Fairness and Disentanglement: A Critical Review of Predominant Bias in Neural Networks

Jiazhi Li *jiazhil@usc.edu*
Ming Hsieh Department of Electrical and Computer Engineering
University of Southern California

Mahyar Khayatkhoei *mkhayat@isi.edu*
Information Sciences Institute
University of Southern California

Jiageng Zhu *jiagengz@usc.edu*
Ming Hsieh Department of Electrical and Computer Engineering
University of Southern California

Hanchen Xie *hanchenx@usc.edu*
Thomas Lord Department of Computer Science
University of Southern California

Mohamed E. Hussein *mehussein@apple.com*
Apple Inc.

Wael AbdAlmageed *wabdalm@clemson.edu*
Holcombe Department of Electrical and Computer Engineering
Clemson University

Reviewed on OpenReview: *https://openreview.net/forum?id=LLiJ1WsL2e&referrer=%5BAuthor%20Console%5D(%2Fgroup%3Fid%3DTMLR%2FAuthors%23your-submissions)*

Abstract

Bias issues of neural networks garner significant attention along with their promising advancement. Among various bias issues, mitigating two predominant biases is crucial in advancing fair and trustworthy AI: (1) ensuring neural networks yield even performance across demographic groups, and (2) ensuring algorithmic decision-making does not rely on protected attributes. However, upon the investigation of 415 papers in the relevant literature, we find that there exists a persistent, extensive but under-explored confusion regarding these two types of biases. Furthermore, the confusion has already significantly hampered the clarity of the community and the subsequent development of debiasing methodologies. Thus, in this work, we aim to restore clarity by providing two mathematical definitions for these two predominant biases and leveraging these definitions to unify a comprehensive list of papers. Next, we highlight the common phenomena and the possible reasons for the existing confusion. To alleviate the confusion, we provide extensive experiments on synthetic, census, and image datasets to validate the distinct nature of these biases, distinguish their different real-world manifestations, and evaluate the effectiveness of a comprehensive list of bias assessment metrics in assessing the mitigation of these biases. Further, we compare

these two types of biases from multiple dimensions, including the underlying causes, debiasing methods, evaluation protocol, prevalent datasets, and future directions. Last, we provide several suggestions aiming to guide researchers engaged in bias-related work to avoid confusion and further enhance clarity in the community.

1 Introduction

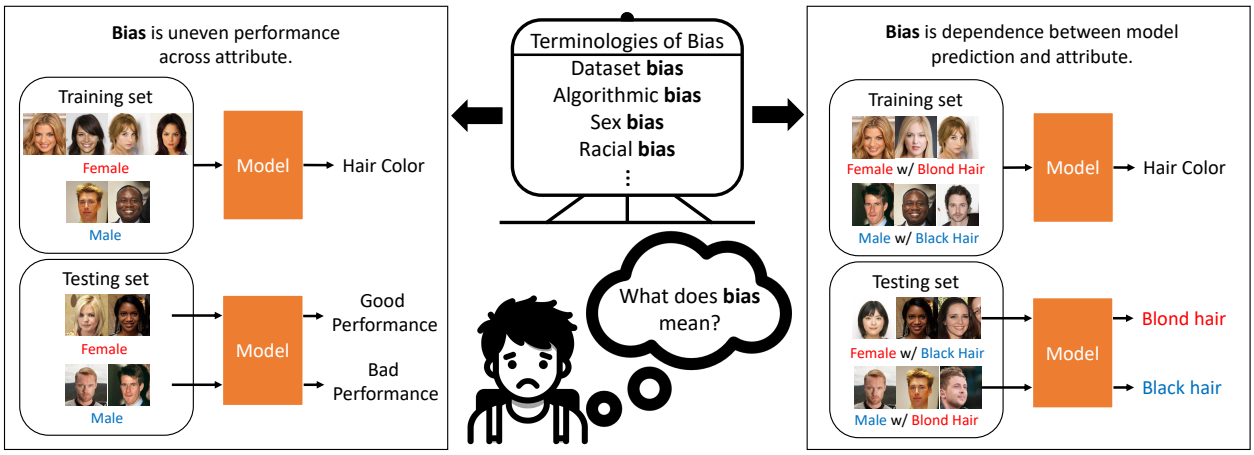


Figure 1: The same set of terminology about bias is interpreted differently by experts, which significantly confuses the understanding of the audience. By investigating 415 papers about prevalent bias issues, we discover that there exists significant confusion regarding these prevalent bias issues. The confusion is evident in several ways, such as ambiguity of terminology, inaccurate motivation, and lack of terminology reuse. Most notably, several studies inaccurately motivate themselves on a particular bias while actually addressing a different type of bias. This prevailing confusion considerably impedes the clarity of related work. Thus, we propose new definitions to unify the existing literature and pave a clear path for future research.

