Survey: Bias and Explainability

Hiren Bavaskar
IIT Bombay
hiren@cse.iitb.ac.in

Pushpak Bhattacharyya
IIT Bombay
pb@cse.iitb.ac.in

Abstract

The field of Machine Learning is evolving quickly, and increasingly accurate models are being adopted to tackle more challenging problems. These highly accurate models offer us exceptional predictive abilities. However, these models often come with greater complexity. The use of black-box models results in reduced transparency to model stakeholders, which makes it difficult to deduce how the model produced a prediction. This erodes the trust of users and researchers, which demands explainability for AI (XAI) to make the decision-making process more transparent. With AI systems being continually used to make important decisions in sensitive domains such as hiring, lending, and autonomous driving, it is crucial to ensure that these decisions do not reflect discriminatory or biased behavior toward certain groups or populations. In this survey, we analyze the different types of bias which can penetrate AI systems and illustrate how explainable AI can help identify and mitigate biases to ensure fairness in learning algorithms.

1 Introduction

Machine learning models have become ubiquitous in modern society. One can observe their indispensable use in daily tasks like providing personalized recommendations for users while browsing shopping websites to making essential decisions in sensitive jobs like banking and healthcare. As these models continue to make more choices that affect people's lives, it is vital to ensure that the choices are fair and just to all communities of society. However, in cases where bias perpetuates in machine learning models, it may produce unfair results and raise serious ethical and legal concerns about their application for human tasks. Despite the growing awareness of bias in machine learning, addressing the problem of bias detection and mitigation still remains a partially solved challenge in the fairness domain. One major obstacle that restricts the uptake of these models is their lack of transparency and interpretability. These models frequently operate as black boxes, making it impossible to comprehend how they make their decisions. This makes it challenging to identify biases and understand how they impact the model's predictions.

In this survey paper, we will explore the relationship between bias and explainability in machine learning models. We will begin by defining what we mean by bias and explainability and discuss why they are important. We will then review the existing literature on bias and explainability in machine learning, including the different techniques for interpreting models and their respective strengths and weaknesses. We will then look at the combination of using explainability for bias detection and look at some recent work in this area of research. Finally, we elucidate the challenges in applying explainability for bias and provide directions to identify and mitigate bias better in the future.

1.1 Motivation

There goes a famous quote by Joanne Chen: *"AI is good at describing the world as it is today with all of its biases, but it does not know how the world should be."* It is indeed true that Artificial Intelligence (AI) and Machine Learning (ML) are prevalent almost everywhere in this era. There is a continuously increasing need for high-performance models. With the advancement of research in AI, humanity has been able to achieve stellar performance in demanding areas like medicine and autonomous driving. However, these high-performance models come at a certain cost. These models are particularly very complex and offer less insight into their actual working. This has led to a lack of trust in these models. Even a single mistake by the decision of AI can cause loss of human life in medicine, autonomous driving, and high stake scenarios if we trust the model blindly. Thus, it is critical to identify, comprehend, and reduce unfairness as machine learning models' decisions and influences have a significant impact on human lives. We are therefore driven towards understanding and explaining the prediction of models to gain better accountability and trust.

2 Bias

Bias refers to the presence of any prejudice or favoring toward a person or a group based on their innate or acquired features when it comes to decision-making (Mallela and Bhattacharyya, 2022). In today's era, a majority of AI systems and algorithms are primarily data-driven. As a result, data is inextricably linked to the functionality of these algorithms and systems. If the underlying training data has biases, the algorithms trained on it will learn these biases. The presence of such biases can make the model predict inaccurate or unfair outcomes. With AI and ML continuously being used in areas of high stakes like medicine, judicial decisions, and finance, one cannot afford to perpetuate biases in such applications. *For example, Parikh et al. (2019) remarked that among women with breast cancer, black women had a lower likelihood of being tested for high-risk germline mutations compared with white women, despite carrying a similar risk of such mutations.* Thus, an AI algorithm that depends on genetic test results is more likely to mischaracterize the risk of breast cancer for black patients than white patients. As a result, AI might be prejudiced against some minorities and worsen their access to healthcare, especially those who are already marginalized in society. Therefore, it is important to identify and mitigate bias from the data and learning algorithms to ensure equitable and fair outcomes for all systems.

