Bias – The Achilles Heel of Artificial Intelligence in Healthcare

Fara Aninha Fernandes,[a,b] Georgi Chaltikyan,[a] Martin Gerdes,[b] Carmen Kraemer,[c] Christian W. Omlin[b]

[a]Technische Hochschule Deggendorf - European Campus Rottal-Inn, Germany
[b]University of Agder, Norway
[c]Aachen, Germany

ABSTRACT

The field of artificial intelligence (AI) has evolved considerably since the end of the 20th century. While this technology shows great promise and potential to solve daily tasks, the question of fairness of predictions and recommendations decisions by AI models needs to be addressed. There have been examples of AI models performing unfair and creating prejudiced decisions, which has led to a growing need to be able to know 'why' and 'how' these models make decisions. This is particularly important in the healthcare field, where the outcomes of AI models play a decisive role in the well-being of patients with potentially life-or-death impact. In addition, a system for detecting and mitigating biases needs to be developed so that the advantages of AI can be utilized in healthcare. A scoping review was carried out to study the source, nature and impact of biases of AI models. Results showed that bias can be data-driven, algorithmic or introduced by humans. These biases propagate deeply rooted societal inequality, misdiagnose patient groups, and further perpetuate global health inequity. Mitigation of biases is proposed at various stages of the machine learning pipeline. These strategies use techniques such as scrutinizing the way data is collected, better representation of patient groups, optimal training of the model and evaluating model performance. In conclusion, it must be ascertained that AI decisions are free of unwarranted biases and justly fair. Therefore, in an effort to mitigate bias, AI models should adopt systems that contain techniques in which biases can be predicted, measured, explained and then mitigated.

KEYWORDS

Artificial intelligence, machine learning, bias, explainability, interpretability, explainable AI

1. Introduction

Artificial intelligence (AI) technology is pervasive in many industries today. As the term 'intelligence' rightly suggests, this technology boasts of 'smart machines' performing mundane as well as complicated tasks, otherwise performed by humans. AI technology shows a lot of promise in the healthcare industry and is tasked with performing sensitive operations. When AI is used in the healthcare field, it must perform its operations with impeccable accuracy. However, the existence of bias in AI models thwarts this effort, resulting in distrust and a lack of confidence in their adoption. Therefore, the issue of bias

2. Background

Artificial intelligence (AI) can be seen at work in a plethora of applications in smartphones, the internet of things and much more. The chatbot ChatGPT is an impressive language model in versatility and capability.[1] An overview of some further AI applications in various industries is provided in Table 1.

Table 1: Overview of some applications using artificial intelligence

Field	*Example*
Clinical Medicine	Intelligent assistants in intensive care units that collect and process electronic data including in-hospital mortality, readmission, length of stay[2] Neural networks that detect patterns in radiographic images and H&E stained slides[3]
Transport	Smart traffic sensors that determine traffic conditions, identify the severity of traffic incidents, predict bus arrival times[4] Autonomous vehicles that incorporate AI software[4]
Banking	Algorithms that check genuine credit card transactions, block risky transactions, verify client identity, interact via chatbots, retrieve information from documents and use robo-advisors[5]
Marketing	Technologies that enable voice purchase requests, virtual assistants, authorization of payments through face recognition[6] Digital assistants and chatbots that facilitate decision-making and offer customized online shopping experiences[6]

AI encompasses a group of algorithms which are capable of solving problems that usually require human intelligence. AI includes machine learning (learns without being explicitly programmed) and a further subclass of machine learning includes deep learning (uses artificial neural networks to learn). Due to its wide capability and consistent decision-making, AI performs on par with humans or in some cases, better than humans.[7] AI has a particular role in healthcare that can be divided into administrative and clinical applications. The focus of using AI in healthcare is to enhance the analysis of medical data, improve treatment outcomes and the efficacy of the healthcare industry.[8] While humans are subject to conjecture, tradition, convenience and habit in clinical decision making, decision rules created with AI support a predictable behavior and have shown to reduce clinical error.[3,7] Researchers found that human cognitive biases can lead to poor agreement between operators (due to variance in cognitive reasoning, and different individual experience) and this vulnerability from human bias can be overcome by AI.[9] Therefore, AI can be used for sensitive tasks in healthcare such as clinical prediction, decision-making and public health policymaking for the benefit of society.[7,9] Table 2 lists some of the uses of AI in healthcare.