Neural networks have shown promising advances in many prediction and classification tasks (Russakovsky et al., 2015; He et al., 2016; Mnih et al., 2013). Along with the impressive capability of neural networks, its societal impact has garnered great attention (Buolamwini & Gebru, 2018; Gong et al., 2021), particularly regarding *protected attributes* (*e.g.*, sex, race, and age), which cannot be used in the decision-making process (Corbett-Davies & Goel, 2018). Failing to carefully consider protected attributes while deploying neural networks can lead to bias issues and severely compromise fairness for specific demographic groups in various real-world applications (Angwin et al., 2022a; Li & AbdAlmageed, 2024). For instance, facial recognition systems may more correctly recognize males than females (Gong et al., 2020). Besides, Artificial Intelligence-assisted bank loan systems may classify a higher proportion of male applicants as having bad credit than female applicants (Zhu et al., 2021).

The underlying bias issues of neural networks, involved in the aforementioned examples, lead to important discussions (Kim et al., 2019a; Liu et al., 2022a; Tartaglione et al., 2021). Specifically, these aforementioned examples highlight the presence of two distinct prevalent types of biases. Without loss of generality, for disambiguation, these two predominant biases can be summarized as follows:

Table 1: Main distinctions between Type I Bias and Type II Bias.

	Type I Bias	Type II Bias
Manifestation	Uneven performance across attributes	Dependence between model prediction and attribute
Use of ground truth	✓	✗
Representative example	Facial recognition systems exhibit lower performance in one demographic group compared to others	Bank loan systems tend to approve loans more frequently for one demographic group compared to others
Possible reason	Insufficient training in the underrepresented group	Correlation between the target Y and the attribute in the training set

- The model yields uneven performance across different demographic attributes, referred to as *Type I Bias*.
- The model depends on demographic attributes to make predictions, referred to as *Type II Bias*.

Although these two prevalent types of biases differ in many aspects, as highlighted in Tab. 1, the current literature often ambiguously groups them under the general term "bias" (*e.g.*, dataset bias, algorithmic bias, sex bias, or racial bias) (Alvi et al., 2018; Ragonesi et al., 2021) and interpret them differently across scenarios. Furthermore, numerous works addressing one type of bias inadvertently cite the other as their motivation (Wang et al., 2019b; Wang & Deng, 2020; Salvador et al., 2022). Additionally, existing survey papers may lack a clear taxonomy that sufficiently distinguishes between them or explicitly acknowledge their fundamental distinctions (Mehrabi et al., 2021a; Wang et al., 2022c; Castelnovo et al., 2022).

Overlooking the distinction between these two types of biases significantly compromises clarity in the current literature and leads to various negative consequences. Specifically, for new researchers, the lingering question of which specific type of bias a paper addresses creates unnecessary confusion. As illustrated in Fig. 1, this confusion arises when researchers encounter multiple works that all reference "bias" without clearly specifying its nature. The widespread confusion surrounding these biases and the lack of clear definitions to separate them results in weak motivation, ambiguous statements, and vague contributions in the existing debiasing work, significantly impeding the clarity of the associated research. Additionally, persistent conflation of these biases, usage of inappropriate references, and unfair comparison between methods addressing different biases can lead to an expanding misunderstanding over time. Besides, this confusion complicates the resolution of bias issues and hinders the advancement of future work in this field.

To that end, the main goal of this paper is to unify the existing literature about Type I Bias and Type II Bias, rectify the common confusion regarding them, and alleviate the cognitive burden for future research. The contributions of this paper can be summarized as follows:

- Proposing General mathematical definitions for Type I Bias and Type II Bias (Sec. 2) and providing a summary of their corresponding related work (Sec. 7). These can be utilized as a roadmap for future work.
- Unifying a comprehensive list of work and relevant fairness criteria under the definition of Type I Bias and Type II Bias (Sec. 4).
- Elucidating the existing phenomena stemming from the confusion between Type I Bias and Type II Bias (Sec. 5.1), and exploring the underlying reasons that contribute to the confusion (Sec. 5.2).
- Conducting extensive experiments to examine the distinction between Type I Bias and Type II Bias (Sec. 6).
- Offering some suggestions to foster a clear community regarding these bias issues (Sec. 8).

