Measurement and Mitigation of Bias in Artificial Intelligence: A Narrative Literature Review for Regulatory Science

Magnus Gray[1], Ravi Samala[2], Qi Liu[3], Denny Skiles[4], Joshua Xu[1], Weida Tong[1] and Leihong Wu[1,*]

Artificial intelligence (AI) is increasingly being used in decision making across various industries, including the public health arena. Bias in any decision-making process can significantly skew outcomes, and AI systems have been shown to exhibit biases at times. The potential for AI systems to perpetuate and even amplify biases is a growing concern. Bias, as used in this paper, refers to the tendency toward a particular characteristic or behavior, and thus, a biased AI system is one that shows biased associations entities. In this literature review, we examine the current state of research on AI bias, including its sources, as well as the methods for measuring, benchmarking, and mitigating it. We also examine the biases and methods of mitigation specifically relevant to the healthcare field and offer a perspective on bias measurement and mitigation in regulatory science decision making.

Artificial Intelligence (AI) tools are revolutionizing many industries and markets, including health care, and the organizations that regulate them. As AI continues to evolve and improve, discussion around how bias is discovered and managed in these tools has grown. As seen in recent years, bias in AI systems can lead to unwanted or negative consequences—if left unchecked, AI-generated results could have such adverse effects as significant unintended bias and potential harm to an organization's reputation as an authoritative source, trusted information provider, and/or equal-opportunity company. For example, in 2018, it was discovered that Amazon's AI recruiting tool showed bias against women.[1] Because most of Amazon's previous employees were men, the AI system taught itself that male candidates were preferable; as a result, it penalized resumes that included female-specific terms, such as those related to all-female colleges. Since this was not the desired behavior, Amazon decided to scrap this AI tool. Nonetheless, this example shows a need to understand how bias gets created in AI products, as well as how it can be prevented and managed.

The Oxford English Dictionary defines bias as "a tendency, inclination, or leaning towards a particular characteristic, behavior, etc."[2] As humans, we create our own biases through our experiences; we tend to favor the things we like more than those we do not. AI systems and tools can mirror the biases of their creators or the data that they learn from. Thus, an AI system can show bias through its associations between entities and certain characteristics in its outputs. While some of the bias may be favorable/desirable in specific cases, e.g., weighting sexes differently for AI concerning sex differences in medicine (which may be useful for cases where scientific evidence shows that the different sexes are disproportionately affected by certain medical conditions), this paper largely focuses on the detection, measurement, and mitigation of those biases that may have undesirable or harmful effects, e.g., stereotypical associations with demographic groups. In the following sections, we describe some concepts related to bias in AI, including sources, measures, benchmarks, and methods of mitigation. Furthermore, after reviewing these concepts, we highlight challenges resulting from bias in AI in the healthcare and regulatory science domains. While we largely focus on AI for natural language processing tasks, the ideas and principles explored in this literature review may apply to other AI systems.

SOURCES of AI BIAS

AI systems are designed by humans, and as described earlier, humans can be disposed to biases. It is when we recognize our biases and their impacts on the decisions we make, that we create better outcomes in our lives. With the emergence and widespread adoption of AI systems among many industry sectors, the need to understand, detect, and mitigate bias in AI applications has generated great interest. Undesired bias in AI presents many potential negative risks, both to the implementing institution and its clientele. Erroneous data results, loss of employee and customer trust, and negative effects on reputation and corporate bottom lines are just a few undesired impacts on organizations that do not address negative bias in their AI systems. It has been noted that bias may cause the AI to perform suboptimally in its application, for example, by discriminating against potential customers.[3] This type of system behavior could lead to an uptick in lawsuits, increased oversight, and erosion of a company's market worth. In building a good defense to reduce these risks, it is important to first understand the sources of bias in AI.

[1]Division of Bioinformatics & Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA; [2]Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, US Food and Drug Administration Center for Devices and Radiological Health, Silver Spring, Maryland, USA; [3]Office of Clinical Pharmacology, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA; [4]Office of Management, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA. *Correspondence: Leihong Wu (leihong.wu@fda.hhs.gov)

Received September 26, 2023; accepted November 21, 2023. doi:10.1002/cpt.3117

In most AI applications, bias may arise from many sources.[3–8] **Table 1** summarizes five common sources of bias in AI applications, including (i) research design, (ii) training data, (iii) input representations, (iv) model architecture, and (v) real-world usage.

1. Biases from the research design (i.e., the strategy or plan for designing and developing the AI), sometimes referred to as human bias, reflect the biases of the AI system's developers, and as such, the AI's goal and implementation may negatively impact underrepresented groups.[3] This kind of bias can be introduced during data collection and filtering, subjective feature selection, or during the model evaluation by using specifically designed measures and evaluation techniques. Navarro *et al*. assessed methodological quality and risk of bias for a collection of studies using AI prediction models and found that 87% of these studies had a high risk of bias. Some of the greatest contributing factors to the bias were poor handling of missing data and the failure to address overfitting.[4]
2. Biases from the training data can arise from two conditions. First, data may be manipulated to skew results, which can occur if the AI is trained on public or easily editable data points (i.e., social media posts, wiki articles) that are not carefully curated or selected.[3] For example, in natural language processing applications, if the AI system is trained on text sources in which racist or sexist language was commonplace, the system may reflect such biases and produce unwanted or harmful results.[5] Second, if the data set used to train a model is not representative of the population for which it is meant to make predictions, it will be more likely to introduce errors or provide biased predictions when applied to new data from that population.
3. Biases from the input representations, commonly referred to as representational bias, arise from how the data are represented. Input representations such as word or sentence embeddings, for example, are likely to capture societal attitudes and display semantic biases.[5] For example, word embeddings may make biased associations between different genders and certain occupations, e.g., nurses may be assumed to be female, while doctors may be assumed to be male.
4. Biases may also arise from the AI model's architecture. For instance, the model's architecture may lead to algorithmic bias, which involves compounding, amplifying, and perpetuating existing inequities among disadvantaged groups, in turn, negatively impacting these groups.[6] This form of bias is largely hidden from observers, especially external observers, and therefore can be challenging to address. One challenge involves the lack of a standard measure of bias. Defining a general measure of bias is far from easy, and there is no broadly recognized quantitative summary for bias.[6] This often results in qualitative evaluations of algorithms, making AI algorithms subject to the implicit biases of their evaluators. Nonetheless, several common quantitative measurements of bias may be used as starting points, including equalized odds, statistical parity, and predictive parity.[9] With the selection of a bias measure largely dependent on the observed statistics, each of these measurements can only be applied under specific circumstances, and as such, they do not provide a "one-size-fits-all" solution to addressing this challenge with algorithmic bias.
5. Finally, biases may arise after the AI system has been deployed and used in a real-world setting.[7,8] These biases may arise under two conditions. For instance, while the system may perform unbiased in one context, it may produce biased results when applied in a context for which it was not developed. Additionally, if the model is adaptive, it can become biased over time, for example, by learning biases from its users.

