Diagn Interv Radiol 2025; DOI: 10.4274/dir.2024.242854

ARTIFICIAL INTELLIGENCE AND INFORMATICS

REVIEW

Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects

Burak Koçak[1]
Andrea Ponsiglione[2]
Arnaldo Stanzione[2]
Christian Bluethgen[3]
João Santinha[4]
Lorenzo Ugga[2]
Merel Huisman[5]
Michail E. Klontzas[6]
Roberto Cannella[7]
Renato Cuocolo[8]

[1]University of Health Sciences, Başakşehir Çam and Sakura City Hospital, Clinic of Radiology, İstanbul, Türkiye

[2]University of Naples Federico II, Department of Advanced Biomedical Sciences, Naples, Italy

[3]University of Zurich, University Hospital Zurich, Diagnostic and Interventional Radiology, Zurich, Switzerland

[4]Digital Surgery LAB, Champalimaud Research, Champalimaud Foundation; Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

[5]Radboud University Medical Center, Department of Radiology and Nuclear Medicine, Nijmegen, Netherlands

[6]University of Crete, School of Medicine, Department of Radiology; University Hospital of Heraklion, Department of Medical Imaging,Crete, Greece; Karolinska Institute, Department of Clinical Science Intervention and Technology (CLINTEC), Division of Radiology, Solna, Sweden

[7]University of Palermo, Department of Biomedicine, Neuroscience and Advanced Diagnostics, Section of Radiology, Palermo, Italy

[8]University of Salerno, Department of Medicine, Surgery and Dentistry, Baronissi, Italy

Corresponding author: Burak Koçak

E-mail: drburakkocak@gmail.com

Received 11 May 2024; accepted 11 June 2024.

Epub: 02.07.2024

Publication date: 03.03.2025

DOI: 10.4274/dir.2024.242854

ABSTRACT

Although artificial intelligence (AI) methods hold promise for medical imaging-based prediction tasks, their integration into medical practice may present a double-edged sword due to bias (i.e., systematic errors). AI algorithms have the potential to mitigate cognitive biases in human interpretation, but extensive research has highlighted the tendency of AI systems to internalize biases within their model. This fact, whether intentional or not, may ultimately lead to unintentional consequences in the clinical setting, potentially compromising patient outcomes. This concern is particularly important in medical imaging, where AI has been more progressively and widely embraced than any other medical field. A comprehensive understanding of bias at each stage of the AI pipeline is therefore essential to contribute to developing AI solutions that are not only less biased but also widely applicable. This international collaborative review effort aims to increase awareness within the medical imaging community about the importance of proactively identifying and addressing AI bias to prevent its negative consequences from being realized later. The authors began with the fundamentals of bias by explaining its different definitions and delineating various potential sources. Strategies for detecting and identifying bias were then outlined, followed by a review of techniques for its avoidance and mitigation. Moreover, ethical dimensions, challenges encountered, and prospects were discussed.

KEYWORDS

Artificial intelligence, machine learning, medical imaging, bias, fairness, radiology

Bias, with its various definitions depending on the context, often denotes systematic errors due to existing inappropriate models, whether intentional or unintentional.[1] Extensive studies of bias in human cognition have included the field of radiology and medical imaging, addressing biases at both personal (e.g., bias during reporting) and societal levels.[2] It is typically linked to conscious or subconscious cognitive preconceptions that may arise during clinical practice, particularly in rapid decision-making scenarios.[3,4]

Advances in artificial intelligence (AI) related to medical imaging, particularly in radiology, present new avenues to enhance patient care across different stages of the patient journey, such as triage, selecting imaging modalities, image quality improvements, risk assessment, diagnosis, and prognostication.[5-7] However, increasing integration of AI into clinical practice comes with new challenges for radiologists, who may not be accustomed to potential biases or systematic errors introduced into their workflow, thereby risking the integrity of outcomes.[8-13]

Medical publication trends indicate a growing interest in bias in AI (Figure 1). This international collaborative review effort aims to provide readers with the fundamental knowledge and potential tools or strategies necessary to navigate bias when dealing with AI for medical imaging, thus mitigating negative impacts on patient management. This study comprehensively reviews bias in AI for medical imaging, covering its fundamentals, detection techniques, prevention strategies, mitigation methods, encountered challenges, ethical concerns, and prospects.

You may cite this article as: Koçak B, Ponsiglione A, Stanzione A, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagn Interv Radiol.* 2025;31(2):75-88.

Definition of bias in artificial intelligence

The concept of bias in machine learning (ML) research and more generally in the field of predictive modeling is intrinsically tied to the concept of variance.[14] In this context, bias can be defined as the distance (or error) between the prediction and the actual target variable, whereas variance signifies the dependence of predictions on the randomness in the training data sampling (Figure 2).[15] Hypothetically, a predictive model can present any combination of high or low bias and variance. From a statistical point of view, the sum of bias (squared) and variance is represented by the mean squared error metric.[16] Interestingly, the concepts of bias and variance are not limited to the domain of statistical or ML modeling alone, but they also affect human learning and have been extensively studied in cognitive sciences.[15]

From a mathematical point of view, noise (the joint probability distribution between training and test/inference samples), bias, and variance are the three components that lead to model performance degradation and negatively affect generalization to new data.[17] Given the somewhat irreducible nature of noise, ML has focused mostly on addressing bias and variance when optimizing model performance during the hyperparameter tuning process. However, it should be made clear that these two entities are interdependent, and reducing one (e.g., variance) typically comes at the expense of increasing the other (i.e., bias), which gives birth to the concept of a bias-variance tradeoff. In recent years, the technical evolution of ML models, and especially the rise of large neural network architectures, has begun to challenge the traditional approach of validation (or cross-validation) error minimization as the ideal strategy to optimize the bias-variance tradeoff during model training.[17-20]

Main points

- In the medical artificial intelligence (AI) context, "bias" refers to systematic errors leading to a distance between prediction and truth, to the potential detriment of all or some patients.
- AI in medical imaging is at risk of being compromised by several types of biases, which could adversely affect patient outcomes.
- Understanding that medical imaging AI systems are prone to biases in various forms is key for their successful incorporation into real-world clinical settings, with greater satisfaction of end-users.
- Proactively identifying and addressing AI bias may prevent its potential negative consequences from being realized later.
- Increasing community awareness about all aspects of bias, such as fundamentals, mitigation strategies, and ethics, may contribute to the development of more effective regulatory frameworks.

Types and sources of bias

One way to comprehend imaging AI bias is by examining sources of bias related to fundamental components of the AI life cycle: study design and dataset (formulating the research question, collection, annotation,

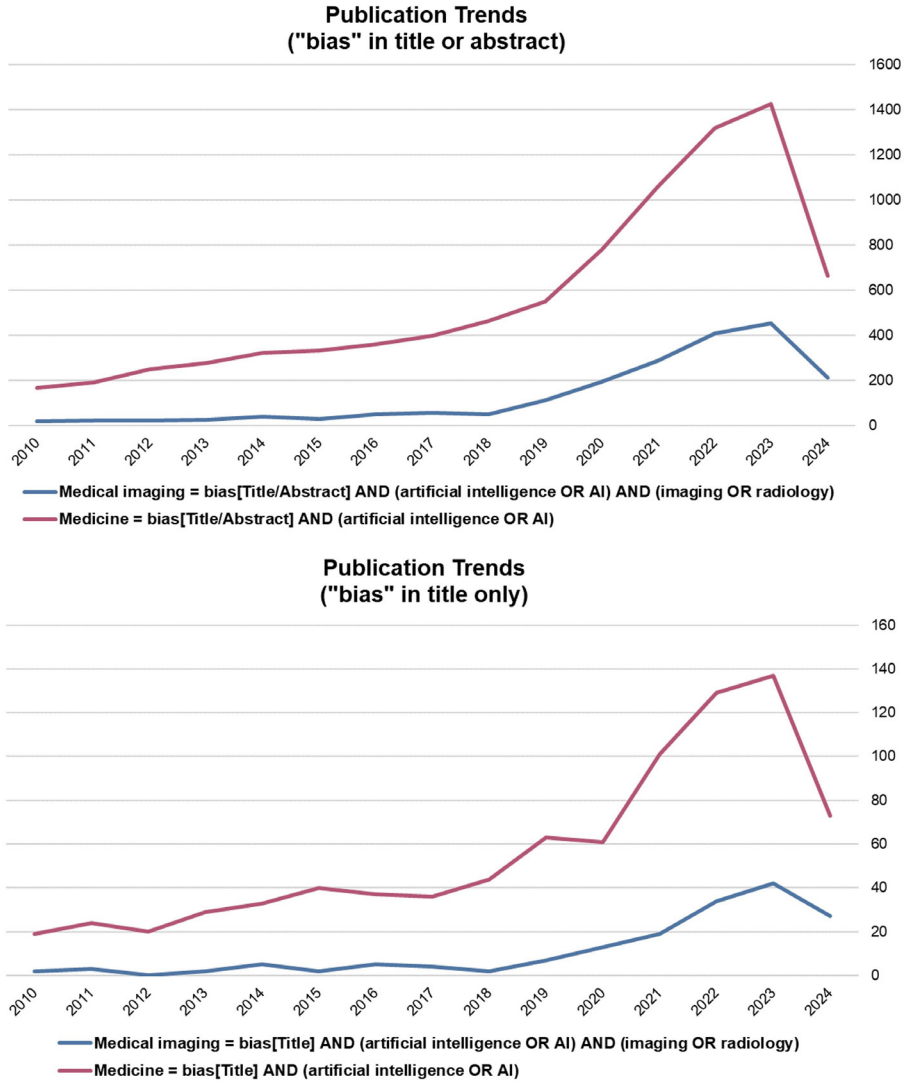

Figure 1. Publication trends about bias in medical imaging artificial intelligence (AI) in comparison with AI in medicine, with different search syntaxes to identify the occurrences of the term "bias" in the title or abstract versus the title alone. Source: PubMed; date of search: May 7, 2024.

