Introduction

This literature review investigates the essential topic of bias in artificial intelligence, examining multiple concepts and strategies across multiple studies. "Bias" can refer to a few different concepts. Originally, bias in machine learning (ML) described a skewed representation of a phenomenon [27]. An example of this is when a nonlinear relationship is represented as linear and therefore consistently overestimates some predictions while underestimating others. In another sense, bias means anything that helps the model determine what kind of decision it is going to make, whether helpful or unhelpful [24]. This kind of bias is present in every machine learning model.

This review is interested in the type of bias that has the potential to benefit or harm people based on some protected characteristic, such as race, age, gender, or disability status. With that goal in mind, when this paper references "bias," it refers to a difference in treatment (usually performance accuracy) between different representing humans with different protected characteristics. This discriminatory bias is hard enough to identify in human decision-making processes and presents a challenge to the data science community as well: hidden biases cannot be corrected. The present review investigates the research question: how can bias be identified and measured in ML/AI models? The following discussion will show that such bias is present in many machine learning systems, answer the research question, and demonstrate that bias mitigation in those systems is a goal that is both worthy and feasible.

The problem of bias

Bias is pervasive in machine learning models across disciplines, and can have real consequences for the people who use the models or those who are affected indirectly. Recent interest in mitigating bias testifies to the prevalence of bias as an issue [20]. One study found that a pedestrian detection model identified pedestrians with darker skin tones less reliably than those with lighter skin tones [5]. Gender and skin tone can have effects on the accuracy of word error rate in speech recognition models [6]. The benchmark study in [29] evaluated the performance of large language models (LLMs) such as GPT-4, Llama 3, and others across many bias areas, including gender, racism, and handicap, by systematically testing them against an expert-curated dataset. All LLMs evaluated were found to have considerable biases, especially in increasingly difficult reasoning tasks.

These are not isolated examples, nor are they victimless or harmless mistakes. Biases in AI modeling, training, and application can result in discriminatory outcomes in fields such as policing, credit scoring, and health evaluations [20, 36, 18]. LLMs can spread misconceptions and provide unfair results in a variety of applications, including content filtering and decision-making systems [29]. The authors of [14] identify two main types of harm: allocative and representational. Allocative harm happens when resources (e.g. credit, job opportunities) are withheld from groups of people. Representational harm is when people are stigmatized or stereotyped based on group characteristics [14]. Stereotyping can then lead to allocative harm by human systems or to emotional trauma. The importance of addressing biases against protected characteristics is increasing quickly as AI systems become more and more integrated in fields such as healthcare, criminal justice, and employment, where fairness is critical.

How bias gets into models

It is no wonder that bias shows up in AI/ML systems; investigating the lifecycle of a model reveals myriad opportunities to introduce bias. Some of these are discussed in this section, though a reader should not assume that what follows is a comprehensive list of opportunities for bias to spring up in a system. Given the diversity of AI/ML training methods and applications, such a list may be impossible. It is important for any data practitioner to carefully consider how their work might introduce harmful bias into their product.

Bias in the ML lifecycle can start before the model itself is even conceived. Research design guides what questions are asked and how – and for whom – they are answered [19, 7]. Sampling bias occurs when certain groups are overrepresented in the sample compared to the overall population. Even if groups are represented proportionally to the population, data bias can exist when there are many more examples of one group than another [14]. Datasets can be labeled with bias if annotation guidelines are unclear, or labelers have preconceived opinions about the data [20]. Feature bias can emerge when predictors are incorporated that may not be appropriate for predicting the response variable. For example, rental history is a good indicator of a person's likelihood to make on-time mortgage payments, but rental payments are mostly not considered by the credit bureaus. This practice limits credit for low-income individuals and disproportionately affects people of color [18]. Similarly, preprocessing methods such as encoding sensitive attributes or applying transformations are likely to introduce bias [26, 31].

Once the data is ready, algorithms can be trained to perpetuate biased human decisions. A classic example is hiring algorithms or ATS software that favors applicants with white-sounding names. Evaluation bias occurs when the metrics that are used to optimize the model favor certain model performance that does not align with fair outcomes across groups [20]. Temporal bias can creep in when the original training data no longer accurately represents reality. Aggregation bias treats as similar multiple different subsets of data that should be treated as distinct, and the result can be an application that works well for only the dominant subgroup, or no group at all [14]. Finally, once the model is trained and ready for use, deployment bias arises if it is utilized for purposes beyond its intended scope [14, 36].

