Published August 31, 2023 as 10.3174/ajnr.A7963

Ethical Considerations and Fairness in the Use of Artificial Intelligence for Neuroradiology

C.G. Filippi, J.M. Stein, Z. Wang, S. Bakas, Y. Liu, P.D. Chang, Y. Lui, C. Hess, D.P. Barboriak, A.E. Flanders, M. Wintermark, G. Zaharchuk, and O. Wu

ABSTRACT

SUMMARY: In this review, concepts of algorithmic bias and fairness are defined qualitatively and mathematically. Illustrative examples are given of what can go wrong when unintended bias or unfairness in algorithmic development occurs. The importance of explainability, accountability, and transparency with respect to artificial intelligence algorithm development and clinical deployment is discussed. These are grounded in the concept of "primum no nocere" (first, do no harm). Steps to mitigate unfairness and bias in task definition, data collection, model definition, training, testing, deployment, and feedback are provided. Discussions on the implementation of fairness criteria that maximize benefit and minimize unfairness and harm to neuroradiology patients will be provided, including suggestions for neuroradiologists to consider as artificial intelligence algorithms gain acceptance into neuroradiology practice and become incorporated into routine clinical workflow.

ABBREVIATION: AI = artificial intelligence

Artificial intelligence (AI) is beginning to transform the practice of radiology, from order entry through image acquisition and reconstruction, workflow management, diagnosis, and treatment decisions. AI will certainly change neuroradiology practice across routine workflow, education, and research. Neuroradiologists are understandably concerned about how AI will affect their subspecialty and how they can shape its development. Multiple published consensus statements advocate the need for radiologists to play a primary role in ensuring that AI software used for clinical care is fair to and unbiased against specific groups of patients.[1] In this review, we focus on the need for developing and implementing fairness criteria and how to balance competing interests that minimize harm and maximize patient benefits when implementing AI solutions in neuroradiology. The responsibility for promoting health care equity rests with the entire neuroradiology community, from academic leaders to private practitioners. We all have a stake in establishing best practices as AI enters routine clinical practice.

Received January 30, 2023; accepted after revision July 7.

From the Department of Radiology (C.G.F.), Tufts University School of Medicine, Boston, Massachusetts; Department of Radiology (J.M.S., S.B.), University of Pennsylvania, Philadelphia, Pennsylvania; Athinoula A. Martinos Center for Biomedical Imaging (Z.W., Y. Liu, O.W.), Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts; Department of Radiological Sciences (P.D.C.), University of California, Irvine, California; Department of Neuroradiology (Y. Lui), NYU Langone Health, New York, New York; Department of Radiology and Biomedical Imaging (C.H.), University of California, San Francisco, San Francisco, California; Department of Radiology (D.P.B.), Duke University School of Medicine, Durham, North Carolina; Department of Neuroradiology/Otolaryngology (ENT) Radiology (A.E.F.), Thomas Jefferson University, Philadelphia, Pennsylvania; Department of Neuroradiology (M.W.), Division of Diagnostic Imaging, MD Anderson Cancer Center, Houston, Texas; and Department of Radiology (G.Z.), Stanford University, Stanford, California.

Please address correspondence to Christopher G. Filippi, MD, Tufts University School of Medicine, Department of Radiology, 800 Washington St, Box 299, Boston, MA 02111; e-mail: cfilippi@tuftsmedicalcenter.org; @sairaallapeikko

Indicates open access to non-subscribers at www.ajnr.org

Indicates article with online supplemental data.

http://dx.doi.org/10.3174/ajnr.A7963

Definitions

"Ethics," in a strict dictionary definition, is a theory or system of values that governs the conduct of individuals and groups.[2] Ethical physicians should endeavor to promote fairness and avoid bias in their personal treatment of patients and with respect to the health care system at large. A biased object yields 1 outcome more frequently than statistically expected, eg, a 2-headed coin. Similarly, a biased algorithm systematically produces outcomes that are not statistically expected. One proposed definition for algorithmic bias in health care systems is "when the application of an algorithm compounds existing inequities in socioeconomic status, race, ethnic background, religion, sex, disability, or sexual orientation to amplify them and adversely impact inequities in health systems."[3] This definition, while not ideal, is a request for developers and end users of AI algorithms in health care to be aware of the potential risk of poorly designed algorithms for not merely reflecting societal imbalances but also amplifying inequities.

"Fairness" can be defined as the absence of favoritism toward specific subgroups of populations.[4] Individual fairness is the principle that any 2 individuals who are similar should be treated equally.[5] In contrast, "group fairness," ie, statistical or demographic

parity, is the principle that the demographics of the group receiving positive or negative treatment are the same as the population as a whole.[5] Considering harm caused by algorithmic bias, ie, allocational (denial of opportunities or resources[6]) or representational (reinforcement of negative stereotypes[7]) harm, may be more intuitive.