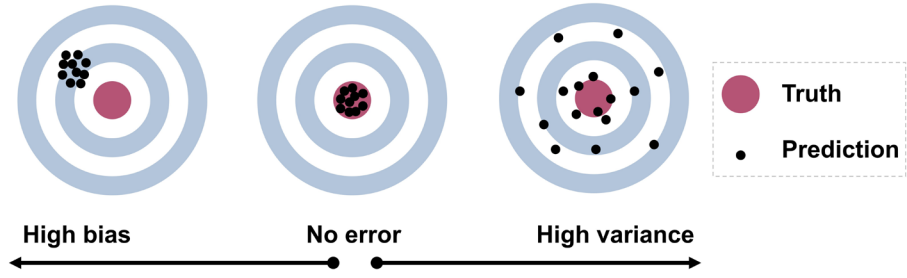

Figure 2. Over-simplified illustration of bias (i.e., systematic error) in contrast to variance, such as random noise.

preprocessing, etc.), modeling (development and evaluation before using in real-world settings), and deployment (implementation in real-world settings). This section focuses on the most common sources of bias that medical imaging professionals, particularly radiologists, may encounter. Accordingly, types and sources of bias and concepts mentioned in this review are given in Figure 3. Table 1 provides a glossary of definition of other bias sources as well, including other related concepts. Table 2 presents fictional examples for selected bias sources.

Bias related to study design and dataset

Bias can emerge when taking the very first step into the development of AI solutions for medical imaging, which is the correct identification of an unmet and relevant clinical need.[21] A valid research question must also

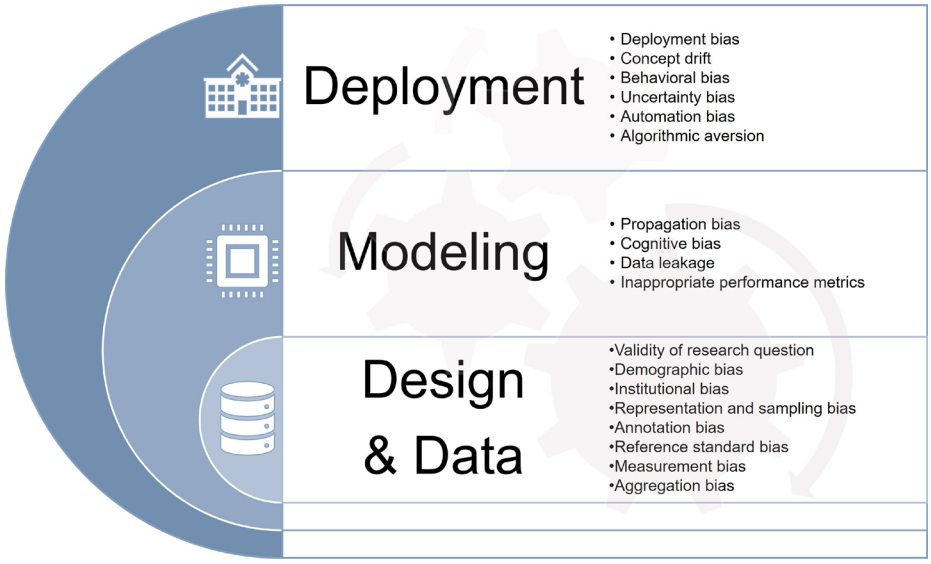

Figure 3. Main types and sources of bias and related concepts highlighted throughout this review. For other common types and sources of bias, please refer to Table 1.

Table 1. Common terminology and concepts related to bias

Terminology	Definition
Aggregation bias	False conclusions or assumptions about individuals compared with the whole population based on inappropriate combinations of distinct groups.
Algorithm fairness	Ensuring equitable outcomes across different demographic groups.
Algorithmic bias	Systematic errors or prejudices in the algorithms.
Algorithmic aversion	Reluctance or skepticism toward relying on artificial intelligence (AI) algorithms.
Annotation bias	Systematic errors mostly introduced by human annotators during the labeling process of training data, mostly related to their experience, subjective interpretation, and cognitive biases concerning the annotation task.
Automation bias	Overreliance on AI results, leading to the neglect of human decision-making.
Behavioral bias	Distortions in user behavior seen across various platforms, contexts, or datasets.
Class imbalance	Disproportionate representation of certain classes within or between the data partitions.
Cognitive bias	Systematic subjective patterns in thinking that can affect the decision-making of individuals due to reliance on heuristics (i.e., shortcut strategies derived from previous experiences to solve a problem or reach a goal).
Concept drift	Changes in correlation between input variables and output predictions over time due to fluctuations in data.
Confirmation bias	Tendency to interpret AI model results in a way that confirms their existing beliefs or expectations.
Data leakage	Exposure of target features or information to the model during training, leading to poor generalizability.
Demographic bias	Systematic errors in models that disproportionately affect specific demographic groups based on factors such as age, gender, or ethnicity.
Deployment bias	Misalignment between the envisioned purpose of a system or algorithm and its actual application.
Distributional shift	Discrepancies between the distribution of data used to train AI models and the distribution encountered in real-world deployment.
Feedback loop bias	Increase of systematic errors over time as the AI model continues to learn from its own predictions and feedback.
Institutional bias	Systematic errors led by differences in practices, protocols, or equipment across institutions.
Measurement bias	Systematic errors related to how particular features are chosen, used, or measured.
Omitted variable bias	Systematic errors appear when one or more relevant variables are omitted, or context is neglected.
Overfitting	Phenomenon where the AI model learns to memorize the training data instead of generalizing on new data.
Propagation bias	Increase of potential systematic errors present in any algorithm or pipeline and being inherited by the final model or even amplified in it.
Representation and sampling bias	Systematic errors in the collection of data, resulting in an unrepresentative sample.
Statistical bias	Discrepancies between actual and predicted values when approximating a specific statistical measure.
Temporal bias	Systematic errors arising over time, such as from the changes in medical imaging technology, protocols, or patient demographics.
Temporal drift	Changes in the distribution or characteristics of data over time, leading to discrepancies between the development and deployment AI performance.
Uncertainty bias	Influence of uncertainty on decision-making stemming from AI models.
Underfitting	Phenomenon where the AI model is too simplistic, failing to adequately capture the complexity of data.

Table 2. Examples based on fictional scenarios for selected bias sources related to medical imaging

Bias source	Example
Annotation bias	A breast artificial intelligence (AI) tool is being developed to assist in analyzing mammograms. As radiologists annotate the images to be used for its development, they primarily focus on identifying malignant masses due to their significance in cancer diagnosis. Benign calcifications, less concerning but still important, may be underrepresented in the annotations made by the radiologists. The resultant tool may have this annotation bias, being more inclined to detect malignant masses and neglecting to adequately recognize benign calcifications, leading to an increased risk of false negatives.
Automation bias	A radiologist or a clinician relies on an AI tool to interpret chest computed tomography (CT) scans. If the AI model is trained on datasets that predominantly include lung nodules, it may develop a bias toward detection of these nodules over other clinically significant findings (e.g., consolidations). By developing a tendency to prioritize the AI tool's output over the entire clinical evaluation, end-users may show an over-reliance on the AI tool, trusting it without thoroughly considering other important information present in the CT scans. This automation bias can result in missing important findings beyond lung nodules.
Confirmation bias	An experienced radiologist uses an AI tool to analyze a prostate magnetic resonance imaging (MRI) scan of a patient with a history of urinary symptoms and elevated prostate-specific antigen levels. As the radiologist examines the imaging results, they may identify certain features that appear to support their initial suspicion of benign prostatic hyperplasia (BPH) based on the observed prostatic enlargement and nodularity. However, the tool also flags some potential small focal lesions or suspicious tissue characteristics, suggestive of prostate cancer. Despite these, the radiologist's focus on confirming their preliminary diagnosis of BPH may lead them to ignore the important alerts provided by the tool. The cognitive bias of the radiologist toward confirming their previous suspicion of BPH influences their MRI interpretation.
Demographic bias	Radiologists utilize an AI tool to analyze abdominal CT scans. If the AI model is trained on datasets that primarily includes younger patients, the AI tool may not be effectively trained to recognize age-related conditions commonly found in older individuals, such as diverticulosis. Consequently, when presented with abdominal CT scans from older patients, the model may experience difficulty in accurately identifying and assessing these pathologies, due to age-related demographic bias.
Feedback loop bias	Radiologists rely on an AI algorithm to assist in analyzing brain MRI scans. If the algorithm is initially trained on datasets mostly featuring images with clear and prominent lesions, such as large tumors, it may develop a bias toward identifying these abnormalities with high accuracy. Users of this tool may subconsciously prioritize confirming the presence of these well-defined lesions, providing feedback that reinforces the AI's accuracy in detecting such cases. Consequently, the model may improve its performance at identifying large lesions while potentially ignoring smaller, subtler, early-stage abnormalities, especially if they were underrepresented in the initial training data. This feedback loop between the AI model and the end-users, such as radiologists, can perpetuate bias, leading to a situation where the AI becomes increasingly adept at detecting certain types of abnormalities while potentially missing others.

be properly formulated so that it can be effectively translated into a fitting task for AI.[22] Any flaw in these essential starting points inevitably generates a bias in the subsequent steps, such as the selection of training datasets, AI model development, and/or deployment.