Detecting bias

Now that the problem of bias is established, the research question presents itself: how can bias be detected in machine learning models, especially AI tools that work like a "black box"? Bias needs to be identified and, if possible, quantified before it can be corrected in ML applications. Every model that can generalize predictions based on training data does so using some sort of inductive bias. Insight into the model's decision-making can help improve its performance and also help identify its biases. This section explores a few quantitative and qualitative methods for detecting bias in AI models, as well as some existing tools that can assist in bias detection.

Quantitative methods

Most quantitative bias detection techniques consist of calculating an aggregate score, or metric, for each category of a demographic group and comparing the scores between groups. The goal is to make sure that the model works equally well on data with diverse features. If the calculated score is similar for both (or all) groups, that indicates less bias. On the other hand, if the scores are very different across groups, that indicates high levels of bias. For example, if a facial recognition tool used by police mismatches black men more often than it mismatches white men, that would indicate a racial bias in the model.

Table 1 lists a sample of common metrics that conform to the method detailed above. It is important to consider that no such list should be considered to be comprehensive, and no single metric can give a holistic view of model bias. Some literature suggests that a base threshold for disparity may be set at 80% [32, 23]. This may be based upon the common use of the four-fifths rule to quantify disparate impact, though such a threshold may be misapplied in computing contexts [11]. Any data practitioner utilizing these bias metrics should carefully consider the context and application of the model, as well as the consequences of error for any subgroups when determining an acceptable threshold.

Metric	**Definition**	**Important for**
Statistical/equal parity	Equal representation for all groups in the dataset	Datasets
Proportional parity	Representation in the dataset proportional to the population	Datasets
Predictive parity	Equal positive predictive value; the chance an observation belongs to its predicted group	All models
Equalized/average odds	Combined equal chances of true positives and false positives	All models
Opportunity equality	Equal probability of a positive prediction given a positive label	Allocative models
Demographic parity	Equal probability of a positive prediction	All models
Counterfactual fairness	Whether a classifier assigns the same label to two inputs equal except for a sensitive trait	All models
False positive rate parity	Equal odds of negatives being identified as positives	Punitive models
False discovery rate parity	Equal proportion of all positive predictions are actually negative	Punitive models
False negative rate parity	Equal odds of positives being identified as negatives	Allocative models
False omission rate parity	Equal proportion of all negative predictions are actually positive	Allocative models

Table 1: Bias metrics [19, 32, 23]

The N-Sigma approach [8], an adaptation of 5-sigma to evaluate bias, takes the means of different demographic groups to be compared (e.g. racial groups) and divides them by the standard deviation of the respective group to yield a measure of distance between the distributions. Finding the distance between the distributions for each demographic group allows for risk levels to be identified instead of a simple ‘yes’ or ‘no’ for whether bias is present. Just like 5-sigma seeks a result that is 5 standard deviations away from the mean, N-sigma can represent different risk levels based on how many standard deviations one group is from the mean of the other(s).

The authors of [211devised a metric based on normalized pointwise mutual information (nPMI) to measure biases on a dataset where there may not be ground truth labels for protected characteristics such as race, gender, or disability status. The method calculates associations between labels predicted by a classifier model to uncover biases relating to certain labels. For example, this method will capture if a model is more likely to predict ‘man’ for the same input as it is likely to predict ‘doctor’ but more likely to predict ‘woman’ for the same input as ‘nurse’ is more likely. The nPMI method does not measure overall accuracy, but it works with unlabeled datasets and uncovers bias that is baked into the model.

Some image classification models were found to have undesirable bias where they would make classifications based on irrelevant background information [30]. To identify this type of bias, the authors of [30] applied image transforms such as Fourier transform, wavelet transform, median filter, and combinations in order to smooth out the noise in the pictures and refine the hidden signal in irrelevant background information. They found that transformations had different effects on curated versus natural datasets. Since curated datasets are more likely to have irrelevant, standard backgrounds and natural datasets are more likely to have 100% relevant information, they concluded that testing the model with the transformed pictures helped find bias without having to crop the background out.

Quantitative methods enable researchers to audit their model’s performance on a lot of data rather quickly. There are some limitations to quantitative bias detection, however. Bias is a complicated phenomenon with multifaceted impacts that cannot be easily compressed into a single number or even a group of numbers. Additionally, the application or scope of a model can greatly affect what metrics are important to look out for. Finally, these metrics work best when the dataset being used to test the model is balanced between the demographic groups being tested [19].