2 Definitions

To define and distinguish these two types of biases, we first establish several key concepts. Let the dataset $\mathcal{D}$ be a set of instances (x, y, z) where each sample $x \in \mathcal{X}$ is annotated with a ground truth label $y \in \mathcal{Y}$ for a downstream task (*e.g.*, identity in face recognition) and an attribute label $a \in \mathcal{A}$ (*e.g.*, sex). A model $f : \mathcal{X} \to \mathcal{Y}$ maps an input x to a predicted label $\hat{y}$. In this section, we introduce formal mathematical definitions for these two types of biases, referred to as Type I Bias and Type II Bias, which will be consistently used throughout the paper. In the following sections, we will review 415 papers to demonstrate that various commonly discussed bias issues can be unified using these definitions and explore the phenomena and reasons behind the existing confusion between these bias issues. Furthermore, in Sec. 6, we present a case study using both synthetic and real-world datasets to demonstrate scenarios in which Type I Bias occurs without Type II Bias, and vice versa.

2.1 Type I Bias

The manifestation of Type I Bias is uneven model performance across different demographic groups (Wang et al., 2019b; Gong et al., 2021; Liu et al., 2022a). Specifically, model performance can be evaluated using various metrics, *e.g.*, error rate (Buolamwini & Gebru, 2018), loss (Hashimoto et al., 2018), accuracy (Kim et al., 2019b), average precision (AP) (Ramaswamy et al., 2021), positive predictive value (PPV), true positive rate (TPR) (Dhar et al., 2021; Adeli et al., 2021), false positive rate (FPR) (Xu et al., 2021c), average false rate (AFR), mean AFR (M AFR) (Ryu et al., 2017), confusion matrix (Gong et al., 2020), F1 score (Adeli et al., 2021), receiver operating characteristic curve (ROC) (Wang & Deng, 2020), area under the ROC (AUC) (Mirjalili et al., 2019). All these metrics can be unified under the format of a distance measure $d(\hat{Y}, Y)$, evaluated based on model prediction $\hat{Y}$ and ground truth label Y. Thus, we can formally define this type of bias as follows:

Definition 1. *Type I Bias. A model f involves* Type I Bias *if f yields uneven performance $d(\hat{Y}, Y)$ across attribute A,*

$$\sup_{a,a' \in \mathcal{A}, d \in \mathcal{M}} |d(\hat{Y}, Y|A = a) - d(\hat{Y}, Y|A = a')| > 0 \tag{1}$$

where a, a' are possible values of A (*e.g.*, female and male), and $\mathcal{M}$ is the set of all potential performance metrics.

2.2 Type II Bias

On the other hand, the manifestation of Type II Bias is dependence between model prediction and attribute (Alvi et al., 2018; Kim et al., 2019a; Wang et al., 2020b; Nam et al., 2020). Specifically, these attributes can be categorized by sensitive/protected attributes (Lokhande et al., 2020) (*e.g.*, sex in creditworthiness prediction) or spurious attributes (Sagawa* et al., 2020) (*e.g.*, texture in object recognition). Both of these scenarios can be unified as the dependence between model prediction and the specific attribute. Thus, we can formally define this type of bias as follows:

Definition 2. *Type II Bias. A model f involves* Type II Bias *if model prediction $\hat{Y}$ is not independent with attribute A,*

$$\sup_{a,a' \in \mathcal{A}} |P(\hat{Y}|A = a) - P(\hat{Y}|A = a')| > 0 \tag{2}$$

where a, a' are possible values of A (*e.g.*, female and male).

3 Method

In this section, we introduce the method used to conduct the investigation on a set of 415 papers that discuss relevant bias issues. Specifically, to construct the initial set of relevant work, we search for the keywords "bias" or "fair" in the title of papers from NeurIPS, ICML, ICLR, and FAccT published before February 2025. We include papers that discuss bias issues whose manifestation aligns with either Type I Bias or Type II Bias (we will detail the unification in Sec. 4). We exclude papers that address other bias issues such as inductive bias (Baxter, 2000; Zietlow et al., 2021), implicit bias (FitzGerald & Hurst, 2017; Camuto et al., 2021), selection bias (Hernán et al., 2004; Akbari et al., 2021), sampling bias (Winship & Mare, 1992; Xu et al., 2022a), spectral bias (Fang & Xu, 2024), exposure bias (Li et al., 2024) or bias-variance (Ha et al., 2024; Chen et al., 2024b). Furthermore, to ensure we do not overlook any relevant papers without these keywords or from other prominent conferences such as CVPR, ICCV, and ECCV, we manually traversal the citation graph of the paper in the initial set and append the relevant papers that are either cited by or cite the papers in the initial set.

Once we identify the scope of the investigated papers, we read these papers to determine which type of bias they address by examining two aspects: the problem statement and evaluation protocol. We will elaborate on the criterion for categorizing papers into our definitions in Sec. 4. To accommodate the recent emerging

direction of addressing unlabeled and unknown bias, we enrich the taxonomy with an additional dimension about the status of attribute A. As shown in Tab. 2, we count the number of papers in each category. Note that the total number is not equal to 415 since some papers address both types of biases. We present the categorization list of all 415 investigated papers in the Appendix.