Algorithmic Bias

We can quantify bias (δ) for AI models $f(x)$ with regard to bias features, z, as

$$\delta = \hat{M}\big(f(x), f(x,z)\big),$$

for which $\hat{M}$ is a distance metric that measures the difference between $f(x)$ and $f(x,z)$. The formula intuitively corresponds to the dictionary definition of "bias" of AI models by measuring how much the model outcomes $f(x,z)$ deviate from expected $f(x)$. Bias features z can be explicit (eg, sex, race, age, and so forth) or implicit (eg, data set imbalance, model architectures, poorly chosen learning metrics).[8]

Algorithmic Fairness

Scientists and companies involved in designing and implementing AI solutions across various industries have recognized the importance of fairness and social responsibility in the software they create, embodied in the concept of fairness, accountability, transparency, and ethics in AI.[9] For commercial algorithms, there are regulatory considerations. For example, the Federal Trade Commission is empowered to prohibit "unfair or deceptive acts or practices in or affecting commerce," which include racially biased algorithms.[10] A bill introduced in Congress (the Algorithmic Accountability Act) would go further by directing the Federal Trade Commission to require impact assessment around privacy, security, bias, and fairness from companies developing automated decision-making systems.[11]

Multiple ways to measure algorithmic fairness have been developed.[12-15] Corbett-Davies and Goel[14] proposed 3 definitions for algorithmic fairness: 1) anticlassification for which protected features (eg, sex, race) are explicitly excluded from the model, 2) classification parity for which model performance is equal across groups organized by protected features, and 3) calibration for which model outcomes are independent of protected attributes. However, the impossibility theorem shows that it is not possible to simultaneously equalize false-positive rates, false-negative rates, and positive predictive values across protected classes while maintaining calibration or anticlassification fairness.[12] If only 1 fairness criterion can be achieved, clinical and ethical reasoning will be required to determine which one is appropriate.[16]

Techniques have been developed to explain poor fairness scores in AI algorithms. One approach applied the decomposition method of "additive features"[17] to quantitative fairness metrics[14,15] (eg, statistical parity).[18] By means of simulation data for features which were purposefully manipulated to result in poor statistical parity, this method identified features that were most responsible for fairness disparities in the outputs of AI algorithms.

AI Algorithms: What Could Possibly Go Wrong?

Prominent examples from outside of medicine can be instructive in understanding how particular problems in AI processes, namely lack of representative data sets and inadequate validation, may lead to unfair outcomes with the potential for serious consequences. A sparsity of training data from geographically diverse sources can lead to both representational harm (through bias amplification)[19] and allocational harm (from algorithms working less accurately).[20,21] A study of facial-recognition programs reported that while all software correctly identified white males (<1% error rate), the failure rate for women of color ranged from 21% to 35%.[22] A ProPublica[23] investigation of an AI algorithm that assessed the risk of recidivism showed that white defendants who re-offended were incorrectly classified as low risk almost twice as often as black offenders. In contrast, black defendants who did not re-offend were almost twice as likely as white defendants to be misclassified as at high risk of violent recidivism. These AI algorithms were inadvertently used to perpetuate institutional racism.[24] There are many theoretical reasons for the poor performance, with nonrepresentative training data being the most likely important factor.

Primum No Nocere

Embedded in the Hippocratic Oath for physicians is the concept of "primum no nocere" (first, do no harm), which applies to technological advances in medicine, including neuroradiology and AI implementation. AI models deployed in health care can lead to unintended unfair patient outcomes and can exacerbate underlying inequity. Not surprising, given massive interest in applying AI to medical imaging, examples of bias specific to neuroradiology are emerging. In a study that analyzed >80 articles that used AI on head CT examinations, >80% of data sets were found to be from single-center sources, which increases the susceptibility of the models to bias and increases model error rates.[25] The prevalence of brain lesions in the training and testing data sets did not match real world prevalence, which will likely overinflate the performances of models.[25] In a meta-analysis of AI articles on intracranial aneurysm detection, the authors concluded that most studies had a high risk of bias with poor generalizability, with only one-quarter of studies using an appropriate reference standard and only 6/43 studies using an external or hold-out test set.[26] They found low-level evidence for using these AI algorithms and that none of the studies specifically tested for the possibility of bias in algorithm development.[26] In a study that used AI models to detect both intracranial hemorrhage and large-vessel occlusion, the algorithm showed similar excellent performance in diverse populations regardless of scanning parameters and geographic distribution, suggesting that it is unbiased.[27] This study did not use independent data sets to test that assertion formally.[27]