Qualitative methods

Understanding how ML models make their decisions using qualitative investigation methods can be a useful step to identify bias and fill in some gaps of quantitative analysis, though there are some limitations. Qualitative methods are inherently less scalable than quantitative methods; only a smaller subset of the data can be investigated, since each instance needs to be evaluated by a human. Even so, important biases that may be hidden from qualitative bias detection methods can be uncovered by probing the model’s performance on a smaller subset of specific data. One example is the investigation by [24] into a ResNet-101 classifier which revealed that classifications of ‘basketball’ heavily relied on pixels displaying the players’ skin. Even if the ResNet did not demonstrate disparity in performance between racial groups, the fact that it appeared to judge sports labels based on skin color would still be an undesirable bias in the model.

Using the cognitive psychology approach, a neural network can be evaluated for its biases the same way a human would be. One strategy in the cognitive psychology realm is the hypothesis elimination approach [28], which involves 3 steps:

1. Make an assumption about how the model will map inputs to outputs
2. Determine what decisions (e.g. categories predicted) the model would make if it used that mapping
3. Compare the actual decisions of the model with the projected outcomes from step (2).

If the actual model output does not match the simulated decisions, it is likely that the posited mapping is not how the model is making decisions. That hypothesis can be eliminated, and another one can be tested. It is important to note that this method does not definitively explain the model's decision-making process but rather identifies a decision-making rule that produces similar results as the model.

Two experiments that we found involved curating datasets to see how people or machines reacted to them. In [28], a one-shot labeling model was given a probe image along with two related images for reference: one that matched the probe image in shape, and one that matched the probe image in color. The images were controlled for size and background. The researchers measured how many times the model predicted the same as the shape-match image and the color-match image, respectively. The more matches for shape that they recorded, the stronger the bias of the model towards shape information as opposed to color information. This experiment introduced the useful idea that small, curated datasets can be used to extract meaningful bias information from a model.

Another cognitive psychology experiment [9], performed on humans, can be extrapolated to test machine learning models. First, the researchers evaluated how people biased one word over another when both words had equal probabilities by themselves but unequal probabilities in the presence of the context word. In a second experiment, the researchers used random combinations of probabilities alone and in the context of the given word. Finally, the researchers presented curated combinations of high and low probabilities of the words themselves and in the context of the given word. The experiments found that if the context word increased the probability of a given word occurring, people were more likely to choose that word, even over a word with higher overall probability. The structure of the experiment demonstrates an effective and thorough method for uncovering bias in a test subject. If a cognitive psychology approach is implemented for bias detection, the model's performance should be evaluated in a variety of different situations.

Another general method for qualitative bias evaluation involves assigning weights to the input features to identify which have the most impact on the model's decision-making process. There are multiple techniques for weighting input features, including: activation maximization, sensitivity analysis, and layer-wise relevance propagation (LRP). Activation maximization involves finding an input (such as an image) for each class that gives the highest probability for that class. The input, called the prototype, can be supplied by a data density model or a generative network. The features of the prototype image should give insight into the most salient features related to the model's decision-making process [1]. Activation maximization is limited, however, in that it may be unclear which features of the prototype are actually the most important to the model's mapping function.

Sensitivity analysis uses a relevance scoring metric (such as gradient over partial gradient) to evaluate the importance of each feature. Notably, sensitivity analysis only models a variation of the mapping function, not the function itself. LRP, on the other hand, assigns a proportion of relevance to each feature map on each layer of the neural network one layer at a time until the relevance is distributed over the original features (pixels, words, etc.). Both sensitivity analysis and LRP can be used to produce a heatmap of the most relevant features. The accuracy of the heatmaps can be ascertained by measuring the decline in performance when the most relevant features are removed. Based on this method, LRP performs better than sensitivity analysis [13, 35].

Existing tools

Aequitas is a bias audit tool built by the Center for Data Science and Public Policy at the University of Chicago [22]. It is accessible via Python API, command line interface (CLI), and through their web application [23]. A user uploads the data, including ground truth labels, predicted outcomes, and demographic features. Aequitas calculates a variety of bias and fairness metrics, comparing outcomes across different subgroups, and generating comprehensive bias reports that highlight any statistically significant disparities identified.