Table 2: The taxonomy of bias issues based on 415 papers.

Type of Bias	Attribute A		Papers	Examples
	Known	Labeled		
Type I Bias	✓	✓	253	Gong et al. (2020; 2021); Wang & Deng (2020)
	✓	✗	-	-
	✗	✗	-	-
Type II Bias	✓	✓	246	Kim et al. (2019a); Zhu et al. (2021); Tartaglione et al. (2021)
	✓	✗	8	Wang et al. (2019a); Bahng et al. (2020); Cadene et al. (2019)
	✗	✗	30	Nam et al. (2020); Zhao et al. (2023a); Jeon et al. (2022)
Survey	-	-	25	Mehrabi et al. (2021a); Du et al. (2020); Castelnovo et al. (2022)

4 Unification

In this section, we clarify how bias issues discussed in existing literature align with our proposed definitions. Generally, we categorize the bias into a specific type of bias in our definition if the presence of this bias implies the existence of bias in our definition. Furthermore, the categorization primarily relies on two key factors: the manifestation of bias issues explicitly addressed (if stated in "Problem Statement" section) and the characteristics of evaluation protocol[1]. Other aspects, such as motivation, related work, method, or bias assessment, are considered secondary factors for categorization. This is because certain papers, despite addressing different manifestations of bias, can exhibit similarities in these aspects, thereby leading to the confusion between these two types of biases, as elaborated in Sec. 5.1.

4.1 Type I Bias

The general form of Type I Bias is characterized by the uneven performance of the target across attributes. This definition can be extended to unify a wide range of papers by specifying the usage of performance metrics and the kind of target. To clarify, several representative descriptions are shown as follows, *e.g.*,

- *"Racial bias indeed degrades the fairness of recognition system and the error rates on non-Caucasians are usually much higher than Caucasians."* (Wang & Deng, 2020)
- *"A certain demographic group can be better recognized than other groups."* (Gong et al., 2021)
- *"Recognition accuracies depend on the demographic cohort."* (Wang et al., 2019b)

By specifying how performance is evaluated, Type I Bias covers a broad range of papers where model performance is evaluated using various criteria such as loss (Hashimoto et al., 2018) and accuracy (Kim et al., 2019b). Furthermore, by specifying the kind of target, this definition can unify a wider range of papers. For instance, considering sex as an attribute, the targets can include identity (Gong et al., 2020; Salvador et al., 2022) (*e.g.*, face recognition), the attribute itself (Buolamwini & Gebru, 2018; Karkkainen & Joo, 2021) (*e.g.*, sex classification), or other targets associated with protected attribute (Stone et al., 2022; Hashimoto et al., 2018) (*e.g.*, facial attribute classification). It is noteworthy that Type I Bias is predominantly discussed in various biometrics tasks (Conti et al., 2022; Klare et al., 2012; Morales et al., 2020). Compared with various types of targets, protected attributes (*e.g.*, sex, race, and age) are mainly considered the term of attribute in Type I Bias.

[1]For instance, Type I Bias involves training sets which yield the long-tail distribution, while Type II Bias typically involves training sets which yields the association between target label and attribute label.

4.2 Type II Bias

The general form of Type II Bias is characterized by the dependence between model prediction and attribute. This definition can be used to unify a broad spectrum of papers by considering the status of the attribute and the kind of attribute. The status of the attribute is categorized into three groups: known and labeled, known but unlabeled, and unknown. Specifically, for known and labeled bias, several methods directly leverage attribute labels to explicitly apply a supervision signal for bias mitigation (Zhu et al., 2021). For known but unlabeled bias, several methods mainly utilize the domain knowledge of specific bias attributes to design the module tailored for this bias attribute (Wang et al., 2019a). For unknown bias, several methods identify and emphasize bias-conflicting samples (those exhibiting the opposite bias present in the training set) to mitigate bias (Zhao et al., 2023a). On the other hand, the kind of attribute mainly encompasses sensitive/protected attributes (Angwin et al., 2022b; Chen & Joo, 2021; Calders & Verwer, 2010) and spurious attributes (Nam et al., 2020; Sagawa* et al., 2020; Zhao et al., 2023a). In the case of sensitive attributes, the reliance on them leads to a disproportionate assignment of specific predictions to particular demographic groups, thereby resulting in unfair treatment. In this category, demographic parity (Dwork et al., 2012), a well-known fairness criterion, is often served as a debiasing objective. We present several representative descriptions as follows, *e.g.*,