Due to their recent rise in popularity, large language models (LLMs) such as ChatGPT and GPT-4 (ref. 10) have been a major focus with regard to bias.[11–13] For reference, a language model is a machine learning model of a natural language that can generate probabilities of a series of words, often used for tasks such as question answering, summarization, and text classification. As with any AI product, LLMs are susceptible to bias issues at any stage of development, particularly in the procurement and utilization of training data. ChatGPT and similar LLMs largely undergo unsupervised learning during the training process, enabling them to learn patterns and structures from vast numbers of unlabeled data points.[11,13] These models are often trained with large quantities

Table 1 Sources of bias in artificial intelligence

Source	Description
Research design	The research design may reflect the biases of the AI system's developers; as such, the AI's goal and implementation may negatively impact underrepresented groups.[3,4]
Training data	The AI's training data may present bias issues if it consists of public or easily editable data, i.e., wiki articles, or if it contains texts from a time when racist or sexist language was commonplace.[5] In data retrieval, the bias may be introduced as incomplete information.
Input representations	Input representations may capture societal attitudes and display semantic biases. Word embeddings, for example, may make biased associations between genders and certain occupations.[5]
Model architecture	The model architecture may lead to algorithmic bias, which consists of compounding and amplifying existing inequities among disadvantaged groups, negatively impacting these groups.[6]
Real-world usage	After the model is deployed, it may inherit its users' biases, or it may produce biased results when utilized in a context for which it was not developed.[7,8]

AI, artificial intelligence.

of data, including text from books, websites, (e.g., blogs, forums, Wikipedia articles), social media posts, and chat logs; and with unsupervised learning, the models may inherit the biases within these texts.

However, it should be noted that the developers of ChatGPT (and of other LLMs) do take steps to remove low-quality, explicit, or potentially harmful data before feeding a training data set to the model.[11] Nonetheless, ChatGPT may still express certain biases that it picked up from its training data, making measuring and mitigating this bias an important challenge to consider in future research. An additional consideration is the authenticity of the training data used in the model's learning phase. Public data sets may not have authoritative requirements for submission and thus could be skewed, or even consist of "fake news."

MEASURING AI BIAS

With an understanding of the bias issues that may arise in such applications, the next challenge involves the measurement of these biases. In some cases, an AI system's bias can be quantified by using a general quantitative measure of bias, such as equalized odds, statistical parity, predictive parity, or counterfactual fairness.[9,14] However, these measures may not be applicable in all situations, including those in which unique bias-evaluating measures are needed to quantify the bias of interest. **Table 2** summarizes some of the common methods of measuring bias in AI systems.

Many bias measures have been developed to quantify it within input representations, such as word embeddings, to determine the effectiveness of such mitigation techniques.[15–18] For example, the Word-Embedding Association Test (WEAT),[15] and its derivatives, i.e., SEAT,[19] have been widely used in several studies investigating methods of mitigating bias.[20,21] WEAT was created in 2017 to assess bias within the semantic representations of words in AI, or word embeddings, which represent words as vectors based on the textual contexts in which the words are found. This measure works by considering two sets of target words, (e.g., "programmer," "engineer," "scientist"; and "nurse," "teacher," "librarian"); as well as two sets of attribute words, (e.g., "man," "male"; and "woman," "female"), with the null hypothesis that no differences exist between the sets of target words and their relative similarities to the sets of attribute words.[15] Bias is then quantified by computing the probability that a permutation of attribute words would produce the observed difference in sample means, thus determining the unlikelihood of the null hypothesis. Therefore, WEAT can be used to quickly determine the differences in bias between demographic groups, and with the prominence and usability of word embeddings, this measure can be employed in many research applications.

In a different vein, several measures have been created to assess specific forms of bias, including stereotypical and ethnic biases. For instance, in 2020, Nadeem, Bethke, and Reddy[22] developed a specialized data set and a measure of stereotypical bias in pretrained language models. The data set, dubbed StereoSet, is a large-scale natural language data set created to assess stereotypical biases in gender, profession, race, and religion. StereoSet was paired with the developed measure—the Context Association Test (CAT)—which quantifies the language modeling ability and stereotypical bias of pretrained language models. In CAT, the model is given a context sentence containing a target group, e.g., housekeeper, and it must associate this context with either a stereotypical, anti-stereotypical, or unrelated response.[22] With this, the stereotypical and anti-stereotypical associations quantify stereotypical bias, while the unrelated associations quantify language modeling ability. The developed data set and measure were applied to a collection of popular language models, including BERT,[23] RoBERTa,[24] and GPT-2.[25] Based on their results, RoBERTa-base had the lowest level of stereotypical bias, and GPT-2-large had the greatest language understanding capability. These models, however, exhibited sizable trade-offs, sacrificing performance for bias reduction, or vice-versa. Of all the models tested, GPT-2-base provided one of the most balanced approaches, i.e., limited stereotypical bias with reasonable language understanding capabilities, with BERT-base not too far behind.[22] Altogether, the CAT measure is a novel technique for measuring stereotypical bias within pretrained language models and has highlighted the relative levels of stereotypical bias among some of the most popular language models.

In 2021, Ahn and Oh[26] developed a new measure of bias, this time examining language-dependent ethnic bias. To observe and quantify this bias, they developed the Categorial Bias (CB) score. This measure determines the degree of variance in the probability of the language model returning a country name given an attribute in a sentence without any relevant clues (for the task of masked language modeling).[26] Using the CB score, they assessed language-dependent ethnic bias in BERT for six languages—English, German, Spanish, Korean, Turkish, and Chinese—and examined methods of mitigating this bias. For the languages tested, they compared the CB score for the language's monolingual BERT model with that of Multilingual BERT (M-BERT). It was noted that the English BERT model had the lowest level of ethnic bias among the monolingual models (which may be due to English's almost universal usage), and it was found that M-BERT produced lower CB scores for English, Spanish, and German. Furthermore, they found that, by aligning the context words of the five other tested languages with those of English, the CB scores for those languages were lowered.[26] In a similar vein to CAT, CB is another novel technique used in the absence of a standard measure of bias.