In the neuroradiology literature, there are currently few studies assessing how bias may affect AI algorithms developed for routine clinical use. In 1 study, training cohort bias in ^{15}O-Water PET CBF calculation was evaluated.[28] The study showed that predictions in patients with cerebrovascular disease were poorer if only healthy controls were used for training models. However, predictions for healthy controls were unaffected if the models were trained only on patient data.[28] Training with data including healthy controls and patients with cerebrovascular disease yielded

the best performance.[28] From these neuroradiology examples, incorporating diverse patient characteristics that reflect target patient populations in the training and validation sets may be a reasonable strategy for mitigating bias.

In health care, there are many potential sources of bias such as age, sex, ethnicity, cultural, geographic, environmental, and socioeconomic status along with additional confounders such as disease prevalence and comorbidities.[1] It is easy to imagine that physical characteristics present in neuroradiology images could affect algorithm performance if not sufficiently represented in training sets. Inadequate sampling or matching disease prevalence could impact performance for different populations. Population-based studies could have inadequate inclusion of diverse data. In neuroradiology, additional sources of bias include heterogeneity of scanners, scanner parameters, acquisition protocols, and postprocessing algorithms.

Other ethical issues in AI use center on clinical deployment. Will the use of algorithms be equitable across hospital systems, or will only large, urban academic hospitals have access to state-of-the-art tools? Other considerations include whether the AI model will perform robustly across time. Medicine, health care practices, and devices are constantly evolving. Models need to be periodically validated on diverse populations and calibrated with data reflecting current clinical practices if they are expected to remain clinically relevant.[29] In medicine, interesting case studies that defy common medical knowledge can improve our understanding of disease and lead to practice changes. One such example is that of a patient who defied the odds of a severe motor vehicle crash to achieve complete recovery.[30] How to incorporate these outlier cases into AI algorithms is unclear. Overall, effective, fair, and ethical applications of AI to neuroradiology problems will require balancing competing demands across multiple domains (Online Supplemental Data).

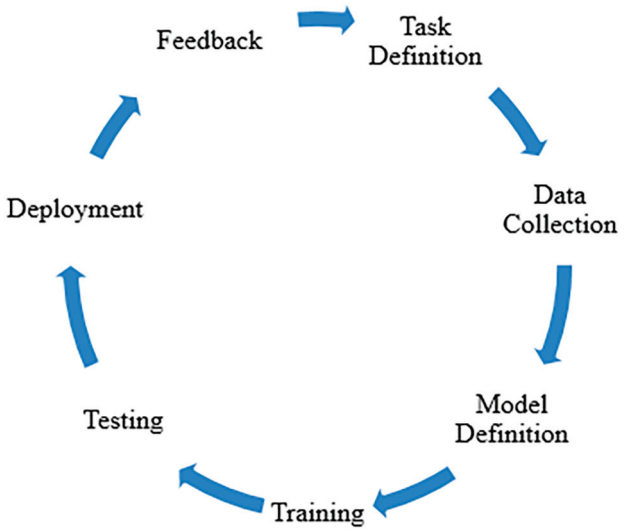

FIG 1. The AI algorithm development lifecycle.

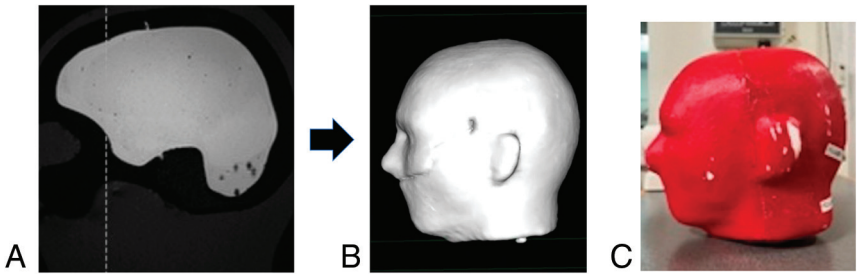

FIG 2. A 3D T1-weighted MR imaging scan (*A*) of an anthropomorphic phantom (*B*), which has no patient-identifying information. A 3D-rendered lifelike reconstruction (*C*) is possible, comparable with the original (photograph courtesy of Jacob C. Calkins, MGH Athinoula A. Martinos Center for Biomedical Imaging). The reconstruction was performed using open source software (Horos Version 3.3.6; Horos Project).