TIBET [3] detects and analyzes biases in text-to-image (TTI) generative models, focusing on the intersectional nature of different bias axes. First, it generates images from a given TTI model and detects possible bias axes for the given prompt (e.g. race, gender, cultural norms). Then, it creates counterfactual prompts along the relevant bias axes to produce images that differ with respect to that bias. Finally, it uses an image comparison model to quantify the difference between the original images created and the alternatives generated by the counterfactual prompts. If the distance between the original set of generated images and each bias counterfactual imageset is roughly the same, there is less likely to be bias along that axis. TIBET also includes qualitative tools to explain the bias that was detected. The code is available on GitHub [2].

AI Fairness 360 is a python package that includes bias detection as well as mitigation algorithms. It offers more than 70 fairness metrics [2] to help quantify fairness, including statistical parity, opportunity equality, and average odds.

BiasAsker is a novel methodology for detecting and quantifying social bias in conversational AI systems, such as ChatGPT and Siri [37]. It may be difficult to identify prompts in these systems due to their black box nature and a lack of comprehensive benchmark datasets that encompass diverse social groupings and prejudices. BiasAsker creates test prompts that successfully trigger biased responses by leveraging a dataset that includes 841 social groups and 8,110 biased characteristics. The broad dataset allows it to develop focused questions that look for biases across multiple domains of social interaction. BiasAsker distinguishes between "absolute bias," in which prejudice is clearly communicated, and "relative bias," in which replies to different prompts are compared to identify subtler indications of bias. BiasAsker also provides metrics and visualization tools for assessing the level and type of these biases. The code is publicly available on GitHub [38].

Bias in datasets

When identifying bias in machine learning models, it is of paramount importance to ensure that the dataset used for evaluation is as balanced as possible [19]. Specialized benchmark datasets such as StereoSet, BOLD, and FairFace exist so that the metrics listed in Table 1 are more likely to measure the bias of the model, not the dataset itself. Statistical and proportional parity, listed in Table 1, are important measures for ensuring that the dataset does not favor one group over another.

Mitigating bias

While the main purpose of this literature review is not to provide methods for bias mitigation, it is worth mentioning a few that provide hope for a less biased future of machine learning. This section will provide a brief overview of some general bias mitigation techniques, among many, for both datasets and ML models.

The best way to ensure an unbiased dataset is to collect data from each subgroup equally. When working with a pre-collected dataset, however, some bias mitigation methods are still available. If one group is under-or over-represented in the data, over- and under-sampling techniques exist to even it out. Data augmentation such as adjusting brightness, cropping and rotating photos can help a model generalize with less training data [20].

If a balanced dataset is available, then retraining a biased model may be a good start to make it more fair [19]. Model parameters can be manually tweaked to increase performance in terms of bias metrics [19, 39]. If features indicating protected characteristics do not contain information relevant to prediction, those features can simply be removed to ensure they do not influence the decision [27]. Other methodologies include fair representation via encoder [27], adversarial training [39], and regularization [39]. The authors of [31] propose a method for removing gender stereotypes from word embeddings preserving useful word associations properties by clustering related concepts. Other debiasing techniques involve manipulating the model output after the fact [32].

Procedures surrounding the implementation of ML systems can limit bias by ensuring that models are utilized appropriately, within their intended scope and with appropriate audits. Debiasing AI/ML models calls for attention beyond technical solutions, and should incorporate legal, social, and ethical views to properly comprehend and address the consequences of AI-driven judgments [3].

Challenges

The challenge of mitigating bias extends from the potential introduction of bias at every phase of model development [14].

The study of bias detection and mitigation is far from complete. Research has focused more on binary outcomes, leaving knowledge gaps for systems involving multi-class and multi-metric analyses [32]. Even with a variety of studies covering different types of models and use cases, fairness measures may not work as intended beyond the experiments they were studied in. As such, there is no catch-all standard for fairness across or even within model families. With such heterogeny in model designs and use cases, each model must be carefully evaluated with a range of metrics and techniques to give a holistic view of its bias and possible discriminatory impact. Bias mitigation is a moving target in that societal norms and stereotypes are ever evolving [20]. Data in this decade may reflect different biases than those from the last, and since models need to be retrained to stay up to date, they also must be continually re-audited.

Intersectionality describes subgroups involving multiple minorities, such as a black woman, who may experience bias and discrimination against her based on not only her race and gender, but on the interaction of these two identity features. Intersectionality poses a challenge for bias detection and mitigation because intersectional groups tend to be small. The authors of [16] propose that fairness metrics should not allow some groups to suffer more discrimination than others and should not encourage "cheating," such as simply collecting less data or gaming the optimization algorithm.