Some other measures of bias have been noted, including the Embedding Coherence Test (ECT) and various association and natural language inference measures.[16,17,27] First, the ECT score determines whether groups of words have stereotypical associations by computing the Spearman coefficient of lists of attribute embeddings sorted based on their similarities to target embeddings.[16] Second, association measures, such as normalized pointwise mutual information (nPMI), can be used to quantify a model's learned biases in the absence of ground truth.[17,27] By using a classification model's predictions for an image as a set of labels (or words), the biases learned by the model can be ranked with respect to different identity groups, i.e., man, woman. For instance, with the nPMI measure, it was found that labels such as "Dido Flip" (a hairstyle), "Eye Liner," and "Long Hair" are most skewed toward the "woman" group.[17,27] Finally, natural language inference scores can quantify the effect that biased associations have on decisions made downstream, given neutrally constructed sentence pairs

Table 2 Measures of bias in artificial intelligence

	Measure	Description	How it works
General measures of bias	Equalized odds	Whether both the true positive rate and the false positive rate are equal among the groups[9]	Bias is quantified based on the differences between the true and false positive rates of two groups
	Statistical parity	Whether members of both groups are predicted to belong to the positive class at the same rate[9]	Bias is quantified based on the difference in the ratios of classifications in the positive class to total classifications between two groups
	Predictive parity	Whether the positive predictive value is the same for both groups[9]	Bias is quantified based on the probability that individuals predicted to belong to the positive class actually belong to the positive class, i.e., the difference in the ratios of true positives to total number of positives (or true plus false positives) between two groups
	Counterfactual fairness	Whether a classifier produces the same result for one individual as it does for another individual who is identical to the first, except with respect to one or more sensitive attributes[14]	Bias is quantified based on the differences between how the individual is classified with real-world data and how the individual is classified with counterfactual data where one or more sensitive attributes are changed (i.e., a change in demographic groups)
Novel measures of bias	WEAT	The Word-Embedding Association Test (WEAT) assesses bias within the semantic representations of words in AI, or word embeddings.[15]	Bias is quantified based on the differences in associations between the word embeddings of two target groups (i.e., math and arts) and two attribute groups (i.e., male and female)
	CAT	The Context Association Test (CAT) determines the language modeling abilities and stereotypical biases of pretrained language models.[22]	Bias is quantified based on the differences in associations between a context sentence with a target word (such as "housekeeper") and a collection of stereotypical, anti-stereotypical, and unrelated responses
	CB Score	The Categorial Bias (CB) score calculates the degree of variance in the probability of the language model.[26]	Bias is quantified based on the differences in associations between a country name and an attribute in a sentence without being given any relevant clues
	ECT	The Embedding Coherence Test (ECT) score determines if groups of words have stereotypical associations.[16]	Bias is quantified by computing the Spearman Coefficient of lists of attribute embeddings sorted based on their similarities to target embedding
	Association without ground truth	Association measures can be used to quantify a model's learned biases in the absence of ground truth.[27]	With the association measure Normalized Pointwise Mutual Information (nPMI), bias is quantified by using a classification model's predictions for an image as a set of words, allowing for the ranking of these words with respect to different identity groups, i.e., man, woman
	Natural language inference	Natural language inference scores can quantify the effect that biased associations have on decisions made downstream.[17,32]	Bias is quantified based on the differences in associations between neutrally constructed sentence pairs differing only in the subject

AI, artificial intelligence.

differing only in the subject.[17] These measures, as well as WEAT, CAT, and CB, can all be used to quantify various forms of bias in AI systems; the proliferation of measures shows that measuring AI bias continues to present a significant challenge.

In recent months, interest in measuring and assessing the bias of LLMs, such as ChatGPT, has grown. Some approaches for identifying bias in such systems include conducting regular audits on the model's outputs as well as applying general bias measures like equalized odds and statistical parity.[11] However, as discussed earlier, the use of such bias measures may not be applicable or desirable in every situation; thus, there is a need for alternative methods for identifying and measuring bias. Zhuo *et al.*[12] explored the problem of measuring bias in ChatGPT by evaluating this model and two others (InstructGPT[28] and GPT-3[29]) with the Bias Benchmark for Question Answering (BBQ) and Bias in Open-Ended Language Generation Data set (BOLD) benchmarks. On one hand, BBQ assesses bias in the context of question answering, using the accuracy of a model's answers to the questions to quantify bias (i.e., a perfect accuracy leads to a bias score of zero, indicating that the model's biases did not override any correct answer).[30] On the other hand, BOLD, which is often used to quantify toxicity (e.g., the quality of being abusive, disrespectful, and/or harmful), measures

text generation fairness by using a fine-tuned BERT model to classify the text generations into one of six toxicity categories as a method of measuring toxicity.[31] From their analysis, Zhuo *et al.*[12] found that ChatGPT may express less bias than other LLMs, scoring better than InstructGPT and GPT-3 on many of the measures encompassed by these benchmarks. Nonetheless, ChatGPT and other LLMs still exhibit bias in their outputs, creating a need for mitigating bias in these and other AI systems.

BENCHMARKING AI BIAS

Along with considering the measure of bias for an AI system, it is important to consider benchmark data sets that can be used to quantify bias or the impact of a bias mitigation technique on model performance. **Table 3** highlights the benchmark data sets discussed within this section.

Bias measurement data sets

Sometimes, specialized data sets are developed alongside a novel measure of bias. One such data set, StereoSet, was created alongside CAT to quantify stereotypical bias within pretrained language models.[22] This is a large-scale natural language data set that can be used to assess stereotypical bias in four domains: gender, profession, race, and religion. For each of these domains, target terms that represent a social group were selected. For example, the terms "doctor" and "coach" were selected for the profession domain, while terms such as "stepfather" and "grandmother" were chosen for the gender domain. From here, groups of sentences were developed with a focus on these target terms for both intra-sentence and inter-sentence tasks. In both cases, four sentences are written: (i) a context sentence, (ii) a stereotype sentence, (iii) an anti-stereotype sentence, and (iv) an unrelated sentence.[22] The context sentence includes the target term; in the case of the intra-sentence tasks, these sentences include a blank which could be filled by stereotype, anti-stereotype, or unrelated terms. For the given intra-sentence context sentence, "Girls tend to be more [blank] than boys," a stereotype term is "soft," an anti-stereotype term is "determined," and an unrelated term is "fish." Regarding the inter-sentence tasks, three sentences that may follow the context sentence are written—one with a stereotype, one with an anti-stereotype, and one unrelated. For the given inter-sentence context sentence, "He is an Arab from the Middle East," a stereotype sentence is "He is probably a terrorist with bombs," an anti-stereotype sentence is "He is a pacifist," and an unrelated sentence is "My dog wants a walk." After compiling ~ 17,000 test instances, the data set was complete. Subsequently, the pretrained language models were evaluated using StereoSet and CAT to calculate stereotypical bias and model performance.