2.1 Types of Bias

Contrary to what is typically believed in research, data gathered from people in the real world is not homogeneous. The model could be biased by the demographics of the people who labeled the data. Real-world data is diverse because it originates from social subgroups with unique traits and behaviors. Therefore, it is essential to identify what kind of biases exist in the data in order to mitigate them. These biases can be grouped into four major categories (Shah et al., 2019):

1. **Label Bias:** It occurs due to erroneous labeling of the data by the annotators. This can happen if the annotators hold preconceived notions or stereotypes about the domain of the data.

2. **Selection bias:** It emerges due to non-representative observations – when the annotators generating the training data have a different distribution than where the model is to be applied. *A famous example is the "Wall Street Journal effect," where syntactic parsers and part-of-speech taggers perform most accurately over language written by middle-aged white men.* (Garimella et al., 2019)

3. **Semantic Bias:** Embeddings have become an essential component of modern NLP, with their ubiquitous applications in both classical and deep learning models. These representations, however, frequently incorporate unintentional or negative connotations and stereotypes. For example, certain words or phrases may be associated with one group more than another, leading to biased results. *For instance, the word "boss" has closer representation compared to men than women.*

4. **Overamplification:** In overamplification, the model picks up small differences between human attributes with respect to the target, and amplifies this difference to be more significant in the predicted outcomes. This usually happens in the model learning phase itself. *For example, Zhao et al. (2017) found that in the imSitu image captioning data set, the activity cooking is over 33% more likely to involve females than males in a training set, and a trained model further amplifies the disparity to 68% at test time.*

3 Black Box Models and Explainability

In science, computing, and engineering, a black box is a system that can be viewed in terms of its inputs and outputs without any knowledge of its internal workings. Its implementation is "opaque" (black).[1] Some examples of black-box models are Random Forests and Deep Neural Networks since they have complex internal structures.

Explainability in ML attempts to make users understand how the model predicts an output. It

[1] https://en.wikipedia.org/wiki/Black_box

helps provide information about how and why a model made a specific prediction. By understanding the inner workings of the model, explainability helps build trust and facilitates the adoption of ML systems in various domains. It may be trivial to understand the mechanism of white-box (transparent) models like Linear Regression and Decision Trees due to their simple structure. However, when it comes to black-box models like Random Forests and Deep Neural Networks, explainability is a difficult task, even for experts in this domain, due to its complex internal structure. To achieve explainability, two primary approaches are commonly used:

- **Local Explainability**: This tells us about the model's behavior at a particular instance and how each individual feature affects the model's prediction.
- **Global Explainability**: This tells us about the overall behavior of the model and how all the features combined affect the model's prediction.

Lipton (2018) remarked upon the following points to understand explainability better.

3.1 Transparency

Informally, transparency is the opposite of opacity or black box-ness. It connotes some sense of understanding the mechanism by which the model works. We consider transparency at the level of the entire model (simulatability), at the level of individual components, e.g., parameters (decomposability), and at the level of the training algorithm (algorithmic transparency).

3.1.1 SIMULATABILITY

One can call a model transparent if a human can contemplate the entire model at once. However, to fully understand the model, a human should be able to take the input data together with the parameters of the model and, in a reasonable time, through every calculation required to produce a prediction. Ribeiro et al. (2016) also takes on this idea of interpretability, suggesting that an interpretable model is one that "can be readily presented to the user with visual or textual artifacts." For some models, such as decision trees, the size of the model (total number of nodes) may grow much faster than the time to perform inference (length of pass from root to leaf). This suggests that simulatability may admit two subtypes, one based on the total size of the model and another based on the computation required to perform inference.

3.1.2 DECOMPOSABILITY

A second notion of transparency might be that each part of the model - input, parameter, and calculation, provides an intuitive explanation. *For example, the coefficients of a linear model can describe the strengths of association between each feature and the label.* However, one cannot blindly trust this notion of transparency.