Table 2: Clinical uses of AI in healthcare [3,8,10]

- Provide recommendations for the prescription of medicines, adherence to clinical practice guidelines
- Analyze CT scans, provide recommendations for radiation treatment
- Diagnose malignancies from photographs, histopathological slides
- Predict eye diseases from retinal scans
- Predict a risk of sepsis
- Assess risk and predict cardiovascular events
- Automate sleep scoring in sleep medicine
- Identify drug-drug interactions and develop personalized treatments
- Improve workflow, reduce healthcare costs, shorten hospital waiting times

AI and the concept of bias:

However, the role of AI is not infallible. There are anecdotal examples of ChatGPT providing faulty answers to relatively simple questions.[11] There have also been a number of case reports where AI systems have been faulty in their decision-making and even suspected of being ‘prejudiced’. This bias is defined from two perspectives – statistical (error with the use of statistical analysis) and cognitive (innate or a learned tendency in favor of or against an individual or group, based on preconceived convictions or preferences).[9,12] Due to the potential of AI in a sensitive area such as health, biased AI algorithms can have serious implications on health outcomes. Obermeyer, Powers, Vogeli and Mullainathan[13] reported the case of an algorithm used by health systems and payers (organizations responsible for paying for healthcare services) to determine patients for high-risk care management. The algorithm made use of a proxy ‘total medical expenditures’ in order to determine the ‘health needs’ of a patient. However, this relationship is often not true since it is possible that a person in need of healthcare may not have spent on medical expenses in the past. It was found that the bias was detected in black patients who had lesser medical expenses, even though they were as sick as white patients. This led to the conclusion that the prediction of costs led to a racial bias. Other examples include AI systems exhibiting bias at interpreting chest radiographs (trained on gender-imbalanced data) and detecting skin cancers (trained on race-imbalanced data).[14]

It is therefore critical that AI systems are not biased in their outcomes, while providing a distinct advantage in the healthcare field. The sources of bias and their consequences should therefore be studied. Vokinger, Feuerriegel and Kesselheim[15] advocated an approach to check for bias along the machine learning (ML) pipeline that consists of the following stages:

- Collection of clinical data, preparation of data for further steps in model development;
- Training of machine learning models to perform an intended task;
- Evaluation of performance of the machine learning model;
- Authorization and deployment of the machine learning model in clinical practice.

Barclay and Zuanazzi[16] describe additional checkpoints along the machine learning pipeline that include: data collection, annotation, modelling, clinical validation, product integration, certification and deployment & service. Using the machine learning pipeline, it is probable that bias can arise at any of the stages.[15–17] The use of checkpoints along the ML pipeline also serves to provide tools to mitigate the biases at their source. A further concern in using AI for healthcare involves the use of 'black boxes' in deep learning models.[18] Since the working of these 'black boxes' remains esoteric, it further disseminates the notion of bias. There are attempts to introduce techniques of 'explainability' and 'interpretability' in order to be able to understand the functioning of these models. AI models for healthcare do not differ from AI models used in other industries. However, the outcomes of AI models used in healthcare have relatively serious consequences if they are flawed (for example, the case reported by Obermayer et al.[13]). Consequently, any approach to mitigate bias must also be researched in order to apply it to AI models for healthcare. This research was done to determine the source of bias in algorithms designed for healthcare and the possibilities for the mitigation of these biases.

3. Methodology

A scoping review of literature on bias in AI algorithms used in healthcare was carried out to answer the following three questions:

1. What are the sources of bias in AI algorithms designed for healthcare and medicine?

2. What are the consequences and pitfalls of biased AI-based medical solutions in healthcare?

3. Can these biases be located and mitigated?

The databases of Google Scholar, PubMed and Scopus were explored using the following keywords: "bias AND (artificial intelligence OR AI OR machine learning) AND (health OR medicine)". The search was carried out to retrieve articles from the databases from the last five years up to March 2023. The selection of articles was following the PRISMA19 model (Figure 1).