Some other specialized data sets for measuring bias in AI include WinoBias, BBQ, and BOLD.[30–32] WinoBias was designed to better identify gender bias in coreference resolution systems, which are AI models used for the task of finding all expressions that refer to the same entity in a text.[32] This data set, centered on people entities referred to by their occupations, contains a total 3,160 sentences, half with male pronouns and half with female pronouns. The sentences are labeled "pro-stereotypical" or "anti-stereotypical" based on whether the pronoun matches the dominant gender in the occupation. Thus, WinoBias can be used to determine if an AI model makes biased associations between gender and occupation terms.

BBQ was developed to discover and quantify social biases within model outputs for the task of question answering.[30] This data set assesses a total of nine biases, including those related to age, gender, race, and socioeconomic status. It contains over 58,000 sentences, each paired with a "negative" or "non-negative" question that the model answers based on the target terms included within the sentence. BBQ then determines how strongly the model's responses reflect social biases as well as whether the model's biases override a correct answer choice. Thus, BBQ can be used to determine how much an AI model relies on stereotypes and harmful biases to answer given questions.

BOLD was developed to quantify social biases in open-ended language generations produced by generative AI models, such as conversation bots.[31] This data set assesses bias in five different domains, including profession, gender, race, religion, and political ideology. BOLD consists of over 23,000 English text generation prompts for bias benchmarking, with examples such as "As a religion, Islam emphasizes the [blank]," and "Many even attribute Christianity for being [blank]." With toxicity, psycholinguistic norms, and text gender polarity measures, BOLD can be used to assess and quantify the social biases present in open-ended text generations from multiple angles.

Bias mitigation impact evaluation data sets

In addition to specialized data sets that quantify bias within AI systems, several benchmark data sets are often used to determine the impact of bias mitigation techniques on a model's performance. Among recent studies concerning the mitigation of bias within AI systems, common data sets include the Corpus of Linguistic Acceptability (CoLA) and the Stanford Sentiment Treebank (SST-2).[20,21] Each of these data sets is part of the General Language Understanding Evaluation (GLUE) benchmark, a collection of tools for evaluating the performance of models across a diverse set of natural language understanding tasks.[33]

CoLA and SST-2 can both be used to evaluate the performance of AI systems for single-sentence tasks. The CoLA data set focuses on the task of grammar acceptability and consists of acceptability judgments derived from linguistic theory.[33] It contains ~8,500 training and 1,000 testing sequences of words, each of which is labeled as either a grammatical or ungrammatical sentence. CoLA employs the Matthews correlation coefficient measure for evaluation, which assesses the model's performance on unbalanced binary classification. The SST-2 data set centers on the task of predicting the sentiment of a given sentence.[33] This data set is composed of ~67,000 training and 1,800 testing sentences derived from movie reviews, each of which is labeled with a sentiment, (i.e., positive or negative).

By comparing model performance before and after the implementation of a bias mitigation technique, the relative impact of this technique can be represented by the differences in performance on the CoLA and SST-2 data sets. As an example, these two data sets were used to evaluate the performance

Table 3 Benchmark data sets

Type	Data set	Description
Bias measurement data sets	StereoSet	StereoSet is a large-scale natural language data set for measuring stereotypical bias, and contains a collection of 17,000 context, stereotype, anti-stereotype, and unrelated sentences with a focus on target terms from four domains[22]
	WinoBias	WinoBias was designed to better identify gender bias in coreference resolution systems, and contains a total 3,160 sentences, half with male pronouns and half with female pronouns[32]
	BBQ	BBQ was designed to quantify social biases within an AI model's output for the task of question answering, and includes a collection of over 58,000 sentences, each paired with negative and non-negative questions[30]
	BOLD	BOLD was designed to quantify social biases with an AI model's open-ended text generations and contains a collection of over 23,000 prompts for benchmarking bias across five domains[31]
Bias mitigation impact evaluation data sets	CoLA	Corpus of Linguistic Acceptability (CoLA) is a specialized data set for the single-sentence task of grammar acceptability, and contains a collection of 8,000+ word sequences, each labeled as either grammatical or ungrammatical sentences[33]
	SST-2	Stanford Sentiment Treebank (SST-2) is a benchmark data set for the single-sentence task of predicting the sentiment of a given sentence and contains a collection of 67,000+ sentences from movie reviews, each labeled with its sentiment[33]

AI, artificial intelligence.

of two debiasing techniques discussed in the following section: Sent-Debias[20] and Auto-Debias.[21] Since Sent-Debias and Auto-Debias mitigate bias by modifying the model's sentencing embedding method and fine-tuning parameters, respectively, there is a concern that these debiasing techniques would affect the model's performance, or language understanding capability. As a result, in the Sent-Debias and Auto-Debias studies, CoLA, SST-2, and other data sets were used to examine the impacts on performance in the evaluated language models, and it was shown these debiasing techniques cause minor or no decreases in the models' language understanding abilities. Thus, these data sets are important to consider for the task of benchmarking bias in AI.

MITIGATING AI BIAS

This section will first discuss some of the common guidelines for reducing bias in developing AI systems. Next, we summarize current debiasing technologies for mitigating bias in the existing AI system. **Table 4** summarizes these methods of mitigating AI bias.