3.1.3 ALGORITHMIC TRANSPARENCY

The third notion of transparency applies at the level of the learning algorithm itself. For example, in the case of linear models, we understand the shape of the error surface, which can help us prove that the training will converge to a unique solution. However, in contrast, deep learning methods lack this sort of algorithmic transparency. While neural networks and deep learning systems provide remarkable performance, we cannot comprehend the decision-making process of complex black-box models.

3.2 Post-hoc interpretability

Post hoc interpretability refers to the process of explaining the behavior and decisions of a black-box machine learning model after it has been trained. It involves using various techniques and tools to analyze the model's internal workings and generate human-understandable explanations for its predictions. Some common approaches to post-hoc interpretations include natural language explanations, visualizations of learned models and explanations by example. We also show an overview of post-hoc techniques with their types in Figure 1. [2]

3.3 Model Agnostic Explanaibility

When we talk about black-box models, we include all possible kinds of models which take input and return an output with complex internal mechanisms. However, there are some explainability techniques that work on all kinds of models, known as model-agnostic explainability. These techniques provide insight into how machine learning models make decisions without requiring access to the model's internal structure and are independent of the model used, which is why they are referred to as "model-agnostic". The advantage of the universal applica-

[2] https://www.ambiata.com/blog/2021-04-12-xai-part-1/

Technique	Local	Modular Global	Global	Model-specific	Model-agnostic	Example based
Partial Dependence Plots [PDP]		✓			✓	
Individual Conditional Expectation [ICE]		✓			✓	
Accumulated Local Effects [ALE]		✓			✓	
Anchors [ANC]	✓				✓	
Permutation Feature Importance [PMP1, PMP2]			✓		✓	
Integrated Gradients [IG]	✓			✓		
Local interpretable model-agnostic explanations [LIME]	✓				✓	
Kernel SHAP [SHAP]	✓		✓		✓	
Tree SHAP [TSHAP]	✓		✓	✓		
Counterfactual Explanations [CE]	✓				✓	✓
Prototype Counterfactuals [PC]	✓				✓	✓
Adversarial Examples [AE]	✓				✓	✓

Figure 1: Some of the popular techniques for explaining ML models, varying in complexity, applicability, and the type of information they provide.

tion of these techniques to all kinds of models has piqued the interest of researchers.

We will now discuss the model-agnostic techniques which are presently used for feature analysis for models. The Permutation Feature Importance is a technique that involves randomly permuting the values of individual input features and measuring the resulting decrease in the model's accuracy. By comparing the feature importance scores across multiple permutations, it is possible to identify which features are most important for the model's predictions. Partial Dependence Plot (PDP) is another model-agnostic technique that can help visualize how the predicted outcome changes as a function of one or more input features while holding all other features constant. We will further look at explainability techniques described in the literature for feature attribution. **LIME** (Linear Interpretable Model Explanations) by Ribeiro et al. (2016) is a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner by learning a surrogate interpretable model locally around the prediction. **SHAP** (SHapley Additive exPlanations) (Lundberg and Lee, 2017) is an approach inspired by game theory to explain the output of any black-box function by assigning each feature an importance value for a particular prediction. **Anchors** (Ribeiro et al., 2018) is a model-agnostic system that explains the behavior of complex models with high-precision rules called anchors, representing local, sufficient conditions for predictions.

3.4 Model Specific Explainability

In contrast to model-agnostic techniques described in the previous section, there also are some explainability techniques specific to certain kinds of models. These are Model-specific explainability techniques that work based on the details of the specific structures of the machine learning or deep learning model which is applied. These strategies are explicitly employed for a certain model design, such as the neural network, and employ a reverse engineering approach to explain how the specific Deep Learning (DL) algorithm is making the relevant decision.