Mitigating Bias and Unfairness

Sources of bias in medical AI have been previously described.[16] In brief, there may be biases in the training data set construction, model training, clinician/patient interaction, and model deployment. It is incumbent on all stakeholders to do their part in mitigating bias and unfairness in the development, deployment, and use of AI models in neuroradiology.

Integration of Fairness, Accountability, Transparency, and Ethics Principles in the AI Cycle

Fairness, accountability, transparency, and ethics principles should be integrated[1,31,32] into the AI development lifecycle (Fig 1, adapted from Cramer et al[31] and the Online Supplemental Data). Diverse stakeholder involvement is critical for all stages. For task definition, one should clearly define the intended long-term effects of the task and model. One should define processes for discovering unintended biases at this stage. This outcome can be achieved by defining fairness requirements.

Data collection that is ethical and transparent and allows sufficient representation of protected groups should be ensured. One should check for biases in data sources. Many neuroradiologic AI applications require labeled data, eg, subarachnoid hemorrhage versus subdural hemorrhage. How and by whom are labels generated? Does it match the expected clinical deployment context? One should check for biases in how data are collected, which could lead to underrepresentation of underserved populations. Data collection should preserve privacy. For example, the collection of high-resolution images enables reconstruction of faces that can potentially be cross-linked to the patient's real identity through face-recognition software[33] as demonstrated in a reconstruction of data from an anthropomorphic phantom (Fig 2).[34] In a PET study for which CT and MR imaging data were collected for standard uptake value quantification, researchers showed that face-recognition software could match facial reconstructions from CT and MR imaging data to their actual face photographs with correct match rates ranging from 78% (CT) to 97%–98% (MR imaging), leading the researchers to advocate for the routine use of de-identification software.[35] When one uses de-identification software, the rates of recognition plummet to 5% for CT and 8% for MR

imaging, without impacting standard uptake value quantification.[35] A recent report used a novel de-identification software that deliberately distorted the ears, nose, and eyes that prevented facial recognition from CT and MR images,[36] which may be a viable solution for this privacy concern.

To address patient privacy concerns, many AI applications use synthetic data for training.[37] These synthetic data sets are typically produced using generative algorithms[38] and have the potential for promoting data-sharing (being unrestricted by regulatory agencies) and for the creation of diverse data sets.[37] However, the use of synthetic data can lead to nonrealistic scenarios[39] or inadvertently reinforce biases.[40,41]

For the model definition stage, model assumptions must be clearly defined, and potential biases, identified. Model architecture must be checked for introduction of biases and whether the cost function has unintended adverse effects.[42]

For the training stage, there are several online free resources to detect and mitigate bias[43] based on statistical definitions of fairness. Fairlearn[44] and AI Fairness 360[45] provide tools to detect and mitigate unfairness. Machine learning–Fairness-Gym takes a slightly different approach using simulation to evaluate the long-term fairness effects of learning agents in a specified environment.[46] The What-If Tool lets one visualize trained models to detect bias with minimal coding.[47] In addition, embedding learning methods that can debias AI models may help mitigate unfairness.[8] For example, Amini et al[8] proposed incorporating learning latent space structures for reweighting data during training to produce a less biased classifier.

For the testing stage, one should ensure that testing data have not leaked into the training data, match the expected deployed clinical context, and sufficiently represent the expected patient population. Potential issues with data-distribution discrepancies[48] can exacerbate unfairness. Variations among data sets can lead to biased learning of features from data sets collected from different sources (ie, domains) under different conditions.[49] Comparing differences between the source domain (where training data were collected) and the target domain (the test data for which the AI model will be used) may help explain any biases that are found. Many advanced domain-matching algorithms have been introduced to improve AI fairness by reducing the domain differences for cross-site data sets.[50,51]

In the deployment stage, continued surveillance of performance in terms of fairness and accuracy is needed. One should determine whether detected errors are one-off or systemic problems. There is no consensus yet on who bears this responsibility. Is it the end-users (radiologists/clinicians), the health care system/hospitals, or the vendors who make and sell the product? How will the algorithms be provided to the medical community? Will they be available equitably to diverse communities? One should ideally be able to explain how the trained AI model makes its decisions and predictions.

For the feedback stage, use and misuse of the system in the real world should be monitored and corrected in a transparent fashion. Fairness metrics[14,15] should be evaluated and then used to refine the model. Accountability for errors needs to be predefined.