The inclusion criteria encompassed only those articles that delineated the use of AI models for healthcare in accordance with the three research questions. Papers that reviewed the use of AI models for specific clinical use cases and those that provided a view of bias in AI and other ML models, but were not specific to healthcare were excluded.

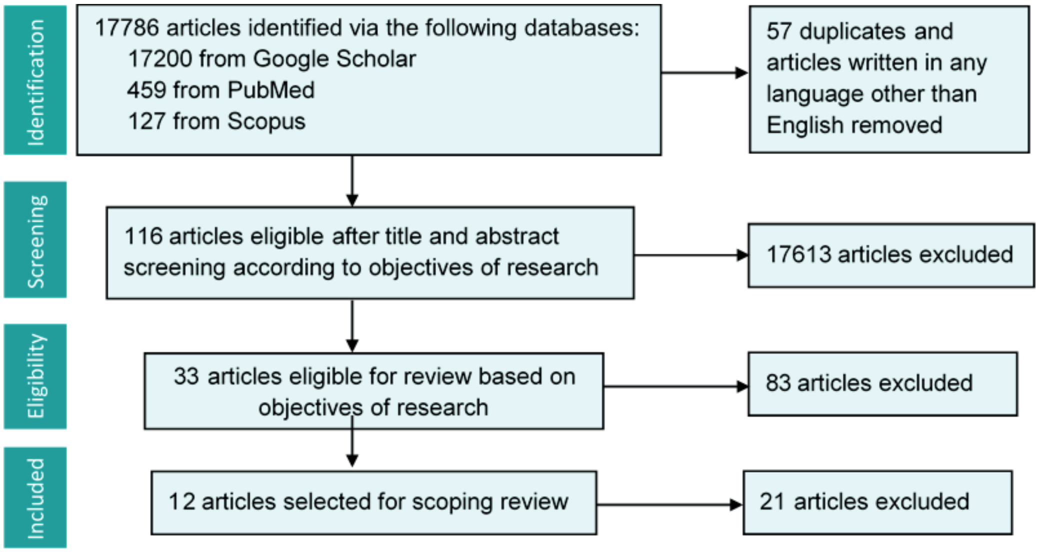
Identification
17786 articles identified via the following databases:
17200 from Google Scholar
459 from PubMed
127 from Scopus
57 duplicates and articles written in any language other than English removed
Screening
116 articles eligible after title and abstract screening according to objectives of research
17613 articles excluded
Eligibility
33 articles eligible for review based on objectives of research
83 articles excluded
Included
12 articles selected for scoping review
21 articles excluded

Figure 1: PRISMA model[19] showing the inclusion and exclusion of records

At the step of identification, only articles written in English were selected. If an article was retrieved from more than one database (i.e. a duplicate), only one such article was retained. The selected articles were then screened by analyzing the title and abstract. At this stage, articles that did not contain the keywords in the title and abstract were excluded. An article was then considered as eligible if the content provided information regarding the occurrence of bias in AI in models, their consequences as well as approaches to mitigation of these biases. Articles that did not pertain to answer any of the three research questions were excluded. In the final phase of selection, only those articles were included that provided information to either two or more questions of this study.

The twelve articles that were selected were reviewed and analyzed. In order to answer the question on the source of biases, the different types of biases encountered in AI models were noted. Similarly, the consequences of these biases and the approaches to mitigation were extracted. Retrieved information was categorized according to the three questions of this research, viz. sources of bias in AI algorithms, consequences and pitfalls of biased AI-based medical solutions in healthcare and mitigation of biases. The extracted information was categorized to follow the machine learning pipeline.

4. Results

The analysis of the selected articles revealed the following key points. These points are divided into sources of bias in AI algorithms, consequences and pitfalls of biased AI-based medical solutions in healthcare and location and mitigation of biases. In each of these three sections, the retrieved data is further categorized according to the stages of the machine learning pipeline.

