Review Article

Healthc Inform Res. 2023 October;29(4):315-322.
https://doi.org/10.4258/hir.2023.29.4.315
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Requirements for Trustworthy Artificial Intelligence and its Application in Healthcare

Myeongju Kim[1], Hyoju Sohn[1], Sookyung Choi[1], Sejoong Kim[1,2,3]

[1]Healthcare Innovation Park, Center for Artificial Intelligence in Healthcare, Seoul National University Bundang Hospital, Seongnam, Korea
[2]Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Korea
[3]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Korea

Objectives: Artificial intelligence (AI) technologies are developing very rapidly in the medical field, but have yet to be actively used in actual clinical settings. Ensuring reliability is essential to disseminating technologies, necessitating a wide range of research and subsequent social consensus on requirements for trustworthy AI. **Methods:** This review divided the requirements for trustworthy medical AI into explainability, fairness, privacy protection, and robustness, investigated research trends in the literature on AI in healthcare, and explored the criteria for trustworthy AI in the medical field. **Results:** Explainability provides a basis for determining whether healthcare providers would refer to the output of an AI model, which requires the further development of explainable AI technology, evaluation methods, and user interfaces. For AI fairness, the primary task is to identify evaluation metrics optimized for the medical field. As for privacy and robustness, further development of technologies is needed, especially in defending training data or AI algorithms against adversarial attacks. **Conclusions:** In the future, detailed standards need to be established according to the issues that medical AI would solve or the clinical field where medical AI would be used. Furthermore, these criteria should be reflected in AI-related regulations, such as AI development guidelines and approval processes for medical devices.

Keywords: Artificial Intelligence, Machine Learning, Healthcare Disparity, Trust, Guideline

Submitted: March 30, 2023
Revised: 1st, September 30, 2023; 2nd, October 6, 2023
Accepted: October 13, 2023

Corresponding Author
Sejoong Kim
Center for Artificial Intelligence in Healthcare, Seoul National University Bundang Hospital, 82 Gumi-ro 173beon-gil, Bundang-gu, Seongnam 13620, Korea. Tel: +82-31-787-7051, E-mail: sejoong@snubh.org (https://orcid.org/0000-0002-7238-9962)

I. Introduction

The field of artificial intelligence (AI) is growing rapidly around the world. The application of AI in healthcare is receiving enormous attention and is actively being studied [1]. As a field where people's lives and well-being are at stake, there are many opportunities to apply AI technology in healthcare due to its vital importance and the large-scale accumulation of computerized medical records [2]. The introduction of AI has made it possible to solve problems that were impossible for traditional technology and quickly perform procedures that used to take a long time. For instance, cerebral infarction can be detected from a head and neck X-ray alone [3], and the occurrence of acute renal injury can be predicted 24 hours in advance by analyzing electronic medical records [3,4]. However, the application and diffusion of research and development outputs, such as medical AI solutions, continue to be challenging. Many AI technologies are not utilized in the medical field, even after being certified by the Ministry of Food and Drug Safety of Republic of Korea. Issues such as insurance reimbursement are also obstacles, but it is necessary to secure reliability, or trustworthiness, before medical AI technology can spread.

Trustworthy AI is currently a hot topic. The issue of whether an AI's judgment and decision-making process can be trusted has been brought to the fore following incidents involving AI-based chatbots such as Tay and Iruda [5]. Microsoft's chatbot Tay spoke as if it held racist, sexist, or far-right political ideologies, emphasizing the challenge of developing AI technology with ethical values [6]. In Korea, Iruda, which was based on the concept of a female college student, became hugely popular, with 400 users within a month of its launch. However, the service was terminated after it caused social controversy by leaking personal information and making comments about sexual minorities [7]. As such, if an AI's output is untrustworthy and causes social problems, people will be reluctant to utilize it, even if significant time and effort have gone into its development.

The reliability of AI in the medical and healthcare fields needs to be discussed more carefully. In the medical field, AI technology is being incorporated into clinical decision support systems (CDSS) to support critical medical tasks such as diagnosis and treatment planning [8]. While the scope of use is limited to assisting healthcare providers, the consequences of misuse can be severe in fields where lives are at stake. For instance, frequent false alarms in settings with urgent patients may cause healthcare providers to become fatigued [9].