Bias in the dataset collection and preparation phases can significantly affect the outcomes of AI systems, particularly in the critical domain of medical imaging. This bias can stem from a variety of sources and can lead to disparities in the performance of AI systems across different patient groups, potentially exacerbating existing health inequalities.[23]

One of the primary sources of bias in medical imaging datasets is demographic imbalance. For example, if a dataset predominantly consists of images from a particular racial or ethnic group, the AI model trained on this dataset may exhibit reduced accuracy when applied to individuals from other groups. This situation can lead to misdiagnoses or delayed diagnoses for underrepresented groups. Similar issues arise with gender, age, and socio-economic status, where AI systems may perform better for the demographic groups that are overrepresented in the training data (Figure 4).[24]

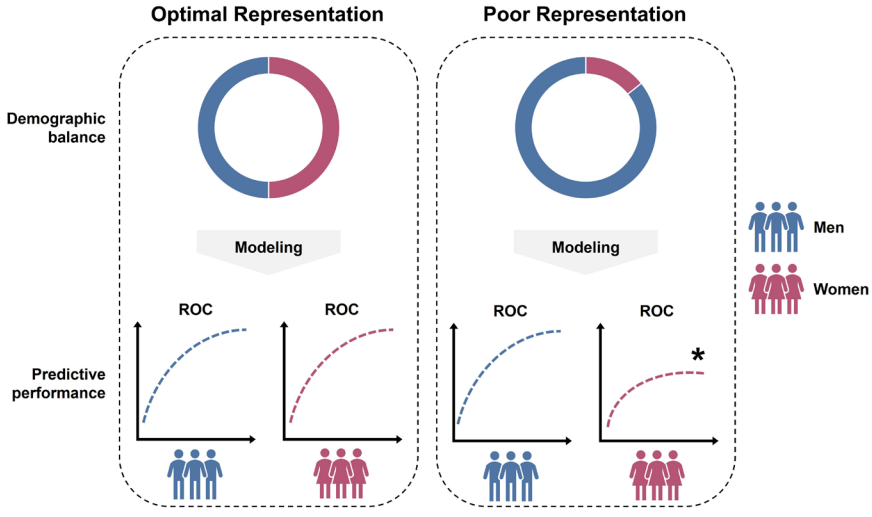

Figure 4. Over-simplified illustration of optimal and poor representation of subgroups, such as gender in this case, and their effect (*) in subsequent modeling. ROC, receiver operating characteristics.

Another critical aspect is the quality and source of the medical images. Bias can be introduced if the images come from a limited number of institutions or geographic locations, as different places may use varying equipment, protocols, and standards for image capture. This can ultimately contribute to covariate shifts (distributional differences of features between training and test sets) (Figure 5). Such variations can cause AI systems to become overfitted to the characteristics specific to the data they were trained on, reducing their generalizability and effectiveness when deployed in different settings.

The preparation of datasets also introduces potential biases (Figure 6). The process of labeling medical images, which is often performed by human experts, can lead to inconsistencies due to subjective interpretation

of what the images represent and in turn to annotation bias. Moreover, if a small group of experts annotates the dataset, their individual biases and level of expertise can influence the labels, affecting the AI model's learning process. A broader concept than annotation bias is reference standard bias, affecting the way instances are labeled and consequently impacting algorithm development.[25] Different reference standards are often available to confirm radiological diagnosis, which may also lead to systematic errors.[26] Some could be highly accurate but also costly and poorly available, whereas others could neglect intermediate findings or be operator-dependent,[27] potentially reducing label applicability and reliability. Additionally, the choice of data preprocessing techniques, such as normalization, augmentation, or cropping, can also influence the model's output by emphasizing certain features over others.[28]

Moreover, bias can stem from broader historical and societal inequities that are reflected in the data. For example, certain diseases may be more prevalent in specific populations due to factors such as access to healthcare, environmental exposures, or genetic predispositions. If these factors are not adequately considered during dataset collection and AI model training, the resulting models may not only perpetuate but also amplify existing disparities.

Bias related to modeling

The development of AI models is a multistep process, and different AI algorithms are frequently employed at different stages, such as image segmentation, feature reduction, and selection.[29] Therefore, potential bias present in any of the algorithms will propagate down the pipeline and be inherited by the final model or even amplified in it, resulting in propagation bias. It should also be considered that, since humans are developing AI models, the latter can also inherit cognitive bias from the former.[3] This is not specific to the model development stage alone and can potentially occur at any point in the AI lifecycle (Figure 7).[30]

AI modeling also includes a validation step, necessary to confirm the performance of the algorithms before actual deployment. This should ideally be verified on publicly available benchmark datasets to ensure a common ground for model testing, as seen in AI challenges. Nevertheless, further testing on independent data remains pivotal to verify that all requisites for deployment are met. In this context, a common and serious source of bias in model validation lies in data leakage.[31] An example of data leak in medical imaging is represented by the inclusion of different scans from the same patient both in the training and validation dataset, which increases the risk of overfitting.

Another aspect to carefully consider is the choice of metrics used to estimate the model's performance, which could introduce bias if those selected do not match the information needed. A case example is the validation of automated segmentation tools, for which specific parameters should be selected based on the segmentation task characteristics (e.g., is it more important to have an accurate segmentation or a precise localization for the task?).[32]

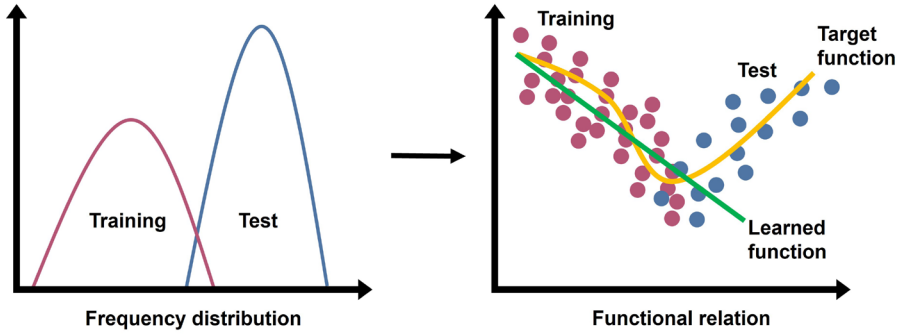

Figure 5. Over-simplified illustration of covariate shift. Distributional differences between training and test sets lead to poor test performance (i.e., poor generalizability) or significant deviation from the learned function.

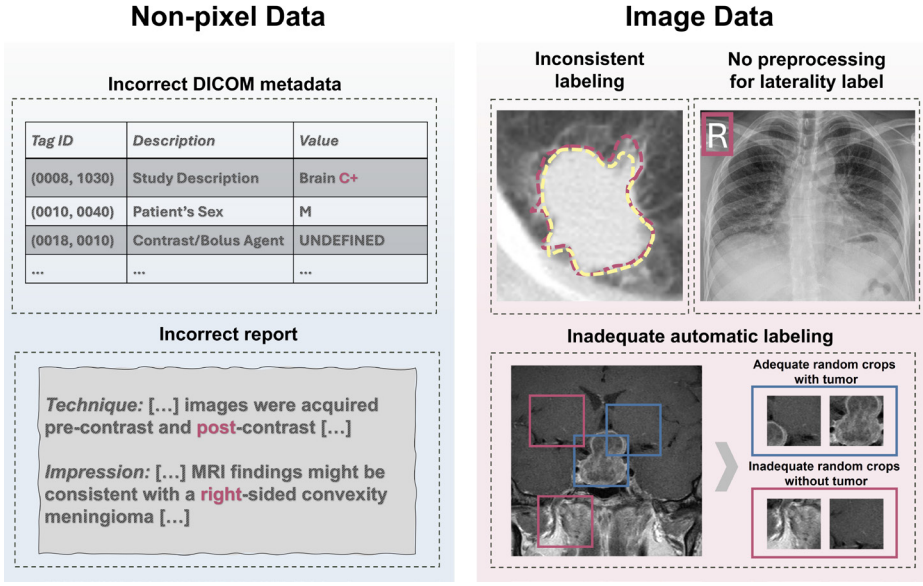

Figure 6. Potential and practical bias sources relevant to medical imaging artificial intelligence based on data type (i.e., non-pixel and image data). Radiological images belong to chest computed tomography (upper left panel), chest X-ray (upper right panel), and pituitary magnetic resonance imaging (lower panel).

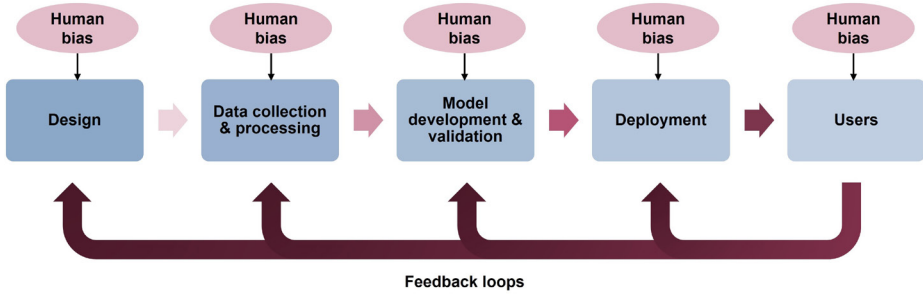

Figure 7. Human bias in the artificial intelligence life cycle.

Finally, the model's performance needs to be put into context, correctly selecting valid baseline alternatives for comparison, such as already recognized diagnostic tests, and formally evaluating with statistical approaches the added value that the model may bring.[33]

Bias related to deployment

Model deployment represents the final phase of AI/ML algorithms for medical imaging, following data collection and evaluation.[34] It involves assessing the model's performance in real-world scenarios, including potential application in clinical practice.[35]

A deployment bias emerges when there is a misalignment between the envisioned purpose of a system or algorithm and its actual application.[36] In medical imaging, this bias can manifest when an algorithm designed for segmentation tasks is utilized by human operators, whether intentionally or inadvertently, as a detection tool instead.[37] Additionally, improper utilization by end-users can also arise when utilizing systems to analyze images from anatomical districts or imaging modalities that differ from those they have been trained and validated with-for example, employing abdominal computed tomography images instead of abdominal magnetic resonance images.