1. **Sources of bias in AI algorithms:**
   The sources of bias in algorithms designed for healthcare can be categorized as shown in Table 3:

Table 3: Sources of bias in AI algorithms for health

Category	*Example*
According to nature of bias[20,21]	Social bias or human-centric elements of bias Statistical bias or methodology-centric elements of bias
According to origin of bias[2,9,20]	Data-driven: Seen in models trained on homogenous / imbalanced / non-representative datasets that suffer from a lack of diversity and missing data Human: Arises from the cognitive ability and accidental behavior that humans are prone to possess Algorithmic: Emerges from outcomes of systematic errors in AI-based systems that affects their ability to classify, estimate risk levels, or make predictions
According to introduction at stages of machine learning pipeline[15]	Stage of data collection Stage of data selection Stage of model development and training Stage of model evaluation Stage of model deployment

The biases introduced at the various stages of model development were further elaborated as follows:

Biases and factors contributing to risk of bias introduced at the stage of data collection: This first checkpoint is the source of data that were used to train the model. Issues included the exclusion of data from under-represented groups of the human population, other missing data, inaccessible data and metadata.[2,15,20–22] Such data is, therefore, not diversified, incomplete and therefore does not represent the entire population. It was also unclear as to how much data the AI system needs in order to provide a correct decision. 'Sampling bias', on the other hand, referred to the use of an inadequate number of cases in the dataset.[2,12,15,23] There was also the problem of data being collected as 'interesting cases' which cannot represent the percentage of normal and abnormal cases in a population (sample selection bias).[3] Unequal distribution also occurred when some patients access more health services, visit multiple hospitals and patient portals more frequently than others, that constitutes to 'misclassification' and adds to bias introduced at this stage.[2,23] Additionally, it was noted that an inadequate amount of data is partly due to restrictions imposed in the access and collection of data for training of algorithms.[7]

Biases introduced at the stage of data selection: The errors in data selection were shown to have a human factor. The bias and errors introduced by humans were due to errant measurements caused by the intrinsic and unintentional habits of people who handle data.[9] Furthermore, the training of algorithms using such data resulted in 'faulty algorithms' that perpetuated societal, racial, and gender (historical) bias.[9,23] Another aspect was the limited amount of correctly labelled data and consequently the use of data with no ground truth ascribed to it that led to poor reproducibility.[3]

Biases introduced at the stage of model development and training: The stage of model development and training represented a significant number of biases. At first, biases introduced in the stages of data collection and selection were further embedded into algorithms.[9,15] Implicit bias was described as a situation that resulted from unforeseen correlations between variables in a model.[12] The use of a biased proxy or a statistically biased estimator resulted in an incorrect, fallacious prediction that was called 'label choice bias' or 'mislabeling'.[22] Similarly, the uneven distribution of features across different groups caused 'feature selection bias' or 'feature leakage'.[21] The annotation of data also led to bias if there was a consistent error.[21,24] Problems of 'underfitting' of the model resulted in the initial stage when the model was untrained or only partially learning information needed to provide an output.[24] On the other hand, an 'overfitting/overtrained' model included salient features and noise that generated high error rates.[24,25] Apart from these biases, the 'aging' of an algorithm and the passage of time since it was trained deteriorates it and resulted in 'temporal bias'.[3,21,23] Unconscious bias and power imbalances also have a human factor when algorithms were designed to solve apparent problems that are neither necessary to be automated nor are considered to be essential by health providers and patients.[23]

Biases introduced at the stage of model evaluation: The stage of model evaluation is a useful checkpoint for the detection as well as mitigation of bias. However, there are instances when the metrics used to evaluate the model resulted in other biased outcomes. It was noted that the boosting of true-positives and false-negatives resulted in a good model for screening, but did not function well for diagnosis.[3,22] Rebalancing of data was another approach to include an equal number of cases and controls, however, this led to overdiagnosis of cases.[3] Another concern was the 'metric/ranking selection bias' where there was a difference between the metrics used to evaluate the model and ill-defined standards for clinical quality.[21] Machine learning models based on artificial neural networks were considered to be 'black boxes' in that any errors and biases in the predictions of the model were hard to detect. [2,3,23]

Biases introduced at the stage of model deployment: A number of biases were attributed to machine learning models after their deployment and clinical implementation. When a model was applied to a population that had different characteristics from the characteristics of the population that was used to train the model, it was shown to exhibit 'population bias' or 'selection bias'.[25] Other terms to describe these circumstances were 'distributional shift' and 'out of sample input'.[3] Human factors contributing to bias at this stage occurred when operators misinterpreted the algorithm's decision. It was noted that 'automation bias' was a phenomenon where a clinician accepted the decision/prediction of an automated system and avoided looking for confirmation (automation complacency).[3,20]