Trust, Radiology, and AI: Guiding Principles

Neuroradiologists need to become educated and involved to ensure that AI is used appropriately in the diagnosis, management, and treatment of patients. For neuroradiologists to trust the use of AI in image interpretation, there needs to be greater transparency about the algorithm. Training data are foundationally critical to algorithm development, explaining why "good" data are so valuable. Therefore, trust-building for neuroradiology starts with the quality of data, its collection and management, its evaluation, the quality of its associated labels, and the protection of patient privacy. To many radiologists, the entire field of AI is opaque, where a "black box" takes images and spews out predictive analytics. For AI to gain widespread acceptance by patients and radiologists, everyone needs to comprehend how a particular trained AI model works.[52]

There are many unresolved issues around the development of AI in radiology. Large amounts of imaging data are needed, which are difficult to share among institutions because there is reticence to engage in data-sharing agreements when imaging data are financially valuable to industry. Additionally, there are data-use agreements and data-sharing agreements that stipulate noncommercial use. However, some might argue that excluding companies from developing products on the basis of de-identified, shared data is itself counterproductive and cannot be enforced in a meaningful way. Federated learning shows promise in disrupting this sharing-based landscape because it alleviates the need to share patient data by training models that gain knowledge from local data that are retained within the acquiring institution at all times.[53,54] However, security concerns[54] such as inferential attacks and "model poisoning" from corruption of the AI model and/or data from $\geq$1 site remain.[55,56] Unfairness in federated learning can be exacerbated by the challenge of simultaneously maintaining accuracy and privacy;[57] however, these potential limitations are being addressed.[58] Informed consent, ownership of data, privacy, and protection of data are major topics that remain in flux without clear best practice guidelines.[59]

For AI algorithm development in academic medical centers, new concepts are necessary. Should we assume that patients who enter a major academic medical center automatically opt-in to allow their anonymized imaging data to be used for research, including AI? Do patients need to explicitly opt-out in writing? If patient data are used to develop AI algorithms, should patients be financially compensated? One viewpoint is that "clinical data should be treated as a form of public good, to be used for the benefit of future patients" once its use for clinical treatment has ended.[60] These questions underscore the need to consider both the patient's and society's rights with respect to the use of such data.

The core principles of ethical conduct in patient research include beneficence (do only good), non-maleficence (do no harm), autonomy, and justice,[61] which must also guide AI development in neuroradiology. In the AI era of neuroradiology, there may be conflicts that evolve around how much decision-making is retained by the neuroradiologist and how much is willingly ceded to an AI algorithm. Floridi and Cowls[52] stated that the "autonomy of humans should be promoted and that the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be protected and re-established." This statement is precisely the major problem that occurred when pilots were unable to override an automated, erroneous AI-driven

navigation system to prevent nosedives, leading to plane crashes with significant loss of life.[62] Justice is conceptually implicit throughout AI development in neuroradiology from the data chosen to train the model to its validation so that no harm or unfairness occurs to certain groups of patients.[52]

Some researchers have articulated the need for a new bioethical consideration specifically to address algorithm development of AI in neuroradiology. Explicability can include explainability (how does it work?) and accountability (who is responsible for how it works?).[52,63,64] It is important that both patients and neuroradiologists understand how imaging tools such as AI algorithms are used to render decisions that impact their health and well-being, particularly around potentially life-saving decisions in which neuroradiology has a clear role. For example, a visual saliency map that delineates on images where the AI algorithm focused its attention to arrive at a prediction (ie, intracranial metastatic lesion on a brain MR imaging examination) would be useful to drive its acceptance by both clinicians and patients.[65] Neuroradiologists need to think like patients and adopt patient-centered practices when AI is deployed. Neuroradiologists should establish a practice to address real or perceived grievances for any unintended harm attributable to AI use.[52] Fear, ignorance, and misplaced anxiety around novel technology can derail the best of scientific intentions and advances, so we need to be prudent as we develop AI and encode bioethical principles into its development and deployment. Transparency can build trust,[66] with both code and data sets made publicly available whenever possible. However, for AI applications involving medical images, one must also balance the need for open science with patient privacy.

Ideally, neuroradiologists should be able to explain in lay language how data are used to build an AI tool, how the AI algorithm rendered a particular prediction, what that prediction means to patient care, and how accurate and reliable those predictions are.[64,65] This explanation will require education in AI from residency through fellowship and a process of life-long learning. The American Society of Neuroradiology (ASNR) convened an AI Task Force to make recommendations around education, training, and research in AI so that the ASNR maintains its primacy as a leader in this rapidly evolving field.