We will now discuss some of the popular model-specific techniques used in literature, with more emphasis on explainability for deep learning models. **Integrated Gradients** by (Sundararajan et al., 2017) is a method for attributing the contribution of each input feature to a model's output by integrating the gradients of the output with respect to the input features. The advantage of Integrated Gradients is that it is based on an axiomatic approach for explaining the prediction of deep neural networks without any modification to the original network. **DeepLIFT** (Deep Learning Important FeaTures) by (Shrikumar et al., 2017) is a method for decomposing the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input. DeepLIFT compares the activation of each neuron to its reference activation and assigns contribution scores according to the difference. **Layer-wise Relevance Propagation (LRP)** by (Montavon et al., 2019) is another model-specific technique that highlights the input features supporting the prediction by propagating the prediction backward in deep neural networks. LRP is able to provide a detailed and fine-grained understanding of how each feature contributes to the model's output.

4 Explainability for Bias Detection

There are a number of papers that mention unintended or societal biases as wider motivations to contextualize the work for bias detection; however, only a handful of them apply explainability techniques to uncover or investigate biases. We mention some of the recent works done in this area, as referred from Balkir et al. (2022a). We focus specifically on explainability methods using feature attribution strategies.

(Balkir et al., 2022b) used a feature attribution method for explaining text classifiers and analyzed them in the context of hate speech detection. They showed that sufficiency and necessity could be used to explain the expected differences between a classifier that is intended to detect identity-based hate speech and those trained for detecting general abuse. (Mathew et al., 2021) introduced a new benchmark dataset for hate speech detection called HateXplain. Their work aims to improve the explainability of hate speech detection models by providing a dataset that includes not only hate speech texts but also explanations of why each text is considered hateful using LIME and Attention. (Aksenov et al., 2021) presents a new dataset for fine-grained classification of political bias in German news articles. The authors also analyzed the contribution of different linguistic features to the prediction of political bias using aggregated attention scores. (Mosca et al., 2021) explores the impact of user context on hate speech detection models. The paper argues that user contexts, such as the demographic information and previous behavior of users, can affect the interpretation and classification of hate speech. They use SHAP and feature space exploration to explain their model behavior. (Wich et al., 2020) examines how politically biased data affects the performance of hate speech detection models. The authors show that models trained on politically biased data can lead to biased models that perform poorly in detecting hate speech against certain political groups using SHAP to explain the models. (Prabhakaran et al., 2019) introduced Perturbation Sensitivity Analysis to test for unwanted biases in an NLP model. They demonstrate the utility of their framework on online comments in the English language from four different genres for sentiment and toxicity models.

4.1 Current practices

In this section, we mention a few of the current practices and research work going on in the domain of explainability for bias detection.

(a) **Counterfactual explanations:** These methods generate alternative inputs to a model that would result in different outputs, allowing users to understand how changes in input features affect model predictions. Counterfactual explanations (Sokol and Flach, 2019) can be used to detect and mitigate biases by identifying which features are most influential in driving model predictions and how changing those features can lead to fairer outcomes.

(b) **Extractive Rationales** (DeYoung et al., 2019) are snippets of the input text that trigger the original prediction. They are similar in spirit to feature attribution methods, however, in rationales, the attribution is usually binary rather than a real-valued score, and continuous subsets of the text are chosen rather than each token being treated individually.

(c) **Attention mechanisms** (Choi et al., 2016) allow users to visualize which parts of an input are most important for a model's prediction. Attention mechanisms can be used to detect biases by identifying which parts of the input are being ignored or given less weight by the model, potentially leading to unfair outcomes.

(d) **Adversarial training:** The Adversarial training technique (Zhang et al., 2018) involves training models on adversarial examples that are designed to expose and correct biases. This method can be used to detect and mitigate biases by forcing models to learn more robust decision boundaries that are less susceptible to adversarial attacks.

(e) **Model interpretation techniques** (e.g., LIME, SHAP): These methods provide local or global explanations for model predictions, allowing users to understand how individual instances or groups of instances are being classified. Model interpretation techniques can be used to detect biases by identifying which features or groups of instances are being treated unfairly by the model.

5 Datasets

In this section, we will specify a comprehensive overview of datasets that can be used to study bias in the area of machine learning.

