represent them and may directly impact them. This praxis of agency to influence social conditions, which can be through influencing algorithmic systems, has been noted as an avenue for guiding interventions on SDOH (Solar and Irwin, 2010), and can increase the trustworthiness of algorithmic systems. This also begins to counter social narratives that ML systems cannot be understood, critiqued, or contested from outside of the systems of power that construct them, which further opens space for communities to engage with ML systems as subjects through reflection and action.

4. Conclusion

Examining the correspondence between algorithmic harm and SDOH frameworks reveals how the harms of health ML algorithms are harms to human health, displaying an inherent tension in the development of health ML. Conceptual SDOH frameworks provide an approach for more comprehensively considering the harms of health ML to health, especially illuminating underemphasized harms and the differential distribution of harms versus benefits. This can allow for a more accurate assessment of the impacts of algorithmic systems, which should inform the development of health ML algorithms. Importantly, the conclusion may be that certain systems should not be built or that certain existing systems should be dismantled. Identifying the SDOH factors that impact health inequities can also enable the development of algorithms that materialize an equitable distribution of healthcare resources, and we simultaneously highlight the need for interventions on the root causes of health inequities.

Acknowledgments

We thank Shravya Gurrapu, Maria Luiza Rocha Bueno, Vyoma Raman, and members of the Health Policy Data Science Lab at Stanford University for valuable feedback on this work. The authors acknowledge funding support from NIH Director's Pioneer Award DP1LM014278.

References

Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, et al. A collaborative, human-centred taxonomy of ai, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294*, 2024.

Emmanuella Ngozi Asabor, Kammarauche Aneni, Sitara Weerakoon, and Ijeoma Opara. Applying a community-engaged participatory machine learning model. *American Journal of Community Psychology*, 74(3-4):262–268, 2024.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual conference of the special interest group for computing, information and society*, volume 1. New York, NY, 2017.

Ruha Benjamin. Assessing risk, automating racism. *Science*, 366(6464):421–422, 2019.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

Benedetta Brevini. Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. *Big Data & Society*, 7(2): 2053951720935141, 2020.

Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4(1):123–144, 2021.

Elizabeth Cohn, editor. *Principles of Community Engagement, 3rd Edition*. U.S. Centers for Disease Control and Prevention, Atlanta, GA, 2025.

Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

Marika M Cusick, Glenn M Chertow, Douglas K Owens, Michelle Y Williams, and Sherri Rose. Algorithmic changes are not enough: Evaluating the removal of race adjustment from the egfr equation. In *Conference on Health, Inference, and Learning*, pages 619–643. PMLR, 2024.

Peter Dauvergne. Is artificial intelligence greening global supply chains? exposing the political economy of environmental costs. *Review of International Political Economy*, 29(3):696–718, 2022.

Hallie Decker and Monica Wendel. Applications of participatory system dynamics methods to public health: a systematic review. *Family & Community Health*, 46:S6–S21, 2023.

Andrés Domínguez Hernández, Shyam Krishna, Antonella Maia Perini, Michael Katell, SJ Bennett, Ann Borda, Youmna Hashem, Semeli Hadjiloizou, Sabeehah Mahomed, Smera Jayadeva, et al. Mapping the individual, social and biospheric impacts of foundation models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 776–796, 2024.

Joseph Donia and James A Shaw. Co-design and ethical artificial intelligence for health: An agenda for critical research and practice. *Big Data & Society*, 8(2):20539517211065248, 2021.

Nwamaka Denise Eneanya, Wei Yang, and Peter Philip Reese. Reconsidering the consequences of using race to estimate kidney function. *Jama*, 322(2):113–114, 2019.

Amelia Fiske, Isabella M Radhuber, Theresa Willem, Alena Buyx, Leo Anthony Celi, and Stuart McLennan. Climate change and health: the next challenge of ethical ai. *The Lancet Global Health*, 13 (7):e1314–e1320, 2025.

Agata Foryciarz, Nicole Gladish, David H Rehkopf, and Sherri Rose. Incorporating area-level social drivers of health in predictive algorithms using electronic health record data. *Journal of the American Medical Informatics Association*, page ocaf009, 2025.

Paulo Freire. Pedagogy of the oppressed (revised). *New York: Continuum*, 356:357–358, 1996.

Iman F Hassan and Thuy Bui. The structural analysis: incorporating structurally competent clinical reasoning into case-based presentations. *Journal of General Internal Medicine*, 37(13):3465–3470, 2022.