Suggestions for Neuroradiologists in AI

Academic neuroradiologists need to lead. It is our responsibility to establish the benchmarks for best practices in the clinical utility of AI in conjunction with our academic partners in imaging societies such as the American College of Radiology and the Radiological Society of North America, as well as federal stakeholders such as the National Institutes of Health, National Institute of Standards and Technology, the Advanced Research Projects Agency, and the Food and Drug Administration. Although guidelines have been published around the ethical implementation of AI code, more work is needed from all relevant stakeholders including neuroradiologists, clinicians, patients, institutions, and regulatory bodies so that consensus builds around best practices that include the new concepts of explainability and accountability while preserving patient privacy and protection against security breaches such as cyberattacks.[1,52,61,65] Quality assurance and quality improvement processes will be needed to detect potential biases in algorithms used in clinical care. Additional processes are needed to redress any perceived grievances and to quantify how AI affects patient outcomes.[67] In the Online Supplemental Data, across the AI development lifecycle, guidelines are listed in the form of essential questions that should be considered and asked around task definition, data collection, model definition, training and testing, and deployment and feedback, particularly when neuroradiologists are asked to evaluate clinical AI tools for their practices.

Summary

In a joint North American and European consortium white paper,[1] the authors made a recommendation that AI in radiology should "promote any use that helps individuals such as patients and providers and should block the use of radiology data and AI algorithms for irresponsible financial gains." Additionally, all AI algorithms must be informed by bioethical principles in which the benefits of AI outweigh the risks and minimize the potential for harm or bad outcomes and minimize the chances that AI will lead to greater health care inequity. Neuroradiologists need to participate fully in this transformative technology and set best practice standards for fair, ethical, and nonbiased deployment of AI in routine neuroimaging practice.

ACKNOWLEDGMENTS

We acknowledge Yilan Gu for assistance in literature research and Jacob Calkins for assistance with the phantom data.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Geis JR, Brady AP, Wu CC, et al. **Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement.** *Radiology* 2019;293:436–40 CrossRef Medline
2. "Ethics". *Merriam-Webster.com Dictionary*. Merriam-Webster, Inc. http://www.merriam-webster.com/dictionary/ethics. Accessed January 3, 2023
3. Panch T, Mattie H, Atun R. **Artificial intelligence and algorithmic bias: implications for health systems.** *J Glob Health* 2019;9:010318 CrossRef Medline
4. "Fairness". *Merriam-Webster.com Dictionary*. Merriam-Webster, Inc. http://www.merriam-webster.com/dictionary/fairness. Accessed January 3, 2023
5. Dwork C, Hardt M, Pitassi T, et al. **Fairness through awareness.** In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, Cambridge, Massachusetts. January 8–10, 2012 CrossRef
6. Crawford K. **The trouble with bias.** *NIPS* 2017;10 https://www.youtube.com/watch?v=fMym_BKWQzk. Accessed January 3, 2023
7. Abbasi M, Friedler SA, Scheidegger C, et al. **Fairness in representation: quantifying stereotyping as a representational harm**. In: *Proceedings of the 2019 SIAM International Conference on Data Mining:* SIAM, Calgary, Alberta, Canada, May 2–4, 2019
8. Amini A, Soleimany AP, Schwarting W, et al. **Uncovering and mitigating algorithmic bias through learned latent structure**. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* 2019;289–95 CrossRef
9. Association for Computing Machinery. **ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)**, Soeul, South Korea. June 21–24, 2022
10. Jillson E. **Aiming for truth, fairness, and equity in your company's use of AI**. *Business Blog Federal Trade Commision*. April 19, 2021. https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai. Accessed January 5, 2023