It must not be assumed, based on group-level fairness statistics, that a model is equally fair for every individual involved [27]. Individual lives are far too complex to be flattened into a set of numbers for an AI model. If an AI/ML model is used to make important decisions related to job opportunities, credit, policing, etc., there must be an appeal process involving human beings that can consider every angle and exception relating to a person's situation [34].

References

[1] A. A. Kuriakose, “Algomox Blog | Bias in Generative AI: Detection, Mitigation, and Management through MLOps,” Bias in Generative AI: Detection, Mitigation, and Management through MLOps, Apr. 16, 2024. https://www.algomox.com/resources/blog/bias_generative_ai_detection_mitigation_mlops/ (accessed Feb. 07, 2025).

[2] A. Chinchure et al., “TIBET,” https://github.com/TIBET-AI/TIBET. (Accessed Apr. 11, 2025).

[3] A. Chinchure et al., “TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models,” arXiv, 2023, doi: 10.48550/arxiv.2312.01261.

[4] B. Moses and A. Dhinakaran, “ML Observability Overview | Machine Learning Observability Resources,” Beyond Monitoring: The Rise of ML Observability, May 19, 2021. https://www.montecarlodata.com/blog-beyond-monitoring-the-rise-of-observability/ (accessed Feb. 07, 2025).

[5] B. Wilson, J. Hoffman, and J. Morgenstern, “Predictive inequity in object detection,” arXiv, 2019, doi: 10.48550/arxiv.1902.11097.

[6] C. Liu et al., “Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions,” arXiv, 2021, doi: 10.48550/arXiv.2111.09983.

[7] C. R. Sugimoto and V. Larivière, *Equity for Women in Science: Dismantling Systemic Barriers to Advancement*, Harvard University Press, Cambridge, MA, USA, 2023.

[8] D. DeAlcala, I. Serna, A. Morales, J. Fierrez, and J. Ortega-Garcia, “Measuring Bias in AI Models: An Statistical Approach Introducing N-Sigma,” in 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Jun. 2023, pp. 1167–1172, doi: 10.1109/COMPSAC57700.2023.00176.

[9] D. Paperno, M. Marelli, K. Tentori, and M. Baroni, “Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood.,” Cogn. Psychol., vol. 74, pp. 66–83, Nov. 2014, doi: 10.1016/j.cogpsych.2014.07.001.

[10] D. Roselli, J. Matthews, and N. Talagala, “Managing bias in AI,” in Companion Proceedings of The 2019 World Wide Web Conference on - WWW ’19, New York, New York, USA, May 2019, pp. 539–544, doi: 10.1145/3308560.3317590.

[11] E. A. Watkins, M. McKenna, and J. Chen, “The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness,” arXiv, 2022, doi: 10.48550/arxiv.2202.09519.

[12] G. Barcelos, “Understanding Bias in Machine Learning Models - Arize AI,” Understanding Bias in Machine Learning Models, Mar. 15, 2022. https://arize.com/blog/understanding-bias-in-ml-models/ (accessed Feb. 04, 2025).

[13] G. Montavon, W. Samek, and K.-R. Müller, “Methods for Interpreting and Understanding Deep Neural Networks,” arXiv, 2017, doi: 10.48550/arxiv.1706.07979.

[14] H. Suresh and J. V. Guttag, “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle,” arXiv, 2019, doi: 10.48550/arxiv.1901.10002.

[15] J. Finocchiaro et al., “Bridging Machine Learning and Mechanism Design towards Algorithmic Fairness,” arXiv, 2020, doi: 10.48550/arxiv.2010.05434.

[16] J. Himmelreich, A. Hsu, K. Lum, and E. Veomett, “The Intersectionality Problem for Algorithmic Fairness,” arXiv, 2024, doi: 10.48550/arxiv.2411.02569.

[17] K. Kärkkäinen, and J. Joo. “FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2019, pp. 1548-1558, arXiv, doi: 10.48550/arXiv.1908.04913.

[18] L. Rice, “Missing Credit: How the U.S. Credit System Restricts Access to Consumers of Color,” in *Who’s Keeping Score? Holding Credit Bureaus Accountable and Repairing a Broken System: Hearing before the U.S. House Committee on Financial Services*, 116th Cong., Feb, 2019, https://nationalfairhousing.org/wp-content/uploads/2019/04/Missing-Credit.pdf (Accessed Apr. 04, 2025).