Concept drift represents an additional source of bias for model deployment (Figure 8). Specifically, it arises when the correlation between input variables, such as images, and output predictions, such as diagnoses, evolves due to fluctuations in data, such as variations in image acquisition hardware or protocols, shifts in disease prevalence, or advancements in gold-standard technologies.[38]

Behavioral bias pertains to the potential distortions in user behavior seen across various platforms, contexts, or datasets.[39] Factors such as past experiences, social stigma, exposure to misinformation, limited healthcare access, and historical context play a role in shaping this bias. In particular, this bias can lead to skewed data cohorts, incomplete information, heightened uncertainty in outcomes, and potential dismissal of algorithm-assisted medical advice.[40]

Uncertainty bias encompasses the influence of uncertainty on decision-making stemming from AI/ML models.[39] Precisely characterizing and estimating uncertainty is pivotal in ensuring the thorough evaluation and transparent reporting of AI/ML models. Nonetheless, human observer decisions relying on AI/ML model outputs and their reported uncertainty may be unduly swayed by the uncertainties inherent in the model's output.[41] Consider this scenario: AI/ML models can be "confidently wrong," meaning they may yield incorrect outcomes with a high level of certainty. Consequently, humans may place greater importance on a prediction that exhibits high certainty, even if it happens to be incorrect, compared with one with lower certainty that is actually correct.

Automation bias refers to the tendency of individuals to rely excessively on automated systems, such as AI algorithms, and to disregard or underutilize their own judgment or critical thinking skills.[42] In the context of AI in medical imaging, automation bias can manifest when clinicians or radiologists place undue trust in the outputs or recommendations provided by AI algorithms, leading them to overlook potentially important information or make errors in diagnosis or treatment planning.[43] Automation bias can occur in busy clinical settings where clinicians may feel pressure to make rapid decisions, leading them to rely on AI-generated results as a shortcut rather than engaging in thorough analysis.[44] Additionally, clinicians may tend to seek out or interpret information in a way that confirms their preexisting beliefs or expectations. If an AI algorithm's recommendation aligns with their initial impressions, they may be more likely to accept it without question. A lack of adequate training or education on how to effectively integrate AI algorithms into workflow may favor automation bias.[45]

Algorithmic aversion refers to a phenomenon where clinicians or healthcare professionals exhibit reluctance or skepticism toward relying on AI algorithms for making diagnostic or treatment decisions in medical imaging.[46] This bias can manifest due to several reasons, such as trust issues on algorithms' reliability, transparency, or interpretability or a lack of familiarity, fear of job displacement, or even ethical and legal concerns.

Bias detection/identification

Detecting bias in AI algorithms necessitates awareness of all sources of bias, including those that have to do with the dataset and the development and evaluation of AI algorithms as well as those related to the deployment of these algorithms, such as human user biases and inference. Methods for bias detection vary according to the type of bias. One of the first strategies that can be used to identify bias related to the dataset is dataset evaluation against a set of predefined criteria (searching for exclusion bias, selection bias, recall bias, observer bias, and prejudice bias) and comprehensive data analysis.[47] Unsupervised analysis of the training dataset, using methods such as principal component analysis and hierarchical clustering, can be used for the detection of patterns in the training dataset that may be otherwise occult, highlighting data skewness. Statistical comparison of model output according to different patient groups or confounders that may exist in the training dataset, such as the gender or age of patients. Potential discrepancies in group results could indicate a source of bias that can affect the final results.[48] Visualization of algorithm output with methods such as class activation heatmaps can help detect discrepancies related to such potential confounders.

The next step in bias detection is the evaluation of the model development process. This starts with a code review that can be

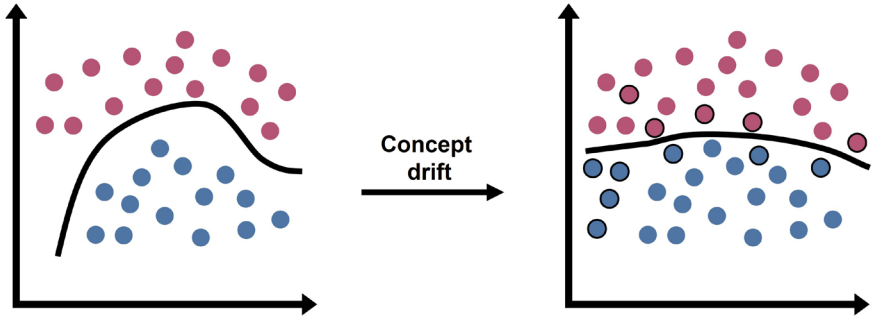

Figure 8. Over-simplified illustration of true concept drift while adding new data over time, resulting in changes in the relationship of input features and the target variable and ultimately in model behavior. In medical imaging, this may result from, for instance, a change of reference standard (e.g., new guidelines) in determining the target variable or a difference in the distribution of underlying data. It is also possible that such changes, particularly changes in data distribution, may result in virtual drifts with no obvious difference in model behavior.

carried out by an independent experienced coder/auditor.[49] Companies such as Google have developed methods for anonymous code review by several experts.[49] Such a code review can be also performed retrospectively by the scientific community for manuscripts published with open-access code.[50] Once the code has been scrutinized for potential bias, comprehensive testing should be initiated. This testing should extend from the evaluation of model performance in populations unseen in the training dataset (e.g., assessment of model performance in a pediatric population even though the algorithm was not trained with child data) to explainability analysis.[51] Simulation methods testing algorithms in various scenarios with Bayesian parameter search have been proposed to identify bias sources of algorithmic performance reduction.[52] Several explainability methods have been used that include saliency maps, such as gradient-weighted class activation mapping (CAM) and integrated gradients. Evaluation of the results of saliency maps necessitates extra care, as concerns have been raised about the reliability of these methods.[53,54]

To detect bias related to the use of the developed algorithm, human factors as well as economic, ethical, and legal factors need to be evaluated. Testing by a variety of user groups with variable experiences and backgrounds can identify human user bias. Receiving feedback with user interviews and monitoring the results per user group can help locate performance outliers or imbalances related to human factors. In addition, deep learning systems that reduce the variability in human actions leading in turn to bias reduction can be useful.[55] Auditing by legal and ethics experts can also reveal issues related to the successful deployment of the model.[56,57]

To identify and flag bias in AI publications, tools have been developed to assist the writing process of AI manuscripts.[58,59] One of these tools is the Prediction Model Risk of Bias ASsesment Tool (PROBAST), which was developed in 2019 to enable the critical evaluation of studies presenting predictive models. The current version of PROBAST evaluates the risk of bias in four potential bias categories: participants, predictors, outcomes, and analysis.[60] Nonetheless, the current version of PROBAST is not suitable for the evaluation of ML studies, and this is the reason that the PROBAST group has initiated the process of developing an AI-specific version of PROBAST called PROBAST-AI, which is still under development.[61] For systematic reviews of AI studies, the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) has been widely used to detect the risk of bias.[62] The QUADAS-2 tool includes 14 questions and provides an estimate of the risk of bias in the study, categorizing it as high, low, or unclear. Reporting guidelines, such as the Fairness Universality Traceability Usability Robustness Explainability-AI and TRIPOD-AI, can assist authors of AI manuscripts in reporting their studies according to the Fairness principle, promoting the identification of bias sources.[58,63,64] When dealing with radiomics studies, the CheckList for EvaluAtion of Radiomics (CLEAR) and METhodological RadiomICs Score (METRICS) have been developed to evaluate the reporting and methodological study quality.[65,66] Among the items evaluated, CLEAR item#7 and METRICS item#1 require adherence to reporting guidelines similar to those mentioned above; CLEAR item#36 and METRICS item#19 require the consideration of confounding factors related to dataset preparation that are closely related to bias.

Avoidance strategies

Ideally, bias should be prevented before it becomes embedded within AI systems. The focus of strategies employed during the planning, data collection, and model training phases of creating AI systems is on prevention, setting a course that avoids the pitfalls of bias rather than correcting for it post-hoc.

To mitigate bias and potentially avoid it, medical AI system development should adhere to ethical AI design principles. Guiding principles, such as transparency, fairness, non-maleficence, and respect for privacy from the outset, are widely included in recommendations and position papers and can help to prevent bias (Figure 9).[67] Transparency increases explainability, interpretability, and similar acts of communication and disclosure, which in the context of bias mitigation applies to the explicit, proactive thought about which training data are used, and how they are collected, processed, and employed. Fairness refers to an impartial treatment without favoritism or discrimination. In the context of preventing bias, fairness can be pursued by creating and upholding design standards that respect diversity, equity, and inclusion. Non-maleficence is a core medical principle. AI systems should never cause foreseeable or unintentional harm, for instance through discrimination or suboptimal patient management, which can be a direct result of biased models.[13] Respect for privacy is an important ethical principle, particularly in healthcare. In the context of mitigating bias, upholding this principle requires careful risk-benefit analyses to balance incorporating more data with the need to provide individuals control over their own data.

By incorporating the above-mentioned considerations early into the design phase, developers can create systems that are less likely to perpetuate or amplify biases. This involves rigorous ethical review processes and early stakeholder consultations to guide the decision-making process. The composition of the involved teams can influence the AI's propensity for bias. Teams that are diverse in terms of gender, ethnicity, culture, and professional background bring a wide array of perspectives to the table, which can help identify and eliminate potential biases early in the development process.