Just as we carefully select trustworthy people, organizations, and companies before entrusting important tasks to them, it is necessary to examine the reliability of medical AI technology using detailed criteria. If medical AI that meets these criteria and has been certified for reliability is preferentially disseminated, it may become possible to maximize the social benefits. This requires the establishment and institutionalization of trustworthy AI standards specialized for the medical and healthcare fields. In this study, we review the requirements for trustworthy AI and examine the current status of its application and related policy initiatives in healthcare.

II. Overview of Requirements for Trustworthy AI

Despite heated social discussions about trustworthy AI, the requirements for it have not yet been clearly established and are discussed inconsistently by different institutions and organizations. The Ministry of Science and Information & Communications Technology in South Korea presents safety, explainability, transparency, robustness, and fairness as core elements of AI reliability. Worldwide, the Fairness, Accountability, and Transparency in Machine Learning (FAT-ML) principles include responsibility, explainability, accuracy, auditability, and fairness [8]. In this report, we will address explainability, fairness, privacy, and robustness among the various requirements under discussion (Table 1).

1. Explainability

1) Explainability and trustworthy medical AI

Explainability refers to presenting the underlying logic of the judgment, decisions, or outputs of the AI in a way that humans can understand; similar terms include interpretability and transparency [10]. In the medical field, this concept provides a basis for healthcare providers to decide whether to refer to the output of the algorithm in their practices [11]. Only after confirming the relationship between the judgment logic of AI algorithms and its output can one verify whether the output is due to a clinically irrelevant feature or a simple error. Especially when conflicts occur between the judgment of healthcare providers and that of AI, understanding the rationales of algorithms' conclusions is crucial for making a final decision. Furthermore, if the conclusion of the AI algorithm is correct, the healthcare provider can learn the judgment process of AI for their professional growth. Otherwise, if the conclusion of the healthcare provider is correct, it can provide an opportunity to identify how to retrain the algo-

Table 1. Requirements for trustworthy artificial intelligence (AI)

Concept	Description
Explainability	The process by which the AI model derives its output can be presented so that users can understand it.
Fairness	The output of the AI model can be presented regardless of specific protected variables.
Privacy	It is possible to avoid problems with personal data that may occur during the development of the AI.
Robustness	The AI model can defend against external attacks and maintain its function and proper performance.

rithm [12].

2) Need for research on explainable AI in deep learning

Much of the research in explainable AI (XAI) has been focused on machine learning algorithms that learn numerical data, and significant progress has been made, such as the development of methods to determine the extent and direction in which a feature affects an algorithm's output [13]. However, only a few of them, such as saliency methods, have been utilized in deep learning algorithms. In many cases, there is a trade-off between predictive power and the ability to explain the judgment logic of the algorithm [14]. As input features become more highly abstracted through deep and complex networks, it becomes challenging to solve the algorithmic black box [10]. Since most of the algorithms that utilize unstructured data are built upon deep learning networks, there have been numerous attempts to develop new XAI methodologies that can be applied to deep learning algorithms. Recently, researchers have been trying to explain the black box by generating adversarial examples. Chang et al. [15] demonstrated that the judgment of AI algorithms that can diagnose ophthalmic disease using fundus images can be explained by reviewing how pathological characteristics of the ophthalmic disease are newly added or deleted when generating adversarial examples. Several methods that can generate minimally perturbating adversarial examples have been suggested as tools to ensure the reliability of AI [16-18].

3) Assessment and reporting of explainability

Because there are no clear criteria or established methods, humans are currently conducting evaluations, but a limitation of this approach is human subjectivity and the possibility of different interpretations. To date, several studies have evaluated the degree of explainability by survey scoring [15,19]. Once explainability is achieved, the decision logic of an algorithm must be provided transparently to the user, which requires an easy-to-understand, user-friendly interface. For instance, Predict is a platform that explains model output by providing users with reports describing different feature representations and judgment logic [20]. However, shortcomings have been identified regarding reproducibility, such as suggesting different rationales for similar cases. As such, procedural and technical improvements are needed to quantify the reliability of an algorithm.