HHS. Social determinants of health, 2024. URL https://health.gov/healthypeople/objectives-and-data/social-determinants-health. *Healthy People 2030*. Office of Disease Prevention and Health Promotion, U.S. Department of Health and Human Services.

Peter S Hovmand. *Community Based System Dynamics*. Springer, 2014.

Amelia Katirai. The environmental costs of artificial intelligence for healthcare. *Asian Bioethics Review*, 16(3):527–538, 2024.

Rohan Khazanchi, Derek R Soled, and Ruqaiijah Yearby. Racism-conscious praxis: a framework to materialize anti-oppression in medicine, public health, and health policy. *The American Journal of Bioethics*, 23(4):31–34, 2023.

Jill A. Kuhlberg, Irene Headen, Ellis A. Ballard, and Jr. Donald Martin. Advancing community engaged approaches to identifying structural drivers of racial bias in health diagnostic algorithms. In *Proceedings of the 2020 Conference of the System Dynamics Society*, July 2020. URL https://proceedings.systemdynamics.org/2020.

Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. Participatory problem formulation for fairer machine learning through community based system dynamics. *ICLR Workshop on Machine Learning in Real Life*, 2020.

Liam G McCoy, Azra Bihorac, Leo Anthony Celi, Matthew Elmore, Divya Kewalramani, Teddy Kwaga, Nicole Martinez-Martin, Renata Prôa, Joel Schamroth, Jonathan D Shaffer, et al. Building health systems capable of leveraging ai: applying paul farmer's 5s framework for equitable global health. *BMC Global and Public Health*, 3(1):39, 2025.

Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

Meredith Minkler and Nina Wallerstein. *Community-Based Participatory Research for Health: From Process to Outcomes*. John Wiley & Sons, 2011.

Kathleen Murphy, Erica Di Ruggiero, Ross Upshur, Donald J Willison, Neha Malhotra, Jia Ce Cai, Nakul Malhotra, Vincci Lui, and Jennifer Gibson. Artificial intelligence for good health: a scoping review of the ethics literature. *BMC medical ethics*, 22(1):14, 2021.

Paige Nong and Jodyn Platt. Patients' trust in health systems to use artificial intelligence. *JAMA Network Open*, 8(2):e2460628–e2460628, 2025.

Andrea G Parker, Laura M Vardoulakis, Jatin Alla, and Christina N Harrington. Participatory ai considerations for advancing racial health equity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2025.

Vinodkumar Prabhakaran and Donald Martin Jr. Participatory machine learning using community-based system dynamics. *Health and Human Rights*, 22(2):71, 2020.

Marissa B. Reitsma, Thomas G. McGuire, and Sherri Rose. Algorithms to improve fairness in medicare risk adjustment. *JAMA Health Forum*, 2025. doi: 10.1001/jamahealthforum.2025.2640.

Mario D Schultz, Ludovico Giacomo Conti, and Peter Seele. Digital ethicswashing: A systematic review and a process-perception-outcome framework. *AI and Ethics*, 5(2):805–818, 2025.

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

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741, 2023.

Orielle Solar and Alec Irwin. A conceptual framework for action on the social determinants of health. Technical report, WHO Document Production Services, 2010.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, 2021.

Nynke van Uffelen, Lode Lauwaert, Mark Coeckelbergh, and Olya Kudina. Towards an environmental ethics of artificial intelligence. *arXiv preprint arXiv:2501.10390*, 2024.

Tijs Vandemeulebroucke. The ethics of artificial intelligence systems in healthcare and medicine: from a local to a global perspective, and back. *Pflügers Archiv-European Journal of Physiology*, 477(4):591–601, 2025.

Tijs Vandemeulebroucke, Yvonne Denier, and Chris Gastmans. The need for a global approach to the ethical evaluation of healthcare machine learning. *The American Journal of Bioethics*, 22(5):33–35, 2022.

Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine*, 383(9):874–882, 2020.

Bram P Wispelwey, Regan H Marsh, Michael Wilson, Sula Mazimba, Scott K Heysell, Imoigele P Aisiku, Akshay S Desai, Cassandra Georges, and Michelle E Morse. Leveraging clinical decision support for racial equity: a sociotechnical innovation. *NEJM Catalyst Innovations in Care Delivery*, 3(4), 2022.