2. Consequences and pitfalls of biased AI-based medical solutions in healthcare:

The consequences and pitfalls of biased algorithms at the various stages of model development were described as follows:

Consequences of biases introduced at the stage of data collection and selection: The use of biased AI algorithms in healthcare was found shown to result in a number of untoward consequences. Bias in data collection and selection resulted in the demonstration of inherent, human-associated and hidden societal biases.[3,9] It was observed from the retrieved papers that misrepresentation of patient groups was a common consequence that led to prejudiced outcomes such as misdiagnoses and unequal treatment of these groups.[9,20] Patients with less access to healthcare had insufficient information stored in electronic health records that led to delayed diagnosis.[2,20] With such glaring discrepancies in the diagnoses of patient groups, it was noted that clinical AI for diverse populations is not guaranteed.[7,9,12] A major finding was the disproportion in the retrieval and use of data from all global regions. AI solutions were more applicable to populations from data-rich regions than those populations from data-poor regions.[7,12] Missing data can lead to unsuitable insurance plans, if they are based on misrepresented health records.[2,20] The use of datasets with a small number of datapoints resulted in unreliability of the model.[23] Data-driven biases were further propagated to the training and development stages of the machine learning pipeline and continue to incorporate unfair predictions and decisions which led to inappropriate treatments.[23] Ethical issues may then arise if AI systems (with data-driven biases) autonomously triage patients to access clinical services.[3]

Consequences of biases introduced at the stage of model development and training: The training of models using datasets with missing data had consequences in that it was not beneficial to patients who were not represented in the dataset.[2] The degradation of the quality of algorithms was also found to be invalid in clinical scenarios.[3] Algorithms trained on insufficient data provided only the mean of the cases and did not account for unique cases (underestimation).[2] The use of data from multiple centers had the potential to cause highly biased datasets.[21] Validity was a major turning point in that models that inherited biases from the previous stage were not externally valid. Selection bias was encountered at this stage and had a profound effect on external validity. The outcomes of such biased ML models (that are trained on a particular study population) are not valid when applied to the target population.[25] The use of 'black box' models do not show transparency in their working and, therefore, make the interpretation of decisions difficult. [2,3,25]

Consequences of biases introduced at the stage of model evaluation: In evaluating ML models, the use of metrics such as 'accuracy' may not reflect real-world situations. Therefore, only using the accuracy as a performance metric to optimize the model led to the creation of faulty models.[3] The use of black box models, where there is no transparency of the model, translates to a loss of confidence and lack of trust by users.[23] A model that consistently produces the same set of results with errors can reflect on the quality of care.[23]

Consequences of biases introduced at the stage of model deployment: On clinical deployment, the difference in characteristics of data used for training the AI model and the characteristics of the population for whom the AI system will be used was questioned.[23] Such models perform inefficiently and are not useful. Algorithms that were designed to tackle problems that are not clinically useful or time-saving, resulted in an unnecessary waste of resources.[23] The AI systems were also subject to the individual interpretations of humans which could further introduce more biases and erroneous predictions and decisions. Automation bias when supplemented by previous biases further questions the reliability and validity of the AI systems.[3] Another aspect to consider is the issue of responsibility when biased algorithms are used.[23]

3. Location and mitigation of biases:

The mitigation of biases is best attempted along the stages of the machine learning pipeline. These mitigation strategies were as follows:

Mitigation of biases at the stage of data collection and data selection: A broader inclusion of patient demographic groups in terms of age, gender, race and ethnicity is an attempt to address biased datasets.[2,7,9,20,21] The sharing of data avoids the monopoly of data-rich databases and contribute to a wider representation of the population.[7] There is also the need to define the target population, capture data from diverse populations and normalize data taken from different sources.[2] Use of classifier performance for the dataset can be used to detect representation bias.[22] Some studies recommended the use of reporting tools like the PROBAST (prediction model risk of bias assessment tool) which functions to check the performance level of models, risk of bias and the suitability of its use in a particular population.[15,20]