2. Fairness

1) Fairness and trustworthy medical AI

Fairness refers to the degree to which the output of an AI model is independent of protected attributes. Protected or sensitive attributes include gender, disability, age, religion, marital status, and educational background, and these attributes are determined according to laws or moral values [19]. It is necessary to confirm that medical AI is fairly developed and applied to ensure that it promotes the welfare of all individuals and groups to the same level. AI technology contributes to equality by increasing access to medical services for marginalized individuals or groups; however, if inequalities are incorporated into the model in the training step, the AI program may render unfair decisions that disadvantage marginalized groups. XAI technologies can be used to address these issues and make it easier to detect model bias by checking whether the algorithm's output was substantially influenced by protected attributes [21,22].

Algorithmic unfairness is most likely to occur when the training data are not representative due to a biased dataset. As an example of racial bias, Black people may not be able to afford expensive tests and are less likely to be included in a hospital's database. The unfairness of the AI model can be caused by algorithmic bias and algorithmic unfairness. Algorithmic bias refers to inequalities that arise during model design, data collection, and sampling. In contrast, algorithmic unfairness refers to inequalities that arise as the learning algorithm learns to be unfair, regardless of biases in the data [23].

2) Types and measures of fairness

Fairness is a concept previously discussed in the fields of ethics and law. In the traditional discourse [24,25], it is divided into equal opportunity fairness, procedural fairness, and consequential fairness [26,27]. The concept of fairness implies multiple dimensions that are challenging to describe and distinguish clearly. In the field of AI, attempts have been made to develop mathematical definitions that can reflect the multidimensionality of fairness, and several types of evaluation metrics have been proposed [28-31].

Fairness currently being discussed in the AI field can be primarily categorized into group fairness, individual fairness, and counterfactual fairness (Table 2). Group fairness is the most actively discussed of these concepts, referring to the idea that different groups should be treated similarly or equally [32]. For example, suppose an algorithm for detecting skin cancer from pathological images shows excellent performance for White people but relatively low

Table 2. Key concepts of model fairness

Concept	Description
Group fairness	Different groups that are not separated by protected variables should be treated similarly.
Individual fairness	Similar individuals should be treated similarly by excluding protected variables.
Counterfactual fairness	A causal relationship inference graph is created, and even if protected variables are reversed in this graph, protection is provided.

Table 3. Metrics of group fairness

Concept	Description
Demographic disparity	The fairness criterion that is met when the result of classification is not dependent on protected variables.
Equalized odds	The fairness criterion that is met when all proportions of classification results are not dependent on the protected variables of the group.
Equality of opportunity	The criterion that is satisfied when the proportion of positive characteristics (eligibility for employment or a financial loan, etc.) as a result of classification is not dependent on the protected variables of the group.
Predictive parity	The criterion that is satisfied when the classification result shows the same positive predictive values for the non-protected group and the subjects of the protected group.

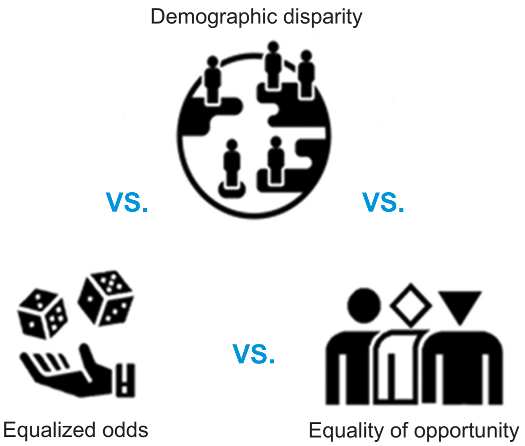

Figure 1. Conflicts between measures of collective fairness.

performance for Black people. In that case, it can be seen as a failure to ensure group fairness [33]. More than 20 metrics have been developed to assess group fairness, including demographic parity, equalized odds, equal opportunity, and predictive parity (Table 3). The metrics of group fairness conflict with each other mathematically, making it difficult to satisfy them simultaneously (Figure 1). It has been statistically proven that scoring well on all three major metrics of group fairness is impossible [34]. Paulus et al. [23] demonstrated that the higher the level of opportunity fairness, the harder it is for an algorithm to obtain equalized odds.

Individual fairness is motivated by the principle that individuals with similar characteristics should be treated similarly and can be measured by calculating distances between individuals [35]. The problem with individual fairness is that it leaves room for evaluator subjectivity in the methodology used to calculate the distance between individuals.