- *"Demographic parity, which is satisfied when the predictions are independent of the sensitive attributes."* (Creager et al., 2019)
- *"Data fairness can be achieved if the generated decision has no correlation with the generated protected attribute."* (Xu et al., 2018)
- *"Ensuring that the positive outcome is given to the two groups at the same rate."* (Madras et al., 2018)

In the case of spurious attributes, depending on them for decision-making will simplify the training process since models may utilize them as shortcut features instead of learning more comprehensive features during training. However, this leads to model predictions heavily relying on these attributes and further poor generalization performance in real-world applications since such spurious correlation between target and attribute does not generally exist. Several representative descriptions are shown as follows, *e.g.*,

- *"If bias features are highly correlated with the object class in the dataset, models tend to use the bias as a cue for the prediction."* (Hong & Yang, 2021)
- *"Since there are correlations between the target task label and the bias label, the target task is likely to rely on the bias information to fulfill its objective."* (Zhu et al., 2021)
- *"If biased data is provided during training, the machine perceives the biased distribution as meaningful information."* (Kim et al., 2019a)

Table 3: The summary of representative fairness criteria.

Category	Notion	Definition	Examples
Fairness w.r.t. Type I Bias	Equalized odds (Hardt et al., 2016)	$P(\hat{Y}=y_1\|A=a_0,Y=y)=P(\hat{Y}=y_1\|A=a_1,Y=y), y\in\{y_0,y_1\}$	Park et al. (2022) Zhang et al. (2023) Conti et al. (2022)
	Equal opportunity (Hardt et al., 2016)	$P(\hat{Y}=y_1\|A=a_0,Y=y_1)=P(\hat{Y}=y_1\|A=a_1,Y=y_1)$	Jung et al. (2022) Yu et al. (2022) Pham et al. (2023)
	Accuracy parity (Quan et al., 2023)	$P(\hat{Y}=Y\|A=a_0)=P(\hat{Y}=Y\|A=a_1)$	Kim et al. (2019b) Zafar et al. (2017a) Quan et al. (2023)
Fairness w.r.t. Type II Bias	Demographic parity (Dwork et al., 2012; Kusner et al., 2017)	$P(\hat{Y}\|A=a_0)=P(\hat{Y}\|A=a_1)$	Creager et al. (2019) Xu et al. (2018) van Breugel et al. (2021)

Table 4: The overview of the literature regarding Type I Bias and Type II Bias.

Category	Description	Subsettings		Examples
Type I Bias	Uneven performance of target across attribute	How is performance evaluated?	Error rate	Buolamwini & Gebru (2018) Sattigeri et al. (2019)
			Loss	Hashimoto et al. (2018)
			Accuracy	Kim et al. (2019b)
			Average precision	Ramaswamy et al. (2021)
			True positive rate	Dhar et al. (2021) Adeli et al. (2021)
			False positive rate	Xu et al. (2021c)
			Mean average false rate	Ryu et al. (2017)
			Confusion matrix	Gong et al. (2020)
			F1 score	Adeli et al. (2021)
			Receiver operating characteristic curve (ROC)	Mirjalili et al. (2018) Qin (2020) Yu et al. (2020)
			Area under the ROC (AUC)	Mirjalili et al. (2019) Gong et al. (2020) Adeli et al. (2021)
		Type of target	Identity	Wang & Deng (2020) Wang et al. (2019b) Gong et al. (2021)
			Attribute itself	Buolamwini & Gebru (2018) Das et al. (2018) Amini et al. (2019)
			Other targets associated with protected attribute	Hashimoto et al. (2018) Adeli et al. (2021) Cheng et al. (2021)
Type II Bias	Dependence between model prediction and attribute	Is attribute known and labeled?	Known and labeled	Zhu et al. (2021) Ragonesi et al. (2021) Kim et al. (2019a)
			Known but unlabeled	Wang et al. (2019a) Bahng et al. (2020) Cadene et al. (2019)
			Unknown	Zhao et al. (2023a) Nam et al. (2020) Jeon et al. (2022)
		Type of attribute	Sensitive attribute/protected attribute	Angwin et al. (2022b) Creager et al. (2019) Madras et al. (2018)
			Spurious attribute	Sagawa* et al. (2020) Tartaglione et al. (2021) Hong & Yang (2021)

4.3 Fairness Criteria

Besides the papers that explore bias issues directly from the perspective of bias itself, there is another group of papers that leverage established fairness criteria (*e.g.*, demographic parity and equalized odds) as their debiasing objectives. In this section, we first adopt the corresponding definitions of fairness from the definition of bias in Definitions 1 and 2, and then demonstrate that relevant papers based on established fairness criteria can be categorized under these definitions. Given that fairness is the opposite of bias, we can derive the fairness definition for each type of bias as follows,