Mitigation of biases at the stage of model development and training: Imbalanced data can be rectified at the stage of model development through the use of 'adversarial debiasing' or 'oversampling', simulated datasets that include missing variables, counterfactual simulations and 'continual learning'.[20,25] These methods generate synthetic data in an attempt to increase the representation of underrepresented classes. Additionally, 'selection bias' can be overcome by the application of an "independent external dataset" to check the suitability of the study population used to train the model with the target population.[25] Furthermore, it was recommended to test and tune models to perform optimally in all population groups and at all stages using the ground truth as a proxy.[2,9,12,20,24] Bias brought about by overfitting can be improved by regularization techniques, cross validation, and data augmentation.[25] Continued monitoring, follow-up, use of feedback loops and retraining the model are necessary to check the results of the machine learning models.[2,3,23]

Mitigation of biases at the stage of model evaluation: The common methods during the stage of model evaluation include the use of performance metrics such as accuracy and F1 score. Models should be trained with the consequences of false negatives (missed diagnoses) and false positives (over-diagnoses).[2,22,23] Methods for testing algorithm performance are detailed in CLAIM (Checklist for Artificial Intelligence in Medical Imaging) and can be applied to detect selection bias.[25] AI algorithms can also be employed to flag decisions that require to be checked.[20,21] Other techniques used are interpretability and explainability. These techniques are specifically introduced to allow human users to understand the results produced by machine learning algorithms. In the field of medical imaging, visual-based methods for explainability include 'saliency maps' that attempt to identify areas and features that contributed to the prediction of the algorithm.[3,22,23]

Mitigation of biases at the stage of model deployment: The first step to mitigate biases at this stage is to have fail-safe methods that will check if socio-demographic characteristics of patients (for whom the model will be used) are representative of the patients included in the training data. This includes 'field-testing' to assess the performance of algorithms in different population groups and clinical settings. Unwanted bias must be recognized from feedback loops, namely transparency and justification of levels of bias for future avoidance and removal.[21] Finally, it needs to be ascertained that clinically useful and meaningful algorithms are designed that are beneficial to the patient, the healthcare provider and to the whole healthcare system through the improvement of outcomes.[2,23]

5. Discussion

Knowledge of the origin or source of bias in the AI algorithm is critical. There are many examples where faulty algorithms were rectified by identifying the source. Recognition of the possible sources of biases at the outset leads to an efficient model development. However, machine learning algorithms that are based on artificial neural networks are subject to 'black-box decision-making'. These are models where it is difficult to detect biases and other errors, prompting researchers to devise techniques in order to evaluate these models. The present research identified three probable sources of bias that are linked to – a) data, b) model design and development, and c) human involvement. Once these sources have been recognized, it is important to check at which stage of the machine learning pipeline they can be introduced. Vokinger, Feuerriegel and Kesselheim[15] in 2021 demonstrated this strategy as well as the possible mitigation techniques at each stage. By using the machine learning pipeline, it was concurred that bias can arise at any of the four stages, propagate along the pipeline and then exacerbate the bias at the final stage. Igoe[17] described this phenomenon as a "trickle-down effect". An ideal workflow would be at first to recognize all possible biases that could arise in the development of a model and apply mitigation techniques right from the start of the machine learning pipeline. Waeed and Omlin[26] utilized the phases of the ML pipeline to demonstrate that XAI can be used to detect, prevent and even mitigate bias. During the course of model development, it is advisable to use techniques (derived from prior knowledge of bias origin) to mitigate expected biases.

A significant emphasis is placed on the data that is used for training algorithms for healthcare. The lack of diversity in datasets is partially due to the limitations in the sharing of medical data. These may be due to privacy laws as well as the lack of interoperability between systems.[14] The use of electronic Health Information Exchange (HIE) is a solution to make large amounts of data accessible to machine learning.[18]

The most obvious consequences of biased AI algorithms are exclusion, unfair allotment of resources, misdiagnoses, fatal outcomes, and lack of external validity. These are largely due to imbalanced data that is used to train the algorithms. Even though one is aware of this fact, the perceived usefulness of AI is once again in doubt, as it cannot be used for broader populations, thus perpetuating global health inequity and existing health disparities. If underrepresented groups are susceptible to the impact of bias, it needs to be addressed in order that AI in health care is applicable to the United Nations Sustainable Development Goal, i.e. 'AI for Good'. An ideal scenario would be to improve the collection of high-quality and correctly labelled data. Kaushal, Altman and Langlotz[14] call for the implementation of a strong infrastructure in terms of technology, regulation, economic and the privacy and safety of data.