Both group fairness and individual fairness have limitations in that it is difficult to assess the exact level of fairness when biases exist during data labeling [36]. One disadvantage of counterfactual fairness is the difficulty of inferring causal relationships, which limits its use. Summarizing this shortcoming of various fairness metrics, it takes work to select one optimal fairness measure for multiple tasks or fields. Therefore, it is necessary to select and apply appropriate fairness metrics according to the dataset or settings in which the AI algorithm is being used.

3) Fairness in healthcare and medical research

In the healthcare and medical fields, most AI studies that have evaluated fairness have focused on group fairness [37]. Some research papers present the results of fairness analyses of the model as part of additional analysis and report the effects of debiasing techniques. Garriga et al. [38] predicted the occurrence of a mental crisis within one month using the electronic medical records of psychiatric ward inpatients and further analyzed algorithmic bias and fairness. The study found bias in the data sampling stage, which showed a low level of demographic equity with a high proportion of Black and mixed-race individuals and modest differences in the performance of the algorithms by race and disability [38-42]. A study by Park et al. [43] evaluated the fairness of a machine learning algorithm for predicting postpartum depression by demographic equity and equality of opportunity metrics and applied techniques such as reweighting and regularization to improve the level of fairness.

3. Privacy

1) Privacy protection and trustworthy medical AI

The issue of privacy protection has long been emphasized throughout the medical field [44]. Medical records include information that can identify individuals, such as social security numbers and birth data, and sensitive information that may compromise personal privacy, such as medical history. Introducing medical AI technology may threaten the protection of privacy afforded by medical data. The training dataset can be reconstructed from only the final output and AI model [45]. When medical data are accumulated at scale in the process of building training datasets, a leak of personal data may have a significant social impact [46].

2) International privacy protection principles

Privacy by Design (PbD) is an internationally accepted privacy principle concerning protecting personal data throughout data collection and utilization. First and foremost, PbD should be adhered to in AI development, and the principles are as follows. PbD states that it is necessary to establish a basis for data collection, such as obtaining patient consent under a specific protocol. Data must be anonymized, pseudonymized, and de-identified, and used only for permitted purposes in accordance with the interests of patients. In addition, it is necessary to analyze the risks of privacy violations in advance and to prepare countermeasures [47].

3) Latest developments in privacy protection

The Korea National IT Industry Promotion Agency has devised a self-inspection checklist for reference during AI development by those who handle personal information, such as AI developers and operators [48]. The checklist has 16 inspection items and 54 confirmation items, emphasizing the need to continue assessing the impact of AI algorithms on personal information. In addition, privacy-enhancing technologies are being developed that implement privacy protection principles, such as minimizing the use of personal information and preventing leakage. Homomorphic encryption technology, which is used for data security, has gained attention because it allows encrypted data to be used for analysis without decryption. Suppose medical data containing sensitive information, such as personal details, are accumulated in one place. In that case, there is a high probability that data will be stolen in one fell swoop if exposed to a person or group who intends malicious use, such as a hacker. Recently, federated learning has emerged as a way to train models separately and combine them, rather than centralizing data, which requires data transfer from one institution to another [49].

4. Robustness

1) Robust and trustworthy medical AI

Robustness means that an algorithm maintains a certain level of performance despite circumstantial changes that may occur during real-world use. Algorithmic robustness must be confirmed at an individual application level, and it is necessary to take measures from the design stage to ensure that performance does not fluctuate in response to changes in the clinical environment, such as user activity, data sets, hardware, or hostile attacks. Recent research on robustness has focused on preventing hostile attacks, ranging from inserting or extracting data by entering AI models to the malicious use of algorithms by third parties [48,49].

2) Adversarial attacks and robustness

An adversarial attack refers to intentionally manipulating data at a level unrecognizable to humans, such that the algorithm outputs wrong results [50]. While this is not currently a socially problematic situation, it has the potential to disrupt the performance of the algorithm. Finlayson et al. [52] showed that an AI algorithm to detect skin cancer failed to diagnose cancer if hostile noise was added to an existing image. Taghanaki et al. [53] demonstrated the vulnerability of an AI algorithm for pneumonia classification based on chest X-ray images by generating adversarial attacks. In this way, altering even a tiny portion of the data can cause the algorithm to produce the opposite result. Therefore, algorithmic defenses against adversarial attacks must be established before the widespread application of medical AI.