Definition 3. *Fairness w.r.t. Type I Bias. A model f is fair w.r.t.* Type I Bias *if f yields even performance $d(\hat{Y}, Y)$ across attribute A,* i.e.*,*

$$\sup_{a,a' \in \mathcal{A}, d \in \mathcal{M}} |d(\hat{Y}, Y|A = a) - d(\hat{Y}, Y|A = a')| = 0 \tag{3}$$

where a, a' are possible values of A (*e.g.*, female and male), and $\mathcal{M}$ is the set of all potential performance metrics.

Definition 4. *Fairness w.r.t. Type II Bias. A model f is fair w.r.t.* Type II Bias *if model prediction $\hat{Y}$ is independent with attribute A,* i.e.*,*

$$\sup_{a,a' \in \mathcal{A}} |P(\hat{Y}|A = a) - P(\hat{Y}|A = a')| = 0 \tag{4}$$

where a, a' are possible values of A (*e.g.*, female and male).

Fairness criteria can be categorized into two key classes: group fairness and individual fairness (Mehrabi et al., 2021a; Wang et al., 2022c; Castelnovo et al., 2022). Specifically, group fairness is founded on the idea that "groups of people may face biases and unfair decisions", whereas individual fairness is grounded in the principle that "similar individuals should receive similar decisions" (Castelnovo et al., 2022). We mainly unify group fairness into our definitions since group fairness is more commonly used in fairness research (Du et al., 2020). Group fairness encompasses several well-known fairness criteria such as demographic parity/statistical parity (Dwork et al., 2012; Kusner et al., 2017), equalized odds/equality of odds (Hardt et al., 2016), equal opportunity/equality of opportunity (Hardt et al., 2016), and accuracy parity (Quan et al., 2023). The categorization of them under our fairness definitions is shown in Tab. 3. Specifically, demographic parity, which requires $P(\hat{Y}|A = a_0) = P(\hat{Y}|A = a_1)$, is consistent with Definition 4 when attribute A is binary. Equalized odds, which requires that both even true positive rate (TPR) ($P(\hat{Y} = y_1|Y = y_1)$) and even false positive rate (FPR) ($P(\hat{Y} = y_1|Y = y_0)$) across A, and equal opportunity, which is the weaker notion of equalized odds that focuses solely on the advantaged outcome where $Y = y_1$, align with Definition 3 since TPR and FPR are included in the set of performance metrics $\mathcal{M}$. Accuracy parity, where accuracy is represented by $P(\hat{Y} = Y)$, also aligns with Definition 3 since accuracy is the element of $\mathcal{M}$.

4.4 Summary

Having unified the prevalent bias issues and well-known fairness criteria under our definitions, in this section, we summarize the main advantages of the proposed definitions. First, the proposed definitions focus on the manifestation of predominant bias, which is clearer and easier to apply compared to definitions based on causes, since the causes of these biases are debatable in some cases (Adeli et al., 2021; Stone et al., 2022; Wang & Deng, 2020). Second, the proposed definitions yield the general form, and by specifying the components in the general form, they can be used to unify a comprehensive list of papers, as summarized in Tab. 4. Third, the proposed definitions, as the first definition to formally define dominant biases, bridge the gap between numerous fairness definitions (Hardt et al., 2016; Kusner et al., 2017; Dwork et al., 2012; Chen et al., 2019; Grgic-Hlaca et al., 2016; Lechner et al., 2021; Quan et al., 2023) and the significant shortage of formal bias definitions. Furthermore, compared with fairness definitions, bias definitions are more practical since encountering bias issues is more common in real-world scenarios, whereas achieving fairness, often considered an ideal benchmark, is rare in practice. Fourth, given that the proposed bias definitions are relatively general, the corresponding fairness definitions are strict, hence aligning with the need for fairness as an ideal standard. Additionally, several well-known fairness criteria can be unified under the proposed fairness definitions.

5 Analyzing Confusion between Type I Bias and Type II Bias

In this section, we identify and analyze the widespread confusion in the literature surrounding two predominant types of bias in neural networks: Type I Bias (uneven performance across demographic groups) and Type II Bias (prediction dependence on protected attributes). First, we highlight five major manifestations of confusion. We then explore three underlying causes. Together, these factors explain why the field continues to conflate fundamentally different bias types, hindering progress toward clear and effective fairness solutions.