[19] M. Gray et al., “Measurement and mitigation of bias in artificial intelligence: A narrative literature review for regulatory science.,” Clin. Pharmacol. Ther., vol. 115, no. 4, pp. 687–697, Apr. 2024, doi: 10.1002/cpt.3117.
[20] M. Shah and N. Sureja, “A Comprehensive Review of Bias in Deep Learning Models: Methods, Impacts, and Future Directions,” Arch. Computat. Methods Eng., vol. 32, no. 1, pp. 255–267, Jan. 2025, doi: 10.1007/s11831-024-10134-2.
[21] O. Aka, K. Burke, A. Bäuerle, C. Greer, and M. Mitchell, “Measuring Model Biases in the Absence of Ground Truth,” arXiv, 2021, doi: 10.48550/arxiv.2103.03417.
[22] P. Saleiro et al., “Aequitas: A Bias and Fairness Audit Toolkit,” arXiv, 2018, doi: 10.48550/arxiv.1811.05577.
[23] P. Saleiro et al., “Bias and Fairness Audit Toolkit,” Center for Data Science and Public Policy, University of Chicago, http://aequitas.dssg.io/ (Accessed Apr. 08, 2025).
[24] P. Stock and M. Cisse, “ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases,” in Computer vision – ECCV 2018: 15th european conference, munich, germany, september 8–14, 2018, proceedings, part VI, vol. 11210, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 504–519.
[25] R. K. E. Bellamy et al., “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” arXiv, 2018, doi: 10.48550/arxiv.1810.01943.
[26] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” arXiv, 2018, doi: 10.48550/arxiv.1802.04422.
[27] S. Dehdashtian et al., “Fairness and Bias Mitigation in Computer Vision: A Survey,” arXiv, 2024, doi: 10.48550/arxiv.2408.02464.
[28] S. Ritter, D. G. T. Barrett, A. Santoro, and M. M. Botvinick, “Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study,” arXiv, 2017, doi: 10.48550/arxiv.1706.08606.
[29] S. Simpson, J. Nukpezah, K. Brooks, and R. Pandya, “Parity benchmark for measuring bias in LLMs,” *AI Ethics*, Dec. 2024, doi: 10.1007/s43681-024-00613-4.
[30] S. T. Erukude, A. Joshi, and L. Shamir, “Identifying bias in deep neural networks using image transforms,” *Computers*, vol. 13, no. 12, p. 341, Dec. 2024, doi: 10.3390/computers13120341.
[31] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” arXiv, 2016, doi: 10.48550/arxiv.1607.06520.
[32] T. P. Pagano et al., “Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods,” *BDCC*, vol. 7, no. 1, p. 15, Jan. 2023, doi: 10.3390/bdcc7010015.
[33] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, “A survey on datasets for fairness-aware machine learning,” *WIREs Data Min & Knowl*, vol. 12, no. 3, May 2022, doi: 10.1002/widm.1452.
[34] V. Eubanks, Automating Eligibility in the Heartland. In *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. St. Martin’s Press, 2018, https://doi.org/10.5204/lthj.v1i0.1386
[35] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models,” arXiv, 2017, doi: 10.48550/arxiv.1708.08296.
[36] X. Ferrer, T. van Nuenen, J. M. Such, M. Cote, and N. Criado, “Bias and Discrimination in AI: A Cross-Disciplinary Perspective,” IEEE Technol. Soc. Mag., vol. 40, no. 2, pp. 72–80, Jun. 2021, doi: 10.1109/MTS.2021.3056293.
[37] Y. Wan, W. Wang, P. He, J. Gu, H. Bai, and M. R. Lyu, “BiasAsker: measuring the bias in conversational AI system,” in Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, New York, NY, USA, Nov. 2023, pp. 515–527, doi: 10.1145/3611643.3616310.
[38] Y. Wan, W. Wang, P. He, J. Gu, H. Bai, and M. R. Lyu, “BiasAsker,” https://github.com/yxwan123/biasasker (Accessed Apr. 12m 2025).

[39] Y. Yang et al., “Enhancing fairness in face detection in computer vision systems by demographic bias mitigation,” in Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, Jul. 2022, pp. 813–822, doi: 10.1145/3514094.3534153.
[40] Y. Zhou, M. Kantarcioglu, and C. Clifton, “Improving Fairness of AI Systems with Lossless De-biasing,” arXiv, 2021, doi: 10.48550/arxiv.2105.04534.