11. **Algorithmic Accountability Act of 2022.** HR 6580. 117th Congress, 2021–2022
12. Chouldechova A. **Fair prediction with disparate impact: a study of bias in recidivism prediction instruments.** *Big Data* 2017;5:153–63 CrossRef Medline
13. Kleinberg J, Mullainathan S, Raghavan M. **Inherent trade-offs in the fair determination of risk scores**. In: *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Berkeley, California. January 9–11, 2017. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik
14. Corbett-Davies S, Goel S. **The measure and mismeasure of fairness: a critical review of fair machine learning.** *arXiv* 2018 arXiv:180800023
15. Verma S, Rubin J. **Fairness definitions explained**. In: *Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*. IEEE Xplore 2018;1–7 CrossRef
16. Rajkomar A, Hardt M, Howell MD, et al. **Ensuring fairness in machine learning to advance health equity.** *Ann Intern Med* 2018;169:866–72 CrossRef Medline
17. Lundberg SM, Lee SI. **A unified approach to interpreting model predictions.** *arXiv* 1705.07874 May 22, 2017
18. Lundberg SM. **Explaining quantitative measures of fairness**. In: *Proceedings of the Fair & Responsible AI Workshop. CHI '20: CHI Conference on Human Factors in Computing Systems*, Honolulu, Hawaii. April 25–30, 2020
19. Zhao J, Wang T, Yatskar M, et al. **Men also like shopping: reducing gender bias amplification using corpus-level constraints**. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. September 7–11, 2017
20. Zou J, Schiebinger L. **AI can be sexist and racist: it's time to make it fair**. *Nature* 2018;559:324–26 CrossRef Medline
21. Sweeney L. **Discrimination in online ad delivery.** *Communications of the ACM* 2013;56:44–54 CrossRef
22. Buolamwini J, Gebru T. **Gender shades: intersectional accuracy disparities in commercial gender classification**. In: *Proceedings of the Conference on Fairness, Accountability and Transparency*, New York, New York. February 23–24, 2018;77–91
23. Angwin J, Larson J, Matta S, et al. **Machine bias: there's software used across the country to predict future criminals. and it's biased against Blacks**. *ProPublica* May 23, 2016 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed January 6, 2023
24. Benjamin R. **The New Jim code: reimagining the default settings of technology & society**. *Race and technology: A Research Lecture Series*. May 1, 2021 to June 30, 2022. Virtual https://www.youtube.com/watch?v=aMuD_lAy2zQ. Accessed July 24, 2022
25. Gunzer F, Jantscher M, Hassler EM, et al. **Reproducibility of artificial intelligence models in computed tomography of the head: a quantitative analysis.** *Insights Imaging* 2022;13:173 CrossRef Medline
26. Din M, Agarwal S, Grzeda M, et al. **Detection of cerebral aneurysms using artificial intelligence: a systematic review and meta-analysis.** *J Neurointerv Surg* 2023;15:262–71 CrossRef Medline
27. McLouth J, Elstrott S, Chaibi Y, et al. **Validation of a deep learning tool in the detection of intracranial hemorrhage and large vessel occlusion.** *Front Neurol* 2021;12:656112 CrossRef Medline
28. Guo J, Gong E, Fan AP, et al. **Predicting (15)O-Water PET cerebral blood flow maps from multi-contrast MRI using a deep convolutional neural network with evaluation of training cohort bias.** *J Cereb Blood Flow Metab* 2020;40:2240–53 CrossRef Medline
29. Gao MM, Wang J, Saposnik G. **The art and science of stroke outcome prognostication.** *Stroke* 2020;51:1358–60 Medline
30. Edlow BL, Giacino JT, Hirschberg RE, et al. **Unexpected recovery of function after severe traumatic brain injury: the limits of early neuroimaging-based outcome prediction.** *Neurocrit Care* 2013;19:364–75 CrossRef Medline
31. Cramer H, Holstein K, Vaughan J, et al. **Challenges of incorporating algorithmic fairness into industry practice. ACM FAT* 2019 Translation Tutorial**. February 22, 2019 https://www.youtube.com/watch?v=UicKZv93SOY, Accessed July 24, 2022
32. Barocas S, Hardt M, Narayanan A. **Fairness in machine learning.** *Nips Tutorial* 2017;1:76–101
33. Schwarz CG, Kremers WK, Therneau TM, et al. **Identification of anonymous MRI research participants with face-recognition software.** *N Engl J Med* 2019;381:1684–86 CrossRef Medline
34. Athinoula A. **Martinos Center for Biomedical Imaging: Martinos Center Anthropomorphic Phantoms.** https://phantoms.martinos.org/Main_Page. Accessed May 21, 2023
35. Schwarz CG, Kremers WK, Lowe VJ, et al; Alzheimer's Disease Neuroimaging Initiative. **Face recognition from research brain PET: an unexpected PET problem**. *Neuroimage* 2022;258:119357 CrossRef Medline
36. Jeong YU, Yoo S, Kim YH, et al. **De-identification of facial features in magnetic resonance images: software development using deep learning technology.** *J Med Internet Res* 2020;22:e22739 CrossRef Medline
37. Arora A, Arora A. **Synthetic patient data in health care: a widening legal loophole.** *Lancet* 2022;399:1601–02 CrossRef Medline
38. Goodfellow I, Pouget-Abadie J, Mirza M, et al. **Generative adversarial nets; NIPS'14**: In: *Proceedings of the 27th International Conference on Neural Information Processing*, Montreal, Quebec, Canada. December 8–13, 2014
39. Ross C, Swetlitz I. **IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show**. *STAT+*. July 25,2018 https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf. Accessed July 19, 2022
40. Onder O, Yarasir Y, Azizova A, et al. **Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review.** *Insights Imaging* 2021;12:51 CrossRef Medline
41. Choi K, Grover A, Singh T, et al. In: *Proceedings of the International Conference on Machine Learning*, Virtual. July 12–18, 2020;1887–98
42. Obermeyer Z, Powers B, Vogeli C, et al. **Dissecting racial bias in an algorithm used to manage the health of populations.** *Science* 2019;366:447–53 CrossRef Medline
43. Agarwal A, Beygelzimer A, Dudík M, et al. **A reductions approach to fair classification**. In: *Proccedings of the International Conference on Machine Learning*, Stockholm, Sweden. July 10–15, 2018;60–69
44. Fairlearn. **Improve fairness of AI systems.** https://fairlearn.org/. Accessed July 24, 2022
45. IBM Research. **AI Fairness 360.** https://ibm.com/opensource/open/projects/ai-fairness-360. Accessed July 24, 2022
46. Google. **The ML Fairness Gym.** https://github.com/google/ml-fairness-gym. Accessed July 24, 2022
47. **What-If Tool.** https://pair-code.github.io/what-if-tool/. Accessed July 24, 2022
48. Koh PW, Sagawa S, Marklund H, et al. **WILDS: a benchmark of in-the-wild distribution shifts**. In: Marina M, Tong Z, eds. In: *Proceedings of the 38th International Conference on Machine Learning* Virtual. July 18-21, 2021:5637–64
49. Yan W, Wang Y, Xia M, et al. **Edge-guided output adaptor: highly efficient adaptation module for cross-vendor medical image segmentation.** *IEEE Signal Process Lett* 2019;26:1593–97 CrossRef
50. Schumann C, Wang X, Beutel A, et al. **Transfer of machine learning fairness across domains.** *arXiv* 2019 arXiv:190609688
51. Joshi N, Burlina P. **AI fairness via domain adaptation.** *arXiv* 2021 arXiv:2104.01109
52. Floridi L, Cowls J. **A united framework of five principles for AI in society.** *Harvard Data Science Review* 2019;1 CrossRef
53. Sheller MJ, Edwards B, Reina GA, et al. **Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data.** *Sci Rep* 2020;10:12598 CrossRef Medline
54. Li T, Sahu AK, Talwalkar A, et al. **Federated learning: challenges, methods, and future directions.** *IEEE Signal Processing Magazine* 2020;37:50–60 CrossRef
55. Wang L, Xu S, Wang X, et al. **Eavesdrop the composition proportion of training labels in federated learning.** *arXiv* 2019 arXiv:1910.06044