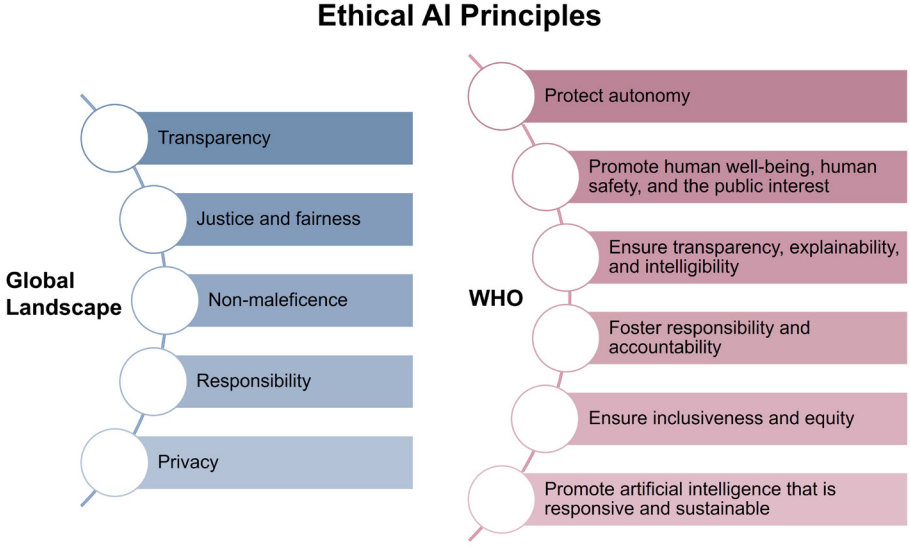

Figure 9. Key ethical artificial intelligence principles. WHO, World Health Organization.

AI systems may transport various types of bias stemming from their underlying training data.[68,69] At the data collection and processing phase, these include measurement bias (how particular features are chosen, used, or measured), omitted variable bias (when one or more relevant variables are omitted or context is neglected), representation and sampling bias (incorrect sampling leads to insufficiently diverse or otherwise non-representative datasets), and aggregation bias (false conclusions about individuals from observing whole populations).[69,70] These issues warrant thoughtful data collection and processing to ensure that datasets are representative of the diversity of the population or phenomena they are intended to model. It requires sourcing data from a wide range of demographics, geographies, and contexts to capture a broad spectrum. Nonetheless, even data collected following these principles may still reflect existing structural and historical biases.

Apart from collecting more data, strategies at the data processing stage may include the creation of more representative training datasets by data augmentation (e.g., by specifically adding underrepresented examples to the data through additional sampling or data generation) or data filtering (e.g., actively undersampling or filtering out undesirable or non-representative samples).[68] Generative AI models, such as large language models or vision-language models capable of synthesizing images, additionally allow for tailored data augmentation by creating new examples that meet a set of targeted criteria.[71-73] An overview of bias avoidance strategies at the data processing phase is presented in Figure 10.

The way data are presented to the model during training (affected by the problem formulation and the labeling methodology) and how model parameters are updated (defined through training setup including the objective function) can introduce bias into the model.[13,68] A classic example is optimizing a model for overall accuracy, which may severely impact the model performance on minority class samples in imbalanced setups. Other techniques, such as pruning, aiming to compress the model may also disproportionately impact underrepresented subsets in the data.[74] Careful design of the training setup can help avoid biases at this stage.

Transparent and comprehensive documentation of the AI system's design choices, data sources, and any assumptions made during development (e.g., through model cards)[75] is crucial and can help spot sources of bias before, during, and after training. Additionally, especially in the context of foundation models, detailed documentation may help developers seeking to use larger models' outputs to train smaller models to prevent propagating bias existing in the teacher model to downstream models.

Mitigation strategies

This section reviews different approaches and algorithms to mitigate biases. Bias mitigation algorithms can be divided into three types according to the phase in which they are applied: in a preprocessing phase, during model training, or after model training.[76] Additionally, algorithms can be categorized according to whether they explicitly or implicitly address bias by accessing or not accessing the bias variables during training.[77]

The bias mitigation algorithms applied in the preprocessing phase are motivated by the fact that many of the errors in ML models arise from biases inherent in the data used to train them. Additionally, these are independent of the model and can be used in a black-box setting by altering the data distribution to increase model fairness.[76] To achieve this effect, discriminatory effects within data are first quantified and then removed or accounted for. Several specific mechanisms for handling discrimination have been proposed to create a fair training distribution.[76]

Re-sampling and re-weighting algorithms focus on rebalancing the class distribution by adjusting the sample probability/loss weight for majority/minority samples.[78-83] Nabi and Shpitser[84] rely on causal inference to estimate the effects of specific variables on the outcome, allowing them to transform the inference problem on a specific distribution into another fair distribution to train the model. Despite addressing what can be considered the root of the fairness issue, this approach may need unrealistic assumptions about the training distribution or result in the loss of information that is implicit in the original data.

Other algorithms, such as distributionally robust optimization[85] and variations,[86] ensembling approaches,[87-89] adversarial debiasing,[90-95] invariant risk minimization,[96] invariant causal predictors,[96,97] limited capacity models,[98-100] and gradient starvation mitigation,[101] have been proposed to mitigate bias during model training by updating the objective function or imposing constraints on the model, with the last two methods implicitly achieving this.[77]

Finally, another set of methods mitigates bias in a post-processing phase after model training by changing prediction based on fairness constraints.[76] Hardt et al.[102] proposed a methodology for achieving equalized odds and equality of opportunity, whereas Pleiss et al.[103] proposed calibrated equalized odds. Woodworth et al.[104] used equalized odds to propose learning non-discriminatory predictors, and Kamiran et al.[105] used decision theory to suggest reject option-based classification and discrimination-aware ensemble for discrimination-aware classification. Lohia et al.[106] proposed a post-processing method for individual and group debiasing. These post-processing methods can be used in black-box settings, similar to preprocessing methods, as they do not require access to model parameters.[76]

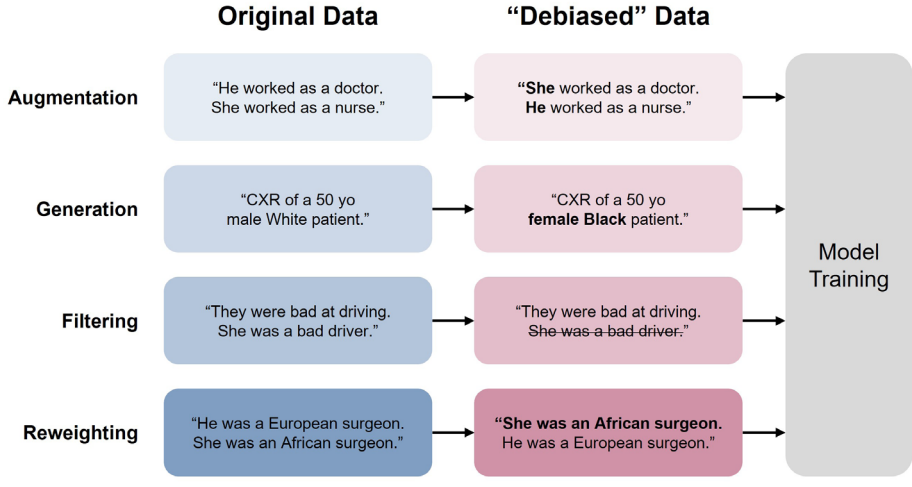

Figure 10. Overview of bias avoidance strategies at the data processing phase. Adapted from Gallegos et al.[68] CXR, chest X-ray.

In addition to active bias mitigation techniques, explainable artificial intelligence (XAI) methods offer insights into the key features influencing a model's predictions and identify and understand the significance of features driving a model's decisions. This understanding is crucial for uncovering limitations and biases in AI applications within medical imaging. These methods help us discern if confounders or biases are present in the model, allowing for their control or removal.[107] In general, XAI methods can be categorized into two main groups: perturbation-based and backpropagation-based explanations. Perturbation-based methods include occlusion,[108] LIME,[109] SHAP,[110] and various forms of perturbations.[111-113] Backpropagation-based methods encompass well-known techniques, such as saliency map visualization,[114] CAMs,[115] and their extensions.[116-118]

Potential challenges

Handling bias in AI systems is crucial for ensuring fairness and equity in decision-making processes. However, there are several challenges in handling bias that can be related to ambiguities in interpreting results, limited diversity in benchmark datasets, and the subjectivity of detecting bias.

Ambiguities in interpreting results can pose significant challenges in the development and clinical use of AI software. These refer to situations where the interpretation of the results is not unique or is open to multiple meanings by the users. Ambiguities can also originate during the different applications of the AI tools from the intended use statement provided by the AI developers, increasing the risk of off-label or erroneous applications of AI in clinical practice.[119] For example, AI software trained for adult fracture detection is at risk of erroneous results if applied in a pediatric population.

Limited diversity in benchmark datasets can represent a significant challenge in AI development and generalizability. This can occur when some diseases or events are collected with underrepresentation or overrepresentation compared with their prevalence in the general population or clinical practice due to the limited patient diversity included in the training data; this causes a class imbalance due to an uneven distribution between the training data and the actual population to which the AI model is applied.[120] As AI tools learn from archival data, a narrowed data source results in AI models that are not generalizable in heterogeneous patient populations with different demographics, clinical characteristics, and disease prevalence, leading to perpetuated bias in the final AI model.[120,121] Publicly accessible benchmarks are essential for comparison for AI models and represent a crucial element of open science.[122] Multicentric databases can potentially overcome this challenge by collecting a large number of diverse and representative data in rarer conditions. Currently, these publicly available datasets are limited to a narrow spectrum of diseases or countries of origin of the patient population.[123] Different demographic and clinical characteristics should be included to ensure a real-world representation in benchmark datasets.[48] However, although sharing data is essential for developing robust AI tools, patient privacy when collecting medical information can pose significant challenges.[124] Furthermore, real-world data are affected by missing or incomplete clinical values in retrospective cohorts and heterogeneity of clinical and laboratory parameters with their standard of reference. Image quality, noise, and acquisition parameters represent additional challenges in handling bias in multicentric cohorts. In the current radiological literature, there are ongoing difficulties in sharing benchmark datasets, with fewer than approximately 6% of all published articles in radiology journals partially or completely sharing the experimental data used to build the AI models.[125] Finally, data labeling for model training can be affected by the human image interpretation and diagnostic performance of the selected reference standard for the investigated condition.[121]

Identifying the source of bias in AI tools is also a relevant challenge. Subjectivity in the detection of bias can be related to personal interpretation and individual perspectives related to the identification of the bias itself. The complexity of AI tools makes it difficult to detect. Moreover, different sources of bias can contribute to the generation of bias, including the data source, algorithm, and users, which makes the identification more cumbersome.[124] Ultimately, identifying and addressing bias in AI will require significant effort for algorithm transparency, data source and processing, and final model utilization.