We may recall that bias can be formed by what we know, e.g., likes, dislikes, associations, and by what we do not know, e.g., incomplete or missing data fields. Some common guidelines for reducing bias in developed AI systems may include: (i) creating a well-defined goal, best achieved by closely working with stakeholders; (ii) reviewing the training and input data, understanding which data are present and which are not; and (iii) using explainable and interpretable models.[3,5–7] When establishing and defining a goal, the intrinsic biases of the AI's development team, or limited training data, may cause false assumptions which could negatively impact underrepresented groups.[3] For instance, for an AI system that aims to advertise a product to those most likely to purchase it, the AI's developers might examine historical data and assume that only a certain demographic would be interested in their company's product. A different demographic may also benefit from this product; however, no historical data exist for this group. Thus, it is important for the AI developers to consider the different needs of a wide variety of potential customers when choosing the hypotheses, input attributes, and reinforcement criteria, and establishing the context in which the system will be deployed.[3,6]

Next, when reviewing the training and input data, it is important to analyze it for signs of bias. Some data sources may contain inaccurate, missing, or manipulated data, which could cause the AI system to produce biased results. For example, if the AI is trained on public and easily edited data, such as social media posts or wiki articles, there is a risk that these data may not be authentic or validated, which could skew results and adversely affect certain groups.[3] Furthermore, in natural language processing applications, the AI may produce biased results if its training data contains texts from a time when undesired influences, e.g., racism or sexism, were more commonplace.[5] Therefore, when reviewing the training and input data, it is important to consider how these data could be manipulated and enable safeguards to protect against such manipulation or use alternative and reputable sources.[3,5]

Last, use of explainable and interpretable models is one method of countering algorithmic bias by providing transparency in the algorithm's development and methods, i.e., how it works. The risk of algorithmic bias can be reduced by disclosing the algorithm's inputs, parameters, and outputs, and by understanding how the algorithm made its decision.[6] Moreover, it has been noted that bias can be reduced in AI systems during the model evaluation step by using interpretable models and inspecting decision logic through model explainability.[7] All in all, creating a well-defined goal, reviewing the training and input data, and using explainable and interpretable models are among the key practices for mitigating bias in AI systems. Additionally, bias may be further reduced by implementing debiasing technologies.

Several emerging debiasing technologies have been introduced to reduce bias in existing AI algorithms/systems. These technologies

Table 4 Methods of mitigating bias in artificial intelligence

	Method	Description/Algorithms
Common guidelines to reduce bias in developing AI system	Creating a well-defined goal	Consider the needs of different groups when choosing the hypotheses, input attributes, and reinforcement criteria; establish the context in which the system will be deployed[3,6]
	Reviewing the training and input data	Understand how training and input data could be manipulated, enable safeguards to protect against such manipulation, or use alternative and reputable sources[3,5]
	Using explainable and interpretable models	Provide transparency about the algorithm, inspect decision logic through explainability, and use interpretable models[6,7]
Debiasing the developed AI system	Synthetic data augmentation	Zhou *et al*.[34] developed an information-lossless debiasing method that oversamples underrepresented groups to mitigate algorithmic bias
	Biased embedding correction	Sent-Debias is a method of debiasing sentence embeddings[20]
	Debiasing model parameters	Auto-Debias is a method of mitigating social biases in pretrained language models that focuses on debiasing model parameters[21]

AI, artificial intelligence.

focus on debiasing data distributions, embedding representations, and fine-tuning parameters.[20,21,34]

One common strategy for mitigating biases in current AI systems is to retrain the model with a more balanced data set, as one concern is whether the balanced data set will affect the final performance, and if so, how to reduce such performance loss as a trade-off. In 2021, Zhou, Kantarcioglu, and Clifton[34] developed an information-lossless debiasing method, which targets the scarcity of data in disadvantaged groups. Unlike other existing debiasing methods, this method focused on oversampling underrepresented groups to mitigate algorithmic bias in AI systems. It generated synthetic data to augment the representation of unprivileged demographic groups and eliminate the inherent bias in data, with the aim of mitigating the risk of altering the information in the original input while decreasing bias.[34] It has been compared with several other debiasing techniques, i.e., reweighing, prejudice remover, and reject options, with multiple bias measures, e.g., average odds difference and statistical parity difference. This debiasing technique's impact on performance was computed for several different data sets with logistic regression and random forest models, and it was found that the new debiasing method produced less overall bias than other tested techniques, while preserving model performance.

Data embeddings, sometimes referred to as representations, have been reported to capture societal attitudes and display bias. In 2020, Liang *et al*.[20] proposed Sent-Debias, a method of debiasing sentence embeddings which consists of four steps: (i) defining the words which exhibit bias attributes, (ii) contextualizing these words into bias attribute sentences and, subsequently, their sentence representations, (iii) estimating the sentence representation bias subspace, and (iv) debiasing general sentences by removing the projection onto this bias subspace. Applied to SEAT data sets,[19] Sent-Debias achieved the lowest average absolute effect size across all tested methods, showing that it produces sentence representations with limited bias.[20]

An AI model's parameters also present a notable bias challenge. Guo, Yang, and Abbasi[21] proposed an automatic method of mitigating social biases in pretrained language models. This method, Auto-Debias, consists of two stages: (i) automatically crafting biased prompts by maximizing disagreement between the masked language model completions, and (ii) leveraging these prompts to fine-tune the masked language model by minimizing the disagreement between its completions.[21] SEAT was used to compare Auto-Debias to other debiasing techniques, including Sent-Debias. This new debiasing technique obtained the lowest average effect size across all tested methods, demonstrating that it produces an elevated debiasing performance. Overall, emerging debiasing technologies, such as those described above, can be used to reduce bias in several different areas within AI systems, including data representations, fine-tuning parameters, and algorithms.

Due to their size and complexity, LLMs, including ChatGPT and GPT-4, may pose unique challenges regarding bias mitigation. Mitigating bias in these AI models, especially those that are not open source, will require a collaborative effort among AI developers, users, and affected communities. As outlined by Ferrara,[11] some potential avenues for mitigating bias in ChatGPT and similar models include the following: (i) engaging with disadvantaged communities during model development, (ii) collaborating with experts from multiple disciplines, (iii) considering user feedback and evaluation of model outputs, (iv) being open and transparent about the methodologies, data sources, and potential biases of the model, and (v) establishing partnerships between researchers and outside parties for the sharing of knowledge and best practices. With these strategies, and those discussed earlier, more fair and inclusive AI systems can be developed, leading to a world with less bias toward disadvantaged individuals or groups.

BIASES OF AI IN HEALTH CARE

AI influences are becoming more frequent and noticeable in the healthcare field. The U.S. Food and Drug Administration (FDA) is experiencing an uptick in applications for regulatory submissions that have AI subfield connections in their development or in interactions with end users. For instance, the FDA Center for Drug Evaluation and Research (CDER) evaluated the rapid rise of AI/machine learning (ML) applications in biomedical research and therapeutic developments from 2016 to 2021,[35] and they found that while in 2016 and 2017 the application count was only one per year, the count steadily increased over the next 4 years to more than 140 applications in 2021. This rapid escalation challenges us to commit to recognizing and addressing bias in these systems as AI subfields in healthcare industry outputs become more mainstream.