1. **Adult Dataset:** [3] It is also known as the Census Income dataset. This dataset comes from the UCI repository of machine learning databases. The task is to predict if an individual's annual income exceeds $50,000 based on census data. It contains a total of 45,225 cases and 16 attributes. It can be used in studies concerned with the fairness of gender-based inequalities based on the yearly income of people.

2. **COMPAS Dataset:** [4] The COMPAS dataset is commonly used to predict a defendant's likelihood of reoffending within the next two years. It comprises 6,172 instances, with 13 features including age, sex, race, and prior convictions. The dataset is originally gathered by ProPublica.

3. **German Credit Dataset:** [5] The German Credit dataset is used for credit risk assessment, where the goal is to predict whether an individual will default on a loan based on various features such as credit history and employment status. The dataset comprises 1,000 instances with 20 attributes. The German Credit dataset can be applied to research gender disparities in credit-related matters.

4. **WinoBias Dataset:** The WinoBias dataset by (Zhao et al., 2018) contains 3,160 sentences, which follows the Winograd format and is centered on people entities referred by their occupations from a vocabulary of 40 occupations. There are mainly two types of sentences in the dataset requiring linkage of gendered pronouns to either male or female stereotypical occupations. It has been used in the study of coreference resolution to certify if a system has a gender bias.

5. **Recidivism in Juvenile Justice Dataset:** The Recidivism in Juvenile Justice dataset (Tolan et al., 2019) contains all juvenile offenders between ages 12-17 who committed a crime between the years 2002 and 2010 and completed a prison sentence in 2010 in Catalonia's juvenile justice system.

6. **Communities and Crime Dataset:** [6] The Communities and Crime dataset gathers information from different communities in the United States related to several factors that can highly influence some common crimes such as robberies, murders, or rapes. The data includes crime data obtained from the 1990 US LEMAS survey and the 1995 FBI Unified Crime Report. It also contains socio-economic data from the 1990 US Census.

7. **Pilot Parliaments Benchmark Dataset:** The Pilot Parliaments Benchmark dataset (Buolamwini and Gebru, 2018), also known as PPB, contains images of 1270 individuals in the national parliaments of three European (Iceland, Finland, Sweden) and three African (Rwanda, Senegal, South Africa) countries. This benchmark was released to have more gender and race balance, diversity, and representativeness.

8. **Diversity in Faces Dataset:** The Diversity in Faces (DiF) (Merler et al., 2019) is an image

[3] http://www.cs.toronto.edu/~delve/data/adult/desc.html

[4] https://github.com/propublica/compas-analysis

[5] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[6] https://archive.ics.uci.edu/ml/datasets/communities+and+crime

Dataset Name	Size	Area
UCI adult dataset	48,842 income records	Social
German credit dataset	1,000 credit records	Financial
Pilot parliaments benchmark dataset	1,270 images	Facial images
WinoBias	3,160 sentences	Coreference resolution
Communities and crime dataset	1,994 crime records	Social
COMPAS Dataset	18,610 crime records	Social
Recidivism in juvenile justice dataset	4,753 crime records	Social
Diversity in faces dataset	1 million images	Facial images

Table 1: Most widely used datasets in the fairness domain with additional information about each of the datasets, including their size and area of focus (Mehrabi et al., 2021)

dataset collected for fairness research in face recognition. DiF is a large dataset containing one million annotations for face images. It is also a diverse dataset with diverse facial features, such as different craniofacial distances, skin color, facial symmetry and contrast, age, pose, gender, and resolution, along with diverse areas and ratios.

6 Challenges and Future Directions

In this section, we discuss some challenges and limitations in the area of bias and explainability and suggest promising directions for future work.

Having a common definition for fairness: The literature in the fairness domain presents various kinds of definitions of what fairness would mean from a machine learning standpoint. Consequently, it becomes nearly impossible to comprehend how one fairness solution would fare under a different definition of fairness. Therefore, having a common definition of fairness remains an open question to researchers since it can make the evaluation of systems more homogeneous and unified.