The mitigation of biases should be attempted by first diligence in data handling and during model development. Potential biases should be anticipated and their mitigation employed. A diverse team approach consisting of engineering and biomedical teams may also be intuitive.[23] A recent approach to mitigation involves the use of interpretability and explainability (XAI) techniques used especially for 'black-box' models. If XAI techniques bring to light the features that influence the outcome of a ML model decision, it will be a tool with direct benefit for the user. Models that are transparent and understandable by humans instill confidence in their utilization, thus encouraging their use. Therefore, a combination of quality mitigation techniques and XAI techniques integrated into algorithms may be the probable solution to the Achilles Heel of artificial intelligence. A list of recommendations to overcome biases in AI algorithms is outlined in Table 4.

Table 4: Recommendations to overcome biases in AI Algorithms

Approach through the collection and selection of data: The collection of better data is first and foremost the most important step to avoid disparities in datasets. These include:

- Datasets should be representative of diverse population groups (address race, gender, age and socio-economic);
- Databases should be improved by addressing the imbalance of country-wise distribution of databases;
- Sharing of data and interoperability should be encouraged.

Approach through best practices in algorithm development: The development of algorithms and the way data is used is critical in determining the usefulness of the model. Best practices include:

- An algorithm should be designed to fulfill the need of clinicians;
- The intended use of the algorithm must be determined;
- Underfitting and overfitting must be considered while training the algorithm;
- Proxies and related labels must be judiciously chosen and rechecked for potential biases;

Approach through best practices in algorithm evaluation: The evaluation phase is a critical checkpoint to find potential biases. Strategies include:

- Performance metrics must be applied to evaluate the model;
- Interpretability and explainability techniques should be used to instill trust in the user;
- Basic criteria of appropriateness, fairness, and bias should be used for evaluation

Approach at the stage of clinical deployment: The final stage includes the following:

- External validation and model re-calibration should be carried out prior to clinical implementation (match the algorithm training sets to the target context and population);
- Manufacturers should be transparent in providing datasheets of the AI system.

6. Conclusion

Artificial intelligence (AI) has a distinct advantage over traditional algorithmic methods. However, just like the metaphorical ‘Achilles Heel’, AI is vulnerable despite its strong capabilities and apparent invincibility with weaknesses of which bias is one. This aspect is of particular concern when using AI in healthcare. Biases originate from data, during model development and evaluation, and after deployment. The existence of these biases challenges the notion of trustworthiness of using AI technology. Bias is a concern to the field of AI, and it is the major reason for its slow adoption despite its promises. Explainability and interpretability must be valid, consistent and reproducible if they are to instill principles of inclusivity, openness, and user trust. They may also be sensible approaches for bias detection and mitigation; however, there still exists the need for further validation and evaluation of these methods. Future research should be focused on further capabilities of XAI methods to be able to detect and hence mitigate bias. In this regard, it is recommended that metrics play a larger role not only in the evaluation of AI algorithms, but also in the evaluation of XAI methods.

References and notes

[1] OpenAI. Introducing ChatGPT. OpenAI. November 30, 2022. Accessed March 23, 2023. https://openai.com/blog/chatgpt

[2] Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern Med.* 2018;178(11):1544-1547. doi:10.1001/jamainternmed.2018.3763

[3] Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf.* 2019;28(3):231-237. doi:10.1136/bmjqs-2018-008370

[4] Kaya O, Schildbach J, Deutsche Bank AG, Schneider S. Artificial intelligence in banking: A lever for profitability with limited implementation to date. Deutsche Bank Research. Published June 4, 2019. https://www.dbresearch.com/PROD/RPS_ENPROD/PROD0000000000495172/Artificial_intelligence_in_banking%3A_A_lever_for_pr.pdf