Ethical considerations

Ethical considerations are important in all steps of the AI pathway, from identifying a use case to post-market surveillance. It is important to ensure the technology promotes well-being, minimizes harm, and distributes benefits and harms justly among all stakeholders.[126] The World Health Organization (WHO) poses six key ethical principles for AI in healthcare in their framework (Figure 9): (1) protect autonomy, (2) promote human well-being, human safety, and the public interest, (3) ensure transparency, explainability, and intelligibility, (4) foster responsibility and accountability, (5) ensure inclusiveness and equity, and (6) promote AI that is responsive and sustainable.[127]

The WHO principles 2 and 5 address bias, mandating that AI tools prioritize human well-being, safety, and public interest. Ensuring AI's safety and efficacy in medical imaging demands rigorous testing, validation, and ongoing monitoring to mitigate harms and biases. Cost-effectiveness analyses and environmental awareness are both crucial to prevent unnecessary burdens on society, patients, and our environment.

Addressing biases in AI, particularly those affecting inclusivity and equity based on gender (identity), ethnicity, and socio-economic status, requires thorough subgroup analyses. The 2020 Dutch case against the "system risk indication" tool, which violated privacy laws and wrongly identified innocent people as fraud suspects, underscores the impact of such biases.[128]

Additionally, the lack of diversity among developers and researchers can worsen these issues, as teams may unconsciously favor perspectives similar to their own. Therefore, enhancing team diversity and unconscious bias training is crucial for mitigating bias in AI development.

Central to data ethics in AI are principles such as informed consent, privacy, data protection, and transparency. Currently, patients can decline being evaluated by AI-based tools according to the right to informed consent for any procedure in the hospital.[129] Patients should be given comprehensive information about how AI is used in their care, including any limitations or biases of the AI system that may affect their treatment. This may, however, eventually become infeasible when AI is deeply integrated into healthcare, and refusing AI may then compromise an individual's access to care. An alternative may then be a human-in-the-loop and a rigorous monitoring system.[130]

Ultimately, to protect patients, the ethical use of AI including mitigating biases needs to be captured in regulations. The recent European Union's AI Act serves as a pioneering legal framework aimed at regulating AI use,

particularly in high-risk applications such as healthcare (as defined in Article 6). Set to fully take effect in 2026, the act governs the development, deployment, and use of AI, ensuring safety, transparency, and adherence to ethical standards across the EU. Article 10 mandates that for high-risk AI systems, training, validation, and testing datasets must be relevant, representative, error-minimal, and complete for their intended use. Additionally, it requires rigorous data governance, including bias examination and mitigation measures, to prevent impacts on health, safety, fundamental rights, or unlawful discrimination, particularly when data outputs affect future inputs. Concerning monitoring, Article 61 of the legislation mandates that developers of high-risk AI systems establish ongoing, systematic post-market surveillance mechanisms. Critiques of the act highlight liability gaps and tension between its vague yet stringent stipulations, potentially stifling innovation and escalating healthcare costs through the compliance burden.[131]

Prospects

Despite the above challenges, proactive efforts are expected to avoid and mitigate bias in AI for medical imaging in the future. Addressing bias in medical imaging AI is a dynamic landscape with many opportunities for innovation. Before going into detail, it should be acknowledged that expecting completely bias-free systems may be unrealistic.

Developing new bias detection methods is a promising future direction. More sophisticated algorithms that can identify and measure biases, including subtle discrimination, may be developed by researchers. Even though AI models are assumed to be biased, AI-based bias auditing tools may be leveraged to help mitigate bias.[132,133] To reflect diverse healthcare landscapes and disparities across countries and regions, initiatives to improve diversity and representativeness in datasets, possibly globally, may support this goal.[123] Such initiatives should aim to reduce AI system biases by compiling larger and more inclusive data repositories from diverse demographic groups and geographic regions.[123]

Additionally, bias or fairness-aware algorithms for medical imaging applications may be promising.[134] These algorithms can ensure equitable outcomes across patient populations. Because collaboration across disciplines is key to progress in this field, experts from computer science, medicine, ethics, and policymaking can collaborate to address bias in AI medical imaging from multiple perspectives.[39] Resultant algorithms must be explainable with transparent methods so these can be further studied and debated in the future.[135-137] AI companies should be encouraged to actively participate in independent research on AI biases and algorithms to improve fairness.

After training, an AI algorithm can be locked or adaptive.[138] Instead of becoming outdated after a few years, the AI model could be updated continuously as it learns from new data. Continuous learning can increase bias if the new data are biased.[139] Continuous monitoring of models should address biases that may arise over time to ensure the integrity of AI medical imaging systems in real-world clinical settings.[10,48,140] By identifying and addressing biases, these systems can improve healthcare outcomes and equity. Independent experts or organizations can audit these regularly.

AI system development and deployment in healthcare should require adherence to certain ethical guidelines and standards, which need to be improved over time considering the dynamic nature of these tools. These guidelines should explicitly deal with AI bias and fairness as well. Stronger regulatory oversight and accountability mechanisms, such as the Food and Drug Administration's action plans and the European Union's AI act, are needed to ensure that AI medical imaging systems meet bias and trustworthiness standards without hindering AI innovation.[141-143]

Final remarks

Understanding that medical imaging AI systems are sensitive to biases is key for their effective real-world integration into clinical practice. As technology progresses, the AI community should prioritize addressing bias throughout the entire AI lifecycle, starting from the research question to data collection, data processing, model development, model evaluation, and eventual real-world deployment. For this purpose, we present collective recommendations in Table 3.

Despite the aspiration for unbiased AI, complete inclusivity of all data types and sources remains an unattainable goal in model development. Nevertheless, by leveraging diverse datasets, integrating fairness-aware systems or bias assessment tools, and promoting interpretability and explainability methods, the future -and also the AI itself- may hold great promise to mitigate bias and enhance patient care outcomes. Even so, developers and clinicians must acknowledge the inherent limitations of AI methodologies and potential biases, similar to traditional diagnostic tools, to ensure the ultimate clinical decisions are based on clinical context and benefit all patients equitably. Being at the forefront of AI implementation, medical imaging professionals, particularly radiologists, are positioned to lead efforts toward unbiased AI integration in healthcare.

By offering a comprehensive review of critical aspects, but without a detailed technical discussion, we hope this review effort raises awareness within the medical imaging community about the importance of identifying and addressing AI bias proactively to prevent its impact from being realized later.

Acknowledgement

Language of this manuscript was checked and improved by Grammarly and partly by QuillBot, which are technologies powered by generative AI. The authors conducted strict supervision when using these tools.

Conflict of interest disclosure

Burak Koçak, MD, is Section Editor in Diagnostic and Interventional Radiology. He had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Roberto Cannella has received support for attending meetings from Bracco and Bayer; research collaboration with Siemens Healthcare. Christian Bluethgen has received support for attending conferences from Bayer AG, CH. He has also received research support from the Promedica Foundation, Chur, CH. Merel Huisman has received speaker honoraria from industry (Bayer); consulting fees (Capvision, MedicalPhit). Other authors have nothing to disclose.

Funding

This study has not directly received any funding. Roberto Cannella: co-funding by the European Union - FESR or FSE, PON Research and Innovation 2014-2020 - DM 1062/2021.

Table 3. Recommendations for addressing bias in artificial intelligence (AI) for medical imaging

Stage of AI	Recommendation
Design	• Ensure that the project team represents a range of perspectives, including radiologists, clinicians, data scientists, engineers, and department administrators, preferably from different demographic backgrounds. • Encourage the entire team for transparency in detecting and reporting potential biases. • Scrutinize research questions to identify any inherent biases or inequalities and address them proactively in the study design. • Consider adhering to established reporting and methodological quality guidelines to ensure transparency and reproducibility.
Data	• Collect data from a wide range of sources to capture diverse patient populations. • Conduct in-depth exploratory data analysis to identify any potential systematic errors that may exist, informing subsequent modeling and mitigation strategies. • Standardize data to ensure consistency across datasets, with effective harmonization techniques. • Implement rigorous quality control measures to maintain the accuracy and reliability of labels and annotations, following established protocols and guidelines. • Continuously monitor data quality and update annotations as needed to reflect any changes or improvements.
Modeling and Evaluation	• Divide the dataset into training, validation, and test sets before any modeling begins, ensuring that each subset is representative of the overall population. • Select evaluation metrics that account for disparities in outcomes across different demographic groups, avoiding metrics that may mask underlying systematic errors. • Consider techniques such as fairness-aware machine learning algorithms and model interpretability methods to mitigate bias and enhance transparency. • Evaluate model fairness using a variety of methods to capture different aspects of bias. • Assess model performance separately for different demographic subgroups to identify any disparities in predictive accuracy or bias. • Continuously retrain and update models to account for evolving datasets and mitigate the perpetuation of historical biases.
Deployment	• Continuously monitor model performance in real-world settings, paying particular attention to disparities in outcomes among different demographic groups. • Conduct thorough evaluation of model performance after any updates or modifications to ensure that biases have not been inadvertently introduced or amplified. • Engage with regulatory bodies to ensure compliance with relevant standards and guidelines and seek periodic audits to validate the fairness and effectiveness of the deployed models. • Try to collect effective feedback from the end-users to identify potential biases or shortcomings in the deployed system and address them promptly.