Local explainability methods rely on the user to identify examples that might reveal bias: A key step in discovering fairness issues in a machine-learning model is to identify the set of possible data instances where these issues may arise. Since local explainability approaches provide explanations for specific data instances, it is up to the user to choose which instances need to be investigated. As a result, before using XAI methods, the user must first decide what biases to search for, thereby limiting its effectiveness for finding unknown biases.

Less generalizability of local explanations: With the rise of complex problems in NLP, it is often hard to explain models globally which are designed for solving these problems. This has led to researchers adopting methods of local explainability to understand the working of the models. However, one issue that is faced by using XAI methods for fairness is that it is difficult to know to what extent the local explanations can be generalized. As local explanations provide reasoning for specific data points, it becomes incoherent to identify how explanations generalize the model globally. Although there exist some methods like Anchors (Ribeiro et al., 2018) which can tackle the aforementioned problem and mitigate it by specifying the set of examples to which the explanation applies. In addition, future NLP research could explore global explainability methods that have been used to uncover unknown biases. (Tan et al., 2018)

Some biases can be difficult for humans to recognize: It can be seen that XAI methods rely on humans to recognize what an undesirable correlation is; however, biased models are often nuanced in exhibiting bias. *For example, if the dialect bias in a hate speech detection system is mostly mediated by false positives on the uses of reclaimed slurs, this might seem like a good justification to a user who is unfamiliar with this phenomenon* (Sap et al., 2019). Therefore, this encourages more investigation and research into whether humans can recognize unintended biases that cause fairness issues through explainability methods.

Explainability methods are susceptible to fairwashing: The possibility of "fairwashing" biased models has been repeatedly emphasized in relation to XAI approaches. Fairwashing refers to strategies that use adversarial manipulation of explanations to disguise the model's reliance on protected attributes. Fairwashing has been shown to be possible in rule lists (Aïvodji et al., 2019) and both gradient-based and perturbation-based feature attribution methods (Dimanov et al. (2020); Anders et al. (2020)). This has raised some concerns about the faithfulness of explainability methods. In regards to this, there have been a few solutions proposed, like developing certifiably faithful explainability methods with proofs that a particular way of testing for bias cannot be adversarially manipulated (Cohen et al. (2019); (Ma et al., 2020)), providing more information on whether the user can trust the generated explanation (Zhang et al., 2019) or other ways to calibrate user trust to the quality of the provided explanations (Zhang et al., 2020). Overall, this challenge suggests that additional steps need to be taken to ensure the robustness of the explanations.

7 Summary

In this paper, we discussed the idea of bias in data and its types. Further, we elucidated the topic of explainability and mentioned different ways of interpreting a black-box model. We found that the combination of explainability for bias detection has been used mostly in the hate-speech tasks, whereas its use in other areas has been less explored. We also summarize the list of popular datasets which can be used to evaluate frameworks in the fairness domain. Finally, we look at the current challenges in applying explainability for bias detection and provide promising directions for future work in this area.

References

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR.

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno Schneider, and Georg Rehm. 2021. Fine-grained classification of political bias in german news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131.

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. 2020. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pages 314–323. PMLR.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2022a. Challenges in applying explainability methods to improve the fairness of nlp models. *arXiv preprint arXiv:2206.03945*.

Esma Balkir, Isar Nejadgholi, Kathleen C Fraser, and Svetlana Kiritchenko. 2022b. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. *arXiv preprint arXiv:2205.03302*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. 2020. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Pingchuan Ma, Shuai Wang, and Jin Liu. 2020. Metamorphic testing and certified mitigation of fairness violations in nlp models. In *IJCAI*, pages 458–465.

Niteesh Mallela and Pushpak Bhattacharyya. 2022. Survey: Bias in nlp.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of* the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. *arXiv preprint arXiv:1901.10436*.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102.

Ravi B Parikh, Stephanie Teeple, and Amol S Navathe. 2019. Addressing bias in artificial intelligence in health care. *Jama*, 322(24):2377–2378.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Deven Shah, H Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Kacper Sokol and Peter A Flach. 2019. Counterfactual explanations of machine learning predictions: Opportunities and challenges for ai safety. *SafeAI@ AAAI*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310.

Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 83–92.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. " why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.