5.1 Manifestations of Confusion

In the previous section, we categorize 415 papers that discuss prevalent biases into two groups based on the manifestation of bias they address. The criteria for this categorization are clearly outlined in Tab. 4. Furthermore, the distinctions between these two types of biases are illustrated in Definitions 1 and 2. However, as summarized in Tab. 5, there is substantial confusion between them in existing literature, which poses challenges for researchers to investigate bias issues. Thus, it is crucial to clarify the confusion and underscore the distinctions between these two types of biases. To this end, in this section, we primarily highlight several prevailing confusions and the potential consequences that arise from overlooking them, based on the investigation of 415 papers.

Table 5: The summary of the existing confusion in the literature regarding bias issues.

Type of confusion	Examples
Ambiguity of Terminology	Wang et al. (2020b); Zhao et al. (2017); Amini et al. (2019)
Inaccurate Motivation	Ragonesi et al. (2021); Alvi et al. (2018); Salvador et al. (2022)
Lack of Terminology Reuse	Stone et al. (2022); Adeli et al. (2021); Wang & Russakovsky (2023)
Abuse of Bias Assessment Metrics	Zhang et al. (2023); Lokhande et al. (2020); Wang et al. (2019c)
Weak Existing Distinction	Wang et al. (2022c); Mehrabi et al. (2021a); Le Quy et al. (2022)

5.1.1 Ambiguity of Terminology

One of the confusions is the ambiguity surrounding the terminology of bias. This ambiguity manifests in three primary ways. First, several papers adopt vague terminology such as “bias issues" or simply “bias" without clarifying the particular type of bias they address (Wang et al., 2020b). Furthermore, other commonly used terms such as “model bias" or “algorithmic bias" are also ambiguous, as they might represent either the bias that manifests in the model or the bias that originates from the model itself. Second, studies often denote bias from varied aspects (Hirota et al., 2022; Markl, 2022). For instance, some papers refer to “demographic bias", “gender bias", or “racial bias", emphasizing bias from the perspective of demographic statistics. In contrast, other works utilize “dataset bias", “model bias", or “algorithmic bias", indicating the source of bias. Third, the existing literature frequently uses the same terms to describe different kinds of biases (Liu et al., 2022a; Ragonesi et al., 2021), as summarized in Tab. 6.

Consequences. The ambiguity of terminology undermines the clarity of the intended statement and may further lead to misdirected debiasing techniques. For instance, in the abstract of the paper (Zhao et al., 2017), the authors claim that:

- *“We find that (a) datasets for these tasks contain significant gender bias and (b) models trained on these datasets further amplify existing bias."* (Zhao et al., 2017)

In this case, the lack of clarity around the term “gender bias" weakens the significance of the findings. Furthermore, the scope of this ambiguity is extensive. Specifically, sections including “Title", “Abstract", “Introduction", and “Related Work" are often impacted, as there may lack sufficient context for a precise interpretation (Sadeghi et al., 2019; Gordaliza et al., 2019). More concerned, the vagueness may persist throughout the entire paper (Amini et al., 2019) if the addressed bias is not disambiguously clarified in the “Problem Statement" or evaluation protocol in the “Experiments" section.

Table 6: The summary of terms commonly used for bias.

Paper	Claimed bias to address (Motivation)	Actual type of bias to address (Technique)	
		Type I Bias	Type II Bias
Wang et al. (2019b)	Racial bias	✓	
Dhar et al. (2021)	Gender bias, skintone bias	✓	✓
Conti et al. (2022)	Gender bias	✓	✓
Wang et al. (2019c)	Gender bias		✓
Zhao et al. (2017)	Gender bias		✓
Wang et al. (2020b)	Gender bias		✓
Amini et al. (2019)	Algorithmic bias	✓	
Liu et al. (2022a)	Dataset bias	✓	
Adeli et al. (2021)	Dataset bias	✓	
Ragonesi et al. (2021)	Dataset bias		✓
Lee et al. (2021)	Dataset bias		✓

5.1.2 Inaccurate Motivation

Another confusion is that existing work addressing these two types of bias inaccurately cites each other for their own motivation. For instance, some studies (Ragonesi et al., 2021; Alvi et al., 2018) that address Type II Bias motivate themselves by the uneven performance in face recognition, a manifestation of Type I Bias. Other work (Wang & Deng, 2021; Salvador et al., 2022) that tackles Type I Bias in debiasing face recognition is motivated by the correlation between model predictions and spurious attributes in facial attribute classification (Alvi et al., 2018), a manifestation of Type II Bias. Furthermore, this confusion is aggravated as some papers are motivated by semi-relevant work. Specifically, as highlighted by (Grother et al., 2019), debiasing face recognition literature (Salvador et al., 2022; Wang et al., 2019b; Wang & Deng, 2020) tends to be motivated by the manifestation of worse accuracy for minority groups in sex classification (Buolamwini & Gebru, 2018), rather than the direct issue of uneven performance in face recognition (Robinson et al., 2020; Pahl et al., 2022).