References

1. Hammond MEH, Stehlik J, Drakos SG, Kfoury AG. Bias in medicine: lessons learned and mitigation strategies. *JACC Basic Transl Sci*. 2021;6(1):78-85. [CrossRef]
2. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. *Radiology*. 1988;167(2):565-569. [CrossRef]
3. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*. 2018;38(1):236-247. [CrossRef]
4. Gopal DP, Chetty U, O'Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. *Future Healthc J*. 2021;8(1):40-48. [CrossRef]
5. Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *Lancet Digit Health*. 2020;2(9):e486-e488. [CrossRef]
6. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*. 2023;13(17):2760. [CrossRef]
7. Tang X. The role of artificial intelligence in medical imaging research. *BJR Open*. 2019;2(1):20190031. [CrossRef]
8. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pac Symp Biocomput*. 2021;26:232-243. [CrossRef]
9. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27(12):2176-2182. [CrossRef]
10. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322(24):2377-2378. [CrossRef]
11. Vrudhula A, Kwan AC, Ouyang D, Cheng S. Machine learning and bias in medical imaging: opportunities and challenges. *Circ Cardiovasc Imaging*. 2024;17(2):e015495. [CrossRef]
12. Banerjee I, Bhattacharjee K, Burns JL, et al. "Shortcuts" causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. *J Am Coll Radiol*. 2023;20(9):842-851. [CrossRef]
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. [CrossRef]
14. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Comput*. 1992;4(1):1-58. [CrossRef]
15. Doroudi S, Rastegar SA. The bias-variance tradeoff in cognitive science. *Cogn Sci*. 2023;47(1):e13241. [CrossRef]
16. Doroudi S. The bias-variance tradeoff: how data science can inform educational debates. *AERA Open*. 2020;6(4):2332858420977208. [CrossRef]
17. Guan X, Burton H. Bias-variance tradeoff in machine learning: theoretical formulation and implications to structural engineering applications. *Structures*. 2022;46:17-30. [CrossRef]
18. Bouchard G. Bias-variance tradeoff in hybrid generative-discriminative models. In: *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*. Cincinnati, OH, USA. 2007:124-129. [CrossRef]
19. Rocks JW, Mehta P. Memorizing without overfitting: bias, variance, and interpolation in overparameterized models. *Phys Rev Res*. 2022;4(1):013201. [CrossRef]
20. Belkin M, Hsu D, Ma S, Mandal S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc Natl Acad Sci*. 2019;116(32):15849-15854. [CrossRef]
21. Gong J, Currano RM, Sirkin D, Yeung S, Holsinger FC. NICE: Four Human-Centered AI principles for bridging the AI-to-clinic translational gap. In: 2021. Accessed April 8, 2024. [CrossRef]
22. Koçak B, Cuocolo R, dos Santos DP, Stanzione A, Ugga L. Must-have qualities of clinical research on artificial intelligence and machine

learning. *Balkan Med J*. 2023;40(1):3-12. [CrossRef]

23. Chen P, Wu L, Wang L. AI fairness in data management and analytics: a review on challenges, methodologies and applications. *Appl Sci*. 2023;13(18):10258. [CrossRef]
24. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci U S A*. 2020;117(23):12592-12594. [CrossRef]
25. Jayakumar S, Sounderajah V, Normahani P, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. *NPJ Digit Med*. 2022;5(1):11. [CrossRef]
26. van der Pol CB, McInnes MDF, Salameh JP, et al. Impact of reference standard on CT, MRI, and contrast-enhanced US LI-RADS diagnosis of hepatocellular carcinoma: a meta-analysis. *Radiology*. 2022;303(3):544-545. [CrossRef]
27. Chen PC, Mermel CH, Liu Y. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. *Lancet Digit Health*. 2021;3(11):e693-e695. [CrossRef]
28. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. *Med Image Anal*. 2020;63:101693. [CrossRef]
29. Stanzione A, Cuocolo R, Ugga L, et al. Oncologic imaging and radiomics: a walkthrough review of methodological challenges. *Cancers (Basel)*. 2022;14(19):4871. [CrossRef]
30. Ashmore R, Calinescu R, Paterson C. Assuring the machine learning lifecycle: desiderata, methods, and challenges. Published online May 10, 2019. [CrossRef]
31. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y)*. 2023;4(9):100804. [CrossRef]
32. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res Notes*. 2022;15(1):210. [CrossRef]
33. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med*. 2022;5(1):48. [CrossRef]
34. Kulkarni V, Gawali M, Kharat A. Key Technology considerations in developing and deploying machine learning models in clinical radiology practice. *JMIR Med Inform*. 2021;9(9):e28776. [CrossRef]
35. Malerbi FK, Nakayama LF, Gayle Dychiao R, et al. Digital education for the deployment of artificial intelligence in health care. *J Med Internet Res*. 2023;25(1):e43333. [CrossRef]
36. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31-38. [CrossRef]
37. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol*. 2024;42(1):3-15. [CrossRef]
38. Roland T, Böck C, Tschoellitsch T, et al. Domain shifts in machine learning based Covid-19 diagnosis from blood tests. *J Med Syst*. 2022;46(5):23. [CrossRef]
39. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *J Med Imaging(Bellingham)*. 2023;10(6):061104. [CrossRef]
40. Olteanu A, Castillo C, Diaz F, Kıcıman E. Social data: biases, methodological pitfalls, and ethical boundaries. *Front Big Data*. 2019;2:13. [CrossRef]
41. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn*. 2021;110(3):457-506. [CrossRef]
42. Khera R, Simon MA, Ross JS. Automation Bias and Assistive AI: Risk of Harm From AI-driven clinical decision support. *JAMA*. 2023;330(23):2255-2257. [CrossRef]
43. Bitterman DS, Aerts HJWL, Mak RH. Approaching autonomy in medical artificial intelligence. *Lancet Digit Health*. 2020;2(9):e447-e449. [CrossRef]
44. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc*. 2012;19(1):121-127. [CrossRef]
45. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation Bias in Mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*. 2023;307(4):e222176. [CrossRef]
46. Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. *Nat Hum Behav*. 2021;5(12):1636-1642. [CrossRef]
47. Rouzrokh P, Khosravi B, Faghani S, et al. Mitigating bias in radiology machine learning: 1. data handling. *Radiol Artif Intell*. 2022;4(5):e210290. [CrossRef]
48. Gichoya JW, Thomas K, Celi LA, et al. AI pitfalls and what not to do: mitigating bias in AI. *Br J Radiol*. 2023;96(1150):20230023. [CrossRef]
49. Murphy-Hill E, Jaspan CN, Egelman C, Cheng L. The Pushback effects of race, ethnicity, gender, and age in code review. *Commun ACM*. 2022;65(3):52-57. [CrossRef]
50. A Akinci D'Antonoli T, Cuocolo R, Baessler B, Pinto Dos Santos D. Towards reproducible radiomics research: introduction of a database for radiomics studies. *Eur Radiol*. 2024;34(1):436-443. [CrossRef]
51. Zhang K, Khosravi B, Vahdati S, et al. Mitigating bias in radiology machine learning: 2. model development. *Radiol Artif Intell*. 2022;4(5):e220010. [CrossRef]
52. McDuff D, Cheng R, Kapoor A. Identifying bias in AI using simulation. Published online September 30, 2018. [CrossRef]
53. Arun N, Gaw N, Singh P, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol Artif Intell*. 2021;3(6):e200267. [CrossRef]
54. Zhang J, Chao H, Dasegowda G, Wang G, Kalra MK, Yan P. Revisiting the trustworthiness of saliency methods in radiology AI. *Radiol Artif Intell*. 2024;6(1):e220221. [CrossRef]
55. Sengupta E, Garg D, Choudhury T, Aggarwal A. Techniques to elimenate human bias in machine learning. In: *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*. Moradabad, India: 2018:226-230. [CrossRef]
56. Naik N, Hameed BMZ, Shetty DK, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg*. 2022;9:862322. [CrossRef]
57. Chen Z. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanit Soc Sci Commun*. 2023;10(1):1-12. [CrossRef]
58. Klontzas ME, Gatti AA, Tejani AS, Kahn CE Jr. AI reporting guidelines: how to select the best one for your research. *Radiol Artif Intell*. 2023;5(3):e230055. [CrossRef]
59. Koçak B, Keleş A, Köse F. Meta-research on reporting guidelines for artificial intelligence: are authors and reviewers encouraged enough in radiology, nuclear medicine, and medical imaging journals? *Diagn Interv Radiol. 2024*. [CrossRef]
60. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med*. 2019;170(1):51-58. [CrossRef]
61. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008. [CrossRef]
62. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529-536. [CrossRef]
63. Lekadir K, Osuala R, Gallin C, et al. FUTURE-AI: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. Published online October 30, 2023. [CrossRef]
64. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. [CrossRef]
65. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al. METhodological RadiomICs Score (METRICS):

a quality scoring tool for radiomics research endorsed by EuSoMII. *Insights Imaging*. 2024;15(1):8. [CrossRef]

66. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14(1):75. [CrossRef]
67. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1(9):389-399. [CrossRef]
68. Gallegos IO, Rossi RA, Barrow J, et al. Bias and Fairness in large language models: a survey. Published online 2023. [CrossRef]
69. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on bias and fairness in machine learning. *ACM Comput Surv*. 2022;54(6):1-35. [CrossRef]
70. Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle. In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM; 2021:1-9. [CrossRef]
71. Carrillo-Perez F, Pizurica M, Zheng Y, et al. Generation of synthetic whole-slide image tiles of tumours from RNA-sequencing data via cascaded diffusion models. *Nat Biomed Eng*. 2024. [CrossRef]
72. Chambon P, Bluethgen C, Delbrouck JB, et al. RoentGen: Vision-language foundation model for chest X-ray generation. Published online 2022. [CrossRef]
73. Wiles O, Albuquerque I, Rebuffi SA, et al. Generative models improve fairness of medical classifiers under distribution shifts. Published online May 31, 2023. [CrossRef]
74. Hooker S, Moorosi N, Clark G, Bengio S, Denton E. Characterising bias in compressed models. Published online 2020. [CrossRef]
75. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM; 2019:220-229. [CrossRef]
76. Feldman T, Peake A. End-to-end bias mitigation: removing gender bias in deep learning. Published online June 20, 2021. [CrossRef]
77. Shrestha R, Kafle K, Kanan C. An investigation of critical issues in bias mitigation techniques. In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE; Waikoloa, HI, USA: 2022:2512-2523. [CrossRef]
78. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Association for Computing Machinery; 2015:259-268. [CrossRef]
79. Cui Y, Jia M, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: 2019:9260-9269. [CrossRef]
80. Zou Y, Yu Z, Kumar BVKV, Wang J. Domain adaptation for semantic segmentation via class-balanced self-training. Published online October 25, 2018. [CrossRef]
81. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: 2017:2999-3007. [CrossRef]
82. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. Hong Kong, 2008:1322-1328. [CrossRef]
83. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357. [CrossRef]
84. Nabi R, Shpitser I. Fair inference on outcomes. Published online January 21, 2018. [CrossRef]
85. Delage E, Ye Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper Res*. 2010;58(3):595-612. [CrossRef]
86. Sagawa S, Koh PW, Hashimoto TB, Liang P. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. Published online April 2, 2020. [CrossRef]
87. Cadene R, Dancette C, Ben-younes H, Cord M, Parikh D. RUBi: reducing unimodal biases in visual question answering. Published online March 23, 2020. [CrossRef]
88. Clark C, Yatskar M, Zettlemoyer L. Don't take the easy way out: ensemble based methods for avoiding known dataset biases. Published online September 9, 2019. [CrossRef]
89. He H, Zha S, Wang H. Unlearn dataset bias in natural language inference by fitting the residual. Published online November 24, 2019. [CrossRef]
90. Kim B, Kim H, Kim K, Kim S, Kim J. Learning not to learn: training deep neural networks with biased data. Published online April 15, 2019. [CrossRef]
91. Grand G, Belinkov Y. Adversarial regularization for visual question answering: strengths, shortcomings, and side effects. Published online June 20, 2019. Accessed April 8, 2024. [CrossRef]
92. Ramakrishnan S, Agrawal A, Lee S. Overcoming language priors in visual question answering with adversarial regularization. Published online November 8, 2018. [CrossRef]
93. Adeli E, Zhao Q, Pfefferbaum A, et al. Representation learning with statistical independence to mitigate bias. Published online October 8, 2019. Accessed April 8, 2024. [CrossRef]
94. Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. Published online January 22, 2018. [CrossRef]
95. Ponsiglione A, Cuocolo R. Radiology under siege? Adversarial attacks against deep learning algorithms. *Eur J Radiol*. 2023;169:111156. [CrossRef]
96. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. Published online March 27, 2020. [CrossRef]
97. Santinha J, Matos C, Figueiredo M, Papanikolaou N. Improving performance and generalizability in radiogenomics: a pilot study for prediction of IDH1/2 mutation status in gliomas with multicentric data. *J Med Imaging (Bellingham)*. 2021;8(3):031905. [CrossRef]
98. Nam J, Cha H, Ahn S, Lee J, Shin J. Learning from failure: training debiased classifier from biased classifier. Published online November 23, 2020. [CrossRef]
99. Sanh V, Wolf T, Belinkov Y, Rush AM. Learning from others' mistakes: avoiding dataset biases without modeling them. Published online December 2, 2020. [CrossRef]
100. Utama PA, Moosavi NS, Gurevych I. Towards Debiasing NLU Models from Unknown Biases. Published online October 13, 2020. [CrossRef]
101. Pezeshki M, Kaba SO, Bengio Y, Courville A, Precup D, Lajoie G. Gradient starvation: a learning proclivity in neural networks. Published online November 24, 2021. [CrossRef]
102. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Published online October 7, 2016. [CrossRef]
103. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On Fairness and calibration. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Accessed April 8, 2024. [CrossRef]
104. Woodworth B, Gunasekar S, Ohannessian MI, Srebro N. Learning non-discriminatory predictors. Published online November 1, 2017. [CrossRef]
105. Kamiran F, Karim A, Zhang X. Decision theory for discrimination-aware classification. In: *2012 IEEE 12th International Conference on Data Mining*. 2012:924-929. [CrossRef]
106. Lohia PK, Ramamurthy KN, Bhide M, Saha D, Varshney KR, Puri R. Bias mitigation post-processing for individual and group fairness. Published online December 14, 2018. [CrossRef]
107. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: an overview for clinical practitioners - Beyond saliency-based XAI approaches. *Eur J Radiol*. 2023;162:110786. [CrossRef]
108. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Published online November 28, 2013. [CrossRef]

109. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. Published online August 9, 2016. [CrossRef]
110. Lundberg S, Lee SI. A unified approach to interpreting model predictions. Published online November 24, 2017. [CrossRef]
111. Fong R, Patrick M, Vedaldi A. Understanding deep networks via extremal perturbations and smooth masks. Published online October 18, 2019. [CrossRef]
112. Dabkowski P, Gal Y. Real time image saliency for black box classifiers. Published online May 22, 2017. [CrossRef]
113. Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017:3449-3457. [CrossRef]
114. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. Published online April 19, 2014. [CrossRef]
115. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Published online December 13, 2015. [CrossRef]
116. Wang H, Wang Z, Du M, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks. Published online April 13, 2020. [CrossRef]
117. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: improved visual explanations for deep convolutional networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018:839-847. [CrossRef]
118. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128(2):336-359. [CrossRef]
119. van Leeuwen KG, Hedderich DM, Harvey H, Schalekamp S. How AI should be used in radiology: assessing ambiguity and completeness of intended use statements of commercial AI products. *Insights Imaging*. 2024;15(1):51. [CrossRef]
120. Tejani AS, Ng YS, Xi Y, Rayan JC. Understanding and Mitigating bias in imaging artificial intelligence. *Radiographics*. 2024;44(5):e230067. [CrossRef]
121. Flory MN, Napel S, Tsai EB. Artificial intelligence in radiology: opportunities and challenges. *Semin Ultrasound CT MRI*. Published online February 2024:S0887217124000052. [CrossRef]
122. Bell LC, Shimron E. Sharing Data Is Essential for the Future of AI in Medical Imaging. *Radiol Artif Intell*. 2024;6(1):e230337. [CrossRef]
123. Tripathi S, Gabriel K, Dheer S, et al. Understanding Biases and Disparities in Radiology AI Datasets: a review. *J Am Coll Radiol*. 2023;20(9):836-841. [CrossRef]
124. Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. *Sci*. 2023;6(1):3. [CrossRef]
125. Venkatesh K, Santomartino SM, Sulam J, Yi PH. Code and data sharing practices in the radiology artificial intelligence literature: a meta-research study. *Radiol Artif Intell*. 2022;4(5):e220081. [CrossRef]
126. 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(2):436-440. [CrossRef]
127. Ethics and Governance of Artificial Intelligence for Health: WHO guidance. Published online 2021. [CrossRef]
128. SyRI legislation in breach of European Convention on Human Rights. Accessed April 11, 2024. [CrossRef]
129. Ploug T, Holm S. The right to a second opinion on artificial intelligence diagnosis—remedying the inadequacy of a risk-based regulation. *Bioethics*. 2023;37(3):303-311. [CrossRef]
130. Pruski M. AI-enhanced healthcare: not a new paradigm for informed consent. *J Bioeth Inq*. 2024. [CrossRef]
131. Duffourc MN, Gerke S. The proposed EU Directives for AI liability leave worrying gaps likely to impact medical AI. *NPJ Digit Med*. 2023;6(1):77. [CrossRef]
132. Chen Y, Clayton EW, Novak LL, Anders S, Malin B. Human-centered design to address biases in artificial intelligence. *J Med Internet Res*. 2023;25(1):e43251. [CrossRef]
133. van Assen M, Beecy A, Gershon G, Newsome J, Trivedi H, Gichoya J. Implications of bias in artificial intelligence: considerations for cardiovascular imaging. *Curr Atheroscler Rep*. 2024;26(4)91-102. [CrossRef]
134. Ferrara C, Sellitto G, Ferrucci F, Palomba F, De Lucia A. Fairness-aware machine learning engineering: how far are we? Empir Softw Eng? *Empir Softw Eng*. 2023;29(1):9. [CrossRef]
135. Palatnik de Sousa I, Vellasco MMBR, Costa da Silva E. Explainable artificial intell. *Sensors (Basel)*. 2021;21(16):5657. [CrossRef]
136. Theunissen M, Browning J. Putting explainable AI in context: institutional explanations for medical. *Ethics Inf Technol*. 2022;24(2):23. [CrossRef]
137. Alikhademi K, Richardson B, Drobina E, Gilbert JE. Can explainable AI explain unfairness? A framework for evaluating explainable AI. Published online June 14, 2021. [CrossRef]
138. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med*. 2020;3:118. [CrossRef]
139. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc*. 2020;27(12):2020-2023. [CrossRef]
140. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283-286. [CrossRef]
141. Vokinger KN, Feuerriegel S, Kesselheim AS. Continual learning in medical devices: FDA's action plan and beyond. *Lancet Digit Health*. 2021;3(6):e337-e338. [CrossRef]
142. Gasser U. An EU landmark for AI governance. *Science*. 2023;380(6651):1203. [CrossRef]
143. Mazzini G, Bagni F. Considerations on the regulation of AI systems in the financial sector by the AI Act. *Front Artif Intell*. 2023;6:1277544. [CrossRef]