56. Pejó B, Biczók G. **Quality inference in federated learning with secure aggregation**. *arXiv* 2020 arXiv:2007.06236
57. Abay A, Zhou Y, Baracaldo N, et al. **Mitigating bias in federated learning.** *arXiv* 2020 arXiv:2012.02447
58. Gu X, Tianqing Z, Li J, et al. **Privacy, accuracy, and model fairness trade-offs in federated learning.** *Computers & Security* 2022;122:102907 CrossRef
59. Mittelstadt BD, Floridi L. **The ethics of big data: current and foreseeable issues in biomedical contexts.** *Sci Eng Ethics* 2016;22:303–41 CrossRef Medline
60. Larson DB, Magnus DC, Lungren MP, et al. **Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework.** *Radiology* 2020;295:675–82 CrossRef Medline
61. Varkey B. **Principles of clinical ethics and their application to practice.** *Med Princ Pract* 2021;30:17–28 CrossRef Medline
62. Mongan J, Kohli M. **Artificial intelligence and human life: five lessons for radiology from the 737 MAX Disasters.** *Radiol Artif Intell* 2020;2:e190111 CrossRef Medline
63. Floridi L, Taddeo M. **What is data ethics?** *Philos Trans A Math Phys Eng Sci* 2016;374:20160360 CrossRef
64. Floridi L, Cowls J, Beltrametti M, et al. **AI4People: an ethical framework for a good AI society—opportunities, risks, principles, and recommendations**. *Minds Mach (Dordr)* 2018;28:689–707 CrossRef Medline
65. Gilpin LH, Bau D, Yuan BZ, et al. **Explaining explanations: an overview of interpretability of machine learning**. In: *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAS)*, Turin, Italy. October 1–3, 2018;80–89 CrossRef
66. Haibe-Kains B, Adam GA, Hosny A, et al; Massive Analysis Quality Control (MAQC) Society Board of Directors. **Transparency and reproducibility in artificial intelligence.** *Nature* 2020;586:E14–16 CrossRef Medline
67. Kingston JKC. **Artificial intelligence and legal liability**. In: Bramer M, Petridis M, eds. *Research and Development in Intelligent Systems XXXIII*. Springer-Verlag International Publishing 2016:269–79