[5] Abduljabbar R, Dia H, Liyanage S, Bagloee SA. Applications of Artificial Intelligence in Transport: An Overview. *Sustainability*. 2019;11(1):189. doi:10.3390/su11010189

[6] Jarek K, Mazurek G. Marketing and Artificial Intelligence. *Central European Business Review.* 2019;8(2). doi: 10.18267/j.cebr.213

[7] Celi LA, Cellini J, Charpignon M-L, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities-A global review. PLOS *Digit Health.* 2022;1(3):e0000022. doi:10.1371/journal.pdig.0000022

[8] Rong G, Mendez A, Bou Assi E, Zhao B, Sawan M. Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering*. 2020;6(3):291-301. doi:10.1016/j.eng.2019.08.015

[9] Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. *Patterns* (N Y). 2021;2(10):100347. doi:10.1016/j.patter.2021.100347

[10] Racine E, Boehlen W, Sample M. Healthcare uses of artificial intelligence: Challenges and opportunities for growth. *Healthcare Management Forum.* 2019;32(5):084047041984383. doi:10.1177/0840470419843831

[11] Eva G. The best / worst / funniest / most absurd etc. ChatGPT responses. December 8, 2022. Accessed March 23, 2023. https://medium.datadriveninvestor.com/the-best-worst-funniest-most-absurd-etc-chatgpt-responses-9094dda976fb

[12] Fletcher RR, Nakeshimana A, Olubeko O. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Front Artif Intell.* 2020;3:561802. doi:10.3389/frai.2020.561802

[13] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342.

[14] Kaushal, A., Altman, R., & Langlotz, C. Health Care AI Systems Are Biased. November 17, 2020. Accessed February 28, 2023. https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/

[15] Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. *Communications Medicine (London)*. 2021;1:25. doi:10.1038/s43856-021-00028-w

[16] Barclay, L., & Zuanazzi, V. Bias in medical imaging AI: Checkpoints and mitigation. November 24, 2021. Accessed March 1, 2023. https://www.aidence.com/articles/bias-in-medical-imaging-ai/?cn-reloaded=1

[17] Katherine J. Igoe. Algorithmic Bias in Health Care Exacerbates Social Inequities — How to Prevent It. March 12, 2021. Accessed February 27, 2023. https://www.hsph.harvard.edu/ecpe/how-to-prevent-algorithmic-bias-in-health-care/

[18] Hobson, C., & Ross, K. Removing Data Bias from AI and Machine Learning Tools in Healthcare White Paper. February 25, 2021. Accessed February 28, 2023. https://www.himss.org/resources/removing-data-bias-ai-and-machine-learning-tools-healthcare-white-paper

[19] Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med.* 2018;169(7):467-473. doi:10.7326/m18-0850

[20] Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. *JAMA*. 2019;322(24):2377-2378. doi:10.1001/jama.2019.18058

[21] Baxter JSH, Jannin P. Bias in machine learning for computer-assisted surgery and medical image processing. *Comput Assist Surg* (Abingdon). 2022;27(1):1-3. doi:10.1080/24699322.2021.2013619

[22] Ganz M, Holm SH, & Feragen A. Assessing Bias in Medical AI. Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML) 2021.https://www.cse.cuhk.edu.hk/~qdou/public/imlh2021_files/64_cameraready_icml_2021_interpretable_machine_learning_in_healthcare_workshop.pdf

[23] Doyen S, Dadario NB. 12 Plagues of AI in Healthcare: A Practical Guide to Current Issues With Using Machine Learning in a Medical Context. *Front Digit Health.* 2022;4:765406. doi:10.3389/fdgth.2022.765406

[24] Zhang K, Khosravi B, Vahdati S, et al. Mitigating Bias in Radiology Machine Learning: 2. Model Development. *Radiol Artif Intell.* 2022;4(5):e220010. doi:10.1148/ryai.220010

[25] Yu AC, Eng J. One Algorithm May Not Fit All: How Selection Bias Affects Machine Learning Performance. *Radiographics.* 2020;40(7):1932-1937. doi:10.1148/rg.2020200040

[26] Saeed W, Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems. Published online January 2023:110273. doi:10.1016/j.knosys.2023.110273