III. Discussion

Extensive research and social consensus on the requirements of explainability, fairness, privacy protection, and robustness are needed for trustworthy medical AI to be deployed and widely used in society. Each clinical setting in which AI is applied will have optimized requirements and standards that must be met, and these requirements and standards must be updated on an ongoing basis.

Depending on the tasks that medical AI solves, such as diagnosis, prognosis prediction, and establishment of treatment plans, optimized requirements may be established. In the future, it will be necessary to establish evaluation standards that can compare the explainability of AI models. Fairness measures optimized for the healthcare and medical fields should also be identified.

In addition to the requirements of trustworthy AI covered in this report, several aspects need to be supplemented in the medical and healthcare field. For medical AI to be trusted by users, performance (e.g., accuracy) must be guaranteed above a certain level, the user interface must be easy to use, and an easy-to-read manual with a standardized format must be produced. The AI-based program should be integrated well into the workflows of existing clinical procedures. In addition, regulations must be established that stipulate who is responsible in the event of an incident or accident caused by medical AI: designers, researchers, medical staff, or patients [45].

The current guidelines for trustworthy AI are designed for the entire domain of AI research. Therefore, it will be necessary to establish development guidelines for AI-based medical devices that account for the specificities of the medical and healthcare fields. Another possible direction would be to include requirements for trustworthy AI in the approval guidelines on AI medical devices. Compliance with requirements could also be made mandatory for high-risk AI technology. Furthermore, evaluating insurance claim reimbursements for medical device use could incorporate assessing the requirements for trustworthy AI.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research is partly based on the KOSMI Issue Report (2022), which was supported by the Ministry of Health and Welfare, Republic of Korea. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics), and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2022R1F1A1060787).

ORCID

Myeongju Kim (https://orcid.org/0000-0003-0543-3019)
Hyoju Sohn (https://orcid.org/0009-0002-8332-8564)
Sookyung Choi (https://orcid.org/0009-0000-3391-9222)
Sejoong Kim (https://orcid.org/0000-0002-7238-9962)

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-43. https://doi.org/10.1136/svn-2017-000101
2. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94-8. https://doi.org/10.7861/futurehosp.6-2-94
3. Jeong HG, Kim BJ, Kim T, Kang J, Kim JY, Kim J, et al. Classification of cardioembolic stroke based on a deep neural network using chest radiographs. EBioMedicine 2021;69:103466. https://doi.org/10.1016/j.ebiom.2021.103466
4. Kim K, Yang H, Yi J, Son HE, Ryu JY, Kim YC, et al. Real-time clinical decision support based on recurrent neural networks for in-hospital acute kidney injury: external validation and model interpretation. J Med Internet Res 2021;23(4):e24120. https://doi.org/10.2196/24120
5. Tidjon LN, Khomh F. Never trust, always verify: a roadmap for Trustworthy AI? [Internet]. Ithaca (NY): arXiv.org; 2022 [cited at 2023 Oct 31]. Available from: https://arxiv.org/abs/2206.11981.
6. Neff G. Talking to bots: symbiotic agency and the case of Tay. Int J Commun 2016;10:4915-31.
7. Choi SS, Hong AR. Identifying issue changes of AI Chatbot 'Iruda' case and its implications. Electron Telecommun Trends 2021;36(2):93-101. https://doi.org/10.22648/ETRI.2021.J.360210
8. Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011;18(3):327-34. https://doi.org/10.1136/amiajnl-2011-000094
9. Graham KC, Cvach M. Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms. Am J Crit Care 2010;19(1):28-34. https://doi.org/10.4037/ajcc2010651
10. Arrieta AB, Diaz-Rodriguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82-115. https://doi.org/10.1016/j.inffus.2019.12.012

11. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci 2021;11(11):5088. https://doi.org/10.3390/app11115088
12. Han HJ. Trends in explainable artificial intelligence (XAI) research in the medical/healthcare domain [Internet]. Pohang, Korea: BRIC View; 2021 [cited at 2023 Oct 31]. Available from: https://www.ibric.org/myboard/read.php?Board=report&id=3751.
13. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): a survey [Internet]. Ithaca (NY): arXiv.org; 2020 [cited at 2023 Oct 31]. Available from: https://arxiv.org/abs/2006.11371.
14. van der Veer SN, Riste L, Cheraghi-Sohi S, Phipps DL, Tully MP, Bozentko K, et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. J Am Med Inform Assoc 2021;28(10):2128-38. https://doi.org/10.1093/jamia/ocab127
15. Chang J, Lee J, Ha A, Han YS, Bak E, Choi S, et al. Explaining the rationale of deep learning glaucoma decisions with adversarial examples. Ophthalmology 2021;128(1):78-88. https://doi.org/10.1016/j.ophtha.2020.06.036
16. Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27-30; Las Vegas, NV. p. 2574-82. https://doi.org/10.1109/CVPR.2016.282
17. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [Internet]. Ithaca (NY): arXiv.org; 2014 [cited at 2023 Oct 31]. Available from: https://arxiv.org/abs/1412.6572.
18. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. Proceedings of 2016 IEEE European Symposium on Security and Privacy (EuroS&P); 2016 Mar 21-24; Saarbruecken, Germany. p. 372-87. https://doi.org/10.1109/EuroSP.2016.36
19. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy 2020;23(1):18. https://doi.org/10.3390/e23010018
20. Chromik M, Butz A. Human-XAI interaction: a review and design principles for explanation user interfaces. In: Ardito C, Lanzilotti R, Malizia A, et al. Human-computer interaction–INTERACT 2021. Cham, Switzerland: Springer; 2021. p. 619-40. https://doi.org/10.1007/978-3-030-85616-8_36
21. Grgic-Hlaca N, Lima G, Weller A, Redmiles EM. Dimensions of diversity in human perceptions of algorithmic fairness [Internet]. Ithaca (NY): arXiv.org; 2022 [cited at 2023 Oct 31]. Available from: https://arxiv.org/abs/2005.00808.
22. Baniecki H, Kretowicz W, Piatyszek P, Wisniewski J, Biecek P. Dalex: responsible machine learning with interactive explainability and fairness in Python. J Mach Learn Res 2021;22(1):9759-65.
23. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med 2020;3:99. https://doi.org/10.1038/s41746-020-0304-9
24. Franck TM. Fairness in international law and institutions. Oxford, UK: Oxford University Press; 1998. https://doi.org/10.1093/acprof:oso/9780198267850.001.0001
25. Rawls J. Justice as fairness: political not metaphysical. In: Corlett JA, editor. Equality and liberty: analyzing Rawls and Nozick. London, UK: Palgrave Macmillan; 1991. p. 145-73. https://doi.org/10.1007/978-1-349-21763-2_10
26. Vidmar N. The origins and consequences of procedural fairness. Law Soc Inq 1990;15(4):877-92. https://doi.org/10.1111/j.1747-4469.1990.tb00607.x
27. Park HM, Kim SH. The multi-dimensionality of theories of justice. Soc Theory 2015;27(2):219-60. https://doi.org/10.17209/st.2015.11.27.219
28. Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, Wang F. Algorithmic fairness in computational medicine. EBioMedicine 2022;84:104250. https://doi.org/10.1016/j.ebiom.2022.104250
29. Awasthi P, Beutel A, Kleindessner M, Morgenstern J, Wang X. Evaluating fairness of machine learning models under uncertain and incomplete information. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 Mar 3-10; Virtual Event, Canada. p. 206-14. https://doi.org/10.1145/3442188.3445884
30. Hinnefeld JH, Cooman P, Mammo N, Deese R. Evaluating fairness metrics in the presence of dataset bias [Internet]. Ithaca (NY): arXiv.org; 2018 [cited at 2023 Oct 31]. Available from: https://arxiv.org/abs/1809.09245
31. Madaio M, Egede L, Subramonyam H, Wortman Vaughan J, Wallach H. Assessing the fairness of AI systems: ai practitioners' processes, challenges, and needs for support. Proc ACM Hum Comput Interact 2022;6(CSCW1):1-26. https://doi.org/10.1145/3512899
32. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Adv Neural Inf Process Syst 2016;29:3315-23.