Consequences. Inaccurate motivation leads to misunderstanding and misalignment in the existing literature. Furthermore, this issue may compound over time, as the subsequent work built upon the papers with such inaccurate motivation will perpetuate the confusion.

5.1.3 Lack of Terminology Reuse

The confusion also manifests in the introduction of new terms in different papers addressing the same bias. For instance, "minority group bias" (Stone et al., 2022), "dataset bias" (Adeli et al., 2021), and "bias as underrepresentation" (Wang & Russakovsky, 2023) are all used to denote uneven performance across attributes (Type I Bias).

- *"Dataset bias is often introduced due to the lack of enough data points spanning the whole spectrum of variations with respect to one or a set of protected variables."* (Adeli et al., 2021)
- *"Minority group bias. When a subgroup of the data has a particular attribute or combination of attributes that are relatively uncommon compared to the rest of the dataset, they form a minority group. A model is less likely to correctly predict for samples from a minority group than for those of the majority."* (Stone et al., 2022)
- *"[...] 'bias' means that one appearance of an object is underrepresented."* (Wang & Russakovsky, 2023)

Similarly, "sensitive attribute bias" (Stone et al., 2022), "task bias" (Adeli et al., 2021), and "bias as spurious correlation" (Wang & Russakovsky, 2023) all signify the dependence between model prediction and attribute (Type II Bias).

- *"Task bias, on the other hand, is introduced by the intrinsic dependency between protected variables and the task."* (Adeli et al., 2021)
- *"Sensitive attribute bias. A sensitive attribute (also referred to as "protected") is one that should not be used by the model to perform the target task, but which provides an unwanted "shortcut" that is easily learned, and results in an unfair model."* (Stone et al., 2022)
- *"[...] considering bias in the form of spurious correlations between the target label and a sensitive attribute which is predictive on the training set but not necessarily so on the test set."* (Wang & Russakovsky, 2023)

Consequences. These inconsistent definitions can further contribute to confusion, with some highlighting the manifestation of the bias while others delving into the underlying causes of the bias. Furthermore, without a unified terminology for the predominant biases, it becomes challenging to systematically gather and compare relevant work.

5.1.4 Abuse of Bias Assessment Metrics

The usage of bias assessment metrics exhibits the confusion in two primary ways. First, the bias assessment metrics, which are designed independently of debiasing methods, are rarely used (Li & Abd-Almageed, 2021; Wang & Russakovsky, 2021). Instead, many works tend to introduce their own metrics to demonstrate the effectiveness of the proposed debiasing method (Wang et al., 2019c; Zhao et al., 2017), which leads to an overwhelming number of metrics. Second, some studies inappropriately employ indirect bias assessment metrics or even metrics that are not designed for the specific bias they address. For instance, several studies (Zhang et al., 2023; Lokhande et al., 2020) motivated by the dependence between model prediction and attributes (the manifestation of Type II Bias) use true positive rate (TPR) difference and false positive rate (FPR) difference for evaluation. However, as highlighted by (Wang & Russakovsky, 2021), metrics such as TPR difference, FPR difference, accuracy difference, and average mean-per-class accuracy difference, are not suitable for evaluating Type II Bias since they fail to consider the dependence between target and attribute in the training set and cannot distinguish between an increase or decrease of dependence in learned representation.

Consequences. The abuse of bias assessment metrics leads to inaccurate evaluations of debiasing performance in relation to the specific type of bias being addressed, hence exacerbating confusion in the field. Furthermore, it also complicates the comparison between different debiasing methods and hinders the construction of a unified evaluation protocol.

5.1.5 Weak Existing Distinction

Despite the evident confusion in the literature, numerous studies, especially survey papers, have not sufficiently distinguished Type I Bias and Type II Bias. Furthermore, the confusion is not only widespread but has also persisted for a significant duration, as shown by the timeframes of the investigated papers. However, the bias taxonomy, presented in surveys over time (Wang et al., 2022c; Mehrabi et al., 2021a; Le Quy et al., 2022), may fail to clearly differentiate between these two types of biases. Alarmingly, a recent and highly cited survey on machine learning bias (Mehrabi et al., 2021a) scarcely cites papers that discuss Type II Bias stemming from spurious correlations between target and attribute, thereby overlooking the distinction from Type I Bias.

Consequences. The weak distinction between these two types of biases in existing surveys will exacerbate the prevailing confusion in this field over time. Consequently, due to the lack of clarity, which surveys were originally designed to provide concerning the categorization of bias issues, these