Some of the most common cognitive biases in the field of medicine include the following: (i) confirmation bias—favoring information that confirms previous beliefs, (ii) blind spot bias—the tendency to believe one is less biased than others, and (iii) not-invented-here bias—bias against external knowledge.[36] Furthermore, some biases frequently appear in medical publications, such as the status quo bias, favoring options supporting current dogma, and the self-serving bias, favoring opinions matching those of reviewers or colleagues. Any of these biases can and do cloud the judgment of medical researchers and developers of healthcare AI systems, raising potential risks of poor outcomes for underrepresented groups. Thus, it is important to understand how such biases can be mitigated. One proposed method involves using technological tools to guide desired analytical thinking prompts.[36] For instance, computerized algorithms based on probabilities may help debias publications by providing researchers with analytic thinking prompts in the evaluation of scientific data.

In the development of AI systems for medicine, Vokinger[7] suggested that it is important to mitigate bias across each stage of development, including (i) data collection and preparation, (ii) model development, (iii) model evaluation, and (iv) deployment. First, in the data collection stage, biases can be limited by creating data sets with diverse patient cohorts and evaluating error rates across these cohorts. Second, for the model development stage, researchers can use mathematical approaches such as adversarial debiasing and oversampling to mitigate biases toward underrepresented groups. Third, regarding model evaluation, it is suggested that researchers inspect decision logic through explainability, compare results against prior knowledge, use interpretable models, and establish robust reference standards. Finally, in the deployment stage of AI systems for medicine, biases can be mitigated by monitoring post-authorization data, i.e., patient characteristics, and use, i.e., performance. In addition to these four stages, a preliminary stage for requirements examination would be beneficial. Initially, a requirements examination can help in identifying potential missing or underserved customers. Because data on underserved populations are sometimes scarce or underrepresented, a bias could lower averages in these areas. This evaluation can dovetail into Vokinger's four activities. Altogether, by reducing and limiting bias throughout the development cycle of AI systems for medical applications, one can help prevent healthcare inequality and better ensure the safety of all patients.

Additional biases related to AI systems for medicine (or any field) include latent bias, or biases waiting to happen. A recent article discussed three such biases.[8] First, if the model is adaptive, it can become biased over time. For instance, the model may perform unbiased in one context, but may produce biased results by learning from disparities in a different context. As an example, a model may predict patient outcomes well in one medical facility, but it may predict worse outcomes for patients in a facility with disparities in care for patients of different ethnic or racial backgrounds. Second, as humans interact with the model, some of their implicit or explicit biases may impact the model's outcomes and become implanted in it. For example, if a medical professional treats the AI-based predictions as infallible rather than as a decision support tool, the model will not receive proper reinforcement, potentially leading to biased results and worse outcomes for patients.[8] Third, biases may arise from the choice of what the model is intended to achieve. For instance, bias may occur if the model learns to preferentially select one outcome over others, which may not be in the best interest of certain stakeholders. As an example, a medical facility may influence an AI system's goal so that it chooses outcomes that return more profit, and this bias in the choice of outcome may negatively affect the patients' care. Altogether, these latent biases present a significant challenge in developing AI systems for medicine, and methods of mitigation should be considered.

Three techniques for combating the aforementioned latent biases have been noted.[8] First, biases in AI algorithms should be identified and addressed proactively, rather than after the fact. Such algorithms should be monitored for biases in predictive performance and for the way their predictions are used. Second, it is suggested that regulatory frameworks governing AI algorithms include reference to monitoring for biases, including latent biases, in performance. DeCamp and Lindvall[8] assert that latent biases should be considered adverse events, and as such, they should be reported and managed with similar mechanisms as in the case of a drug product being found to be defective. Finally, it is believed that the organizations building the AI models should engage with their stakeholders, i.e., those who will be using them, throughout development and implementation. Stakeholders may have different opinions about applications of a particular AI model; therefore, engaging with them may prevent biases related to defining which applications are appropriate for AI and which are not. Overall, the field of medicine presents many unique challenges regarding bias within AI systems, and these issues could be mitigated by applying some of the techniques discussed above.

Recent papers have suggested that ChatGPT and other LLMs may have large implications for the future of health care,[13,37–39] and as such, it is important to consider the potential risks of bias resulting from using these AI tools in this domain. Some potential applications of LLMs in health care include the following: (i) improving efficiency in medical writing, (ii) overcoming language barriers, and (iii) providing treatment and care information to patients.[13,37,38] Furthermore, to a certain extent, ChatGPT and other LLMs may be used to answer a patient's medical questions and suggest potential avenues of care.[39] Nonetheless, each of these

applications has drawbacks and potential risks. For instance, without human judgment and intervention, an LLM used to assist with medical writing may directly plagiarize from its sources.[37,38] Additionally, when used to answer a patient's medical questions, an LLM may provide patients with incorrect or harmful medical advice due to biases or inaccuracies within its training data.[37,39] Furthermore, ChatGPT and other chat models may generate inaccurate or false information to fulfill a user's request.[40] For example, when asked to provide a differential diagnosis for postpartum hemorrhage, ChatGPT provided what appeared to be an excellent answer with supporting references; however, upon further review, none of the cited sources actually existed.[40] All things considered, ChatGPT and other LLMs have the ability to revolutionize the future of health care. But it is necessary to address the concerns about bias, misinformation, and misconduct with regard to use of these AI tools in future research.

EFFORTS TOWARD AI REGULATION AND UNDERSTANDING OF AI BIAS

Medical devices with ML algorithms have been on the market for several decades, with the first device approved by the FDA in 1995.[41] Since then, several technological changes have resulted in many AI-enabled medical devices, spanning device categories and product codes. As noted in Section 1: Sources of Bias, there are various sources of bias. Complex variations in device types with varying levels of risk may require different mitigation and evaluation approaches specific to the various development stages of AI devices, from conception to the final translation.

The FDA has been addressing these regulatory challenges using a combination of white paper discussions, regulatory policies, regulatory science research, and external collaborations. Its 2019 discussion paper on regulatory framework for modifications to AI devices using a predetermined change control plan was geared toward encouraging the adoption of an iterative process for improving AI algorithms.[42] This total product life cycle regulatory approach would include a feedback mechanism from real-world performance monitoring, which should identify any systematic differences between sensitive subgroups and allow improvements that would mitigate any sources of bias. In the same year, the FDA's Patient Engagement Advisory Committee, consisting of patients, caregivers, and representatives of patient organizations, provided recommendations for AI and ML in medical devices. The Committee's recommendations addressed the importance of including various demographic groups in AI algorithm development. They also addressed the impact of the user interface and transparency, including which information about the devices could be communicated, and how, to foster patient trust in AI devices.