33. Srivastava M, Heidari H, Krause A. Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019 Aug 4-8; Anchorage, AK. p. 2459-68. https://doi.org/10.1145/3292500.3330664
34. Saravanakumar KK. The impossibility theorem of machine fairness: a causal perspective [Internet]. Ithaca (NY): arXiv.org; 2020 [cited at 2023 Oct 31]. Available from: https://arxiv.org/abs/2007.06024.
35. Dwork C, Ilvento C. Fairness under composition [Internet]. Ithaca (NY): arXiv.org; 2018 [cited at 2023 Oct 31]. Available from: https://arxiv.org/abs/1806.06122.
36. Binns R. On the apparent conflict between individual and group fairness. Proceedings of the 2020 Conference on Fairness, Accountability, And Transparency; 2020 Jan 27-30; Barcelona, Spain. pp. 514-24. https://doi.org/10.1145/3351095.3372864
37. Meng C, Trinh L, Xu N, Enouen J, Liu Y. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Sci Rep 2022;12(1):7166. https://doi.org/10.1038/s41598-022-11012-2
38. Garriga R, Mas J, Abraha S, Nolan J, Harrison O, Tadros G, Matic A. Machine learning model to predict mental health crises from electronic health records. Nat Med 2022;28(6):1240-8. https://doi.org/10.1038/s41591-022-01811-5
39. Trewin S, Basson S, Muller M, Branham S, Treviranus J, Gruen D, et al. Considerations for AI fairness for people with disabilities. AI Matters 2019;5(3):40-63. https://doi.org/10.1145/3362077.3362086
40. Huq AZ. Racial equity in algorithmic criminal justice. Duke Law J 2019;68(6):1043.
41. Hu L, Kohler-Hausmann I. What's sex got to do with fair machine learning? [Internet]. Ithaca (NY): arXiv.org; 2020 [cited at 2023 Oct 31]. Available from: https://arxiv.org/abs/2006.01770.
42. Chohlas-Wood A, Nudell J, Yao K, Lin Z, Nyarko J, Goel S. Blind justice: algorithmically masking race in charging decisions. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society; 2021 May 19-21; Virtual Event, USA. p. 35-45. https://doi.org/10.1145/3461702.3462524
43. Park Y, Hu J, Singh M, Sylla I, Dankwa-Mullan I, Koski E, Das AK. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA Netw Open 2021;4(4):e213909. https://doi.org/10.1001/jamanetworkopen.2021.3909
44. Meingast M, Roosta T, Sastry S. Security and privacy issues with health care information technology. Conf Proc IEEE Eng Med Biol Soc 2006;2006:5453-8. https://doi.org/10.1109/IEMBS.2006.260060
45. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28(1):31-8. https://doi.org/10.1038/s41591-021-01614-0
46. Bartoletti I. AI in healthcare: ethical and privacy challenges. In: Riano D, Wilk S, ten Teije A, editors. Artificial intelligence in medicine. Cham, Switzerland: Springer; 2019. pp. 7-10. https://doi.org/10.1007/978-3-030-21642-9_2
47. Cavoukian A. Privacy by design [Internet]. Toronto, Canada: Information and Privacy Commissioner of Ontario; 2010 [cited at 2023 Oct 31]. Available from: https://privacysecurityacademy.com/wp-content/uploads/2020/08/PbD-Principles-and-Mapping.pdf.
48. Personal Information Protection Commission. Artificial intelligence (AI) personal information self-checklist [Internet]. Seoul, Korea: Personal Information Protection Commission; 2021 [cited at 2023 Oct 31]. Available from: https://www.korea.kr/common/download.do?fileId=197266311&tblKey=GMN.
49. Scheibner J, Raisaro JL, Troncoso-Pastoriza JR, Ienca M, Fellay J, Vayena E, et al. Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. J Med Internet Res 2021;23(2):e25120. https://doi.org/10.2196/25120
50. Bai T, Luo J, Zhao J, Wen B, Wang Q. Recent advances in adversarial training for adversarial robustness [Internet]. Ithaca (NY): arXiv.org; 2021 [cited at 2023 Oct 31]. Available from: https://arxiv.org/abs/2102.01356.
51. Qiu S, Liu Q, Zhou S, Wu C. Review of artificial intelligence adversarial attack and defense technologies. Appl Sci 2019;9(5):909. https://doi.org/10.3390/app9050909
52. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363(6433):1287-9. https://doi.org/10.1126/science.aaw4399
53. Taghanaki SA, Das A, Hamarneh G. Vulnerability analysis of chest X-ray image classification against adversarial attacks. In: Stoyanov D, Taylor Z, Kia SM, et al. Understanding and interpreting machine learning in medical image computing applications. Cham, Switzerland: Springer; 2018. p. 87-94. https://doi.org/10.1007/978-3-030-02628-8_10