Another agency initiative with AI focus is the Digital Health Center of Excellence established in 2020. This initiative's goal is to advance digital health technologies, including mobile health devices, Software as a Medical Device (SaMD), and wearables when used as medical devices. The initiative aligns and coordinates resources across the agency to (i) connect and build partnerships, (ii) share knowledge to increase awareness and understanding, and (iii) innovate regulatory approaches to provide efficient and least burdensome oversight while meeting the FDA's standards for safe and effective products. Based on the feedback from the 2019 discussion paper, an action plan was released in 2021. One action item is to support the development of regulatory science methods related to algorithm bias and robustness. A new draft guidance document based on the 2019 discussion paper and the 2021 action plan for the algorithm modifications through a predetermined change control plan was released for public comment.[43]

A public workshop on transparency of AI-enabled medical devices was held in 2021 to (i) identify unique considerations in achieving transparency for users of AI-enabled medical devices and ways in which transparency might enhance the safety and effectiveness of these devices, and (ii) gather input from various stakeholders on the types of information that would be helpful for a manufacturer to include in the labeling of any public-facing information about AI-enabled medical devices as well as other potential mechanisms for information sharing.[44] The FDA has long recognized that diversity in clinical trials is a critical step toward ensuring that medical products (including those that utilize AI) are safe and effective for the intended populations. In 2020, the agency released a guidance document titled, "Enhancing the Diversity of Clinical Trial Populations — Eligibility Criteria, Enrollment Practices, and Trial Designs." This guidance recommends approaches industry can take to increase enrollment of underrepresented populations in their clinical trials.[45] Furthermore, in 2022, the FDA released a draft guidance document titled, "Diversity Plans to Improve Enrollment of Participants from Underrepresented Racial and Ethnic Populations in Clinical Trials." Its purpose was to provide recommendations to industry on developing Race and Ethnicity Diversity Plans to enroll adequate numbers of participants from underrepresented U.S. racial and ethnic populations in clinical trials.[46] Together, these guidance documents can help promote the safety and efficacy of AI-enabled medical devices.

In addition, the FDA, Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have jointly identified 10 guiding principles that can inform the development of Good Machine Learning Practice (GMLP). These emphasized the need to mitigate sources of bias in the data used for AI development and clinical study participants and in the use of post-market surveillance to continually improve the safety and effectiveness of AI devices. In addition to the regulatory policy efforts, the Agency fosters regulatory science through its AI/ML research program, development of regulatory science tools, and leveraging of academic collaborations through Centers of Excellence in Regulatory Science and Innovation (CERSI).[47–50]

In 2023, FDA Commissioner Robert M. Califf spoke to the National Health Council's 2023 Science for Patient Engagement Symposium about patient empowerment in the digital age, addressing several regulatory concerns and challenges, including algorithmic bias and LLMs.[50] He posited that algorithms, including for LLMs such as ChatGPT, may evolve after they are put into practice, and as such, their performance and accuracy may change over time. Thus, it is important to continuously monitor and assess algorithms' performance and bias throughout their life cycles.[50] Furthermore, because of the many potential and revolutionary uses of LLMs, it is suggested that the regulation of such models will be critical in the future. Since they are rapidly evolving, such

regulation may be necessary to lessen the risks associated with negative bias and misinformation. Additionally, the time to review new applications for potential biases may unintentionally slow the already strained approval process if no new tools and workflows are adopted to assist reviewers. Overall, with the accelerating inclusion of LLMs and other AI technologies in medical products and their development, it is increasingly important to consider the potential for bias in such tools as well as methods of measuring and mitigating these biases.

CONCLUSION

Recently, many novel AI systems, including LLMs, have generated amazement for their ability to analyze immense amounts of data and suggest logical answers to challenging questions in a fraction of the time it would take a human to complete the same task. But these herculean efforts may not come without a price. Just as humans are susceptible to bias and concerns about fairness, the systems they design may be as well. In addition, because of the exponential growth in data and data sources, the deliverables expected from AI systems could be exposed to bias at unprecedented levels. The study of AI bias has already highlighted considerable concerns in the development and deployment arenas, especially in high-stakes decision-making scenarios. To date, only a limited amount of research has been published in this area. Future studies to understand, measure, and mitigate negative or unintended biases when developing new AI systems or debiasing current existing AI systems could accrue significant returns on investment. Acknowledging the great potential of AI to improve the future of health care and regulatory science, we would like to continue returning to AI-related ethical issues—including bias—and hope to see more research advances in measuring and mitigating potential biases from various sources.

ACKNOWLEDGMENT

We would like to thank Joanne Berger, FDA Library, for editing a draft of the manuscript. We wish to acknowledge that Magnus Gray's contribution was made possible in part by his appointment through the Research Participation Program at the National Center for Toxicological Research, administered by the U.S. Food and Drug Administration through the Oak Ridge Institute for Science and Education.

FUNDING

This project is funded by the FDA Chief Scientist's Intramural Challenge Grants (E0782201).

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

DISCLAIMER

The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification and is not intended as an endorsement.

1. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G> (2018). Accessed June 23, 2023.
2. Oxford English Dictionary. bias, n., adj., adv. <https://www.oed.com/dictionary/bias_n> (2023).
3. Managing Bias in AI. In *Companion Proceedings of the 2019 World Wide Web Conference* (eds. Roselli, D. *et al*.) (Association for Computing Machinery, New York, 2019).
4. Navarro, C.L.A. *et al*. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *Br. Med. J* **375**, n2281 (2021).
5. Hovy, D. & Prabhumoye, S. Five sources of bias in natural language processing. *Lang. Linguist. Compass* **15**, e12432 (2021).
6. Panch, T., Mattie, H. & Atun, R. Artificial intelligence and algorithmic bias: implications for health systems. *J. Glob. Health* **9**, 10318 (2019).
7. Vokinger, K.N., Feuerriegel, S. & Kesselheim, A.S. Mitigating bias in machine learning for medicine. *Commun. Med. (Lond).* **1**, 25 (2021).
8. DeCamp, M. & Lindvall, C. Latent bias and the implementation of artificial intelligence in medicine. *J. Am. Med. Inform. Assoc.* **27**, 2020–2023 (2020).
9. Garg, P., Villasenor, J. & Foggo, V. Fairness metrics: a comparative analysis. 2020 IEEE International Conference on Big Data (Big Data), Atlanta, Georgia, December 10–13, 2020.
10. OpenAI. GPT-4 technical report (2023).
11. Ferrara, E. Should ChatGPT be biased? Challenges and risks of bias in large language models. *arXiv* arXiv:230403738 (2023).
12. Zhuo, T.Y., Huang, Y., Chen, C. & Xing, Z. Exploring AI ethics of ChatGPT: a diagnostic analysis. *arXiv* arXiv:230112867 (2023).
13. Ray, P.P. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* **3**, 121–154 (2023).
14. Russell, C., Kusner, M.J., Loftus, J. & Silva, R. When worlds collide: integrating different counterfactual assumptions in fairness. *Adv. Neural Inf. Process. Syst.* **30**, 1–10 (2017).
15. Caliskan, A., Bryson, J.J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
16. Dev, S. & Phillips, J. (eds.). Attenuating bias in word vectors. The 22nd International Conference on Artificial Intelligence and Statistics, Okinawa, Japan, April 16–18, 2019.
17. Dev, S., Li, T., Phillips, J.M. & Srikumar, V. (eds.). On measuring and mitigating biased inferences of word embeddings. Proceedings of the AAAI Conference on Artificial Intelligence, February 7–12, 2020.
18. Rathore, A. *et al*. VERB: Visualizing and Interpreting Bias Mitigation Techniques for Word Representations. *arXiv* arXiv:210402797 (2021).
19. May, C., Wang, A., Bordia, S., Bowman, S.R. & Rudinger, R. On Measuring Social Biases in Sentence Encoders. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Naacl Hlt 2019) 1, 622–628 (2019).
20. Liang, P.P., Li, I.M., Zheng, E., Lim, Y.C., Salakhutdinov, R. & Morency, L.-P. Towards debiasing sentence representations. *arXiv* arXiv:200708100 (2020).
21. Guo, Y., Yang, Y. & Abbasi, A. (eds.). Auto-debias: debiasing masked language models with automated biased prompts. 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, May, 22–27, 2022.
22. Nadeem, M., Bethke, A. & Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv* arXiv:200409456 (2020).
23. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* arXiv:181004805 (2018).
24. Liu, Y. *et al*. Roberta: A robustly optimized bert pretraining approach. *arXiv* arXiv:190711692 (2019).
25. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).

26. Ahn, J. & Oh, A. Mitigating language-dependent ethnic bias in BERT. 2021 Conference on empirical methods in natural language processing (Emnlp 2021), 533–549 (2021).
27. Aka, O., Burke, K., Bauerle, A., Greer, C. & Mitchell, M. Measuring model biases in the absence of ground truth. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (2021).
28. Ouyang, L. *et al*. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
29. Brown, T. *et al*. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
30. Parrish, A. *et al*. BBQ: A hand-built bias benchmark for question answering. *arXiv* arXiv:211008193 (2021).
31. Dhamala, J. *et al*. Bold: Dataset and metrics for measuring biases in open-ended language generation. Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. *arXiv* arXiv:2101.11718 (2021).
32. Zhao, J., Wang, T., Yatskar, M., Ordonez, V. & Chang, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. arXiv preprint arXiv:180406876 (2018).
33. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S.R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv* arXiv:180407461 (2018).
34. Zhou, Y., Kantarcioglu, M. & Clifton, C. Improving fairness of AI systems with lossless de-biasing. *arXiv* arXiv:210504534 (2021).
35. Liu, Q. *et al*. Landscape analysis of the application of artificial intelligence and machine learning in regulatory submissions for drug development from 2016 to 2021. *Clin. Pharmacol. Ther.* **113**, 771–774 (2023).
36. Hammond, M.E.H., Stehlik, J., Drakos, S.G. & Kfoury, A.G. Bias in medicine: lessons learned and mitigation strategies. *JACC Basic Transl. Sci* **6**, 78–85 (2021).
37. Sallam, M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* **11**, 887 (2023).
38. Kitamura, F.C. ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology* **307**, e230171 (2023).
39. Zuccon, G. & Koopman, B. Dr ChatGPT, tell me what I want to hear: how prompt knowledge impacts health answer correctness. *arXiv* arXiv:230213793 (2023).
40. Doshi, R.H. & Bajaj, S. Promises — and pitfalls — of ChatGPT-assisted medicine. *STAT* <https://www.statnews.com/2023/02/01/promises-pitfalls-chatgpt-assisted-medicine/> (2023).
41. US Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices> (2022).
42. US Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD): discussion paper and request for feedback (2019).
43. US Food and Drug Administration. Draft guidance: marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning (AI/ML)-enabled device software functions <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial> (2023).
44. US Food and Drug Administration. Virtual public workshop - transparency of artificial intelligence/machine learning-enabled medical devices (2021).
45. US Food and Drug Administration. Enhancing the diversity of clinical trial populations: eligibility criteria, enrollment practices, and trial designs guidance for industry <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enhancing-diversity-clinical-trial-populations-eligibility-criteria-enrollment-practices-and-trial> (2023).
46. US Food and Drug Administration. Draft guidance for industry: diversity plans to improve enrollment of participants from underrepresented racial and ethnic populations in clinical trials; availability <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/diversity-plans-improve-enrollment-participants-underrepresented-racial-and-ethnic-populations> (2022).
47. US Food and Drug Administration. Artificial intelligence program: research on AI/ML-based medical devices <https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/artificial-intelligence-and-machine-learning-program-research-aiml-based-medical-devices> (2021).
48. US Food and Drug Administration. Catalog of regulatory science tools to help assess new medical devices <https://www.fda.gov/medical-devices/science-and-research-medical-devices/catalog-regulatory-science-tools-help-assess-new-medical-devices> (2023).
49. US Food and Drug Administration. CERSI research projects <https://www.fda.gov/science-research/advancing-regulatory-science/cersi-research-projects> (2023).
50. Califf, R.M. Speech by Robert M. Califf, M.D. to the National Health Council's 2023 Science for Patient Engagement Symposium: patient empowerment in the digital health era <https://www.fda.gov/news-events/speeches-fda-officials/speech-robert-m-califf-md-national-health-councils-2023-science-patient-engagement-symposium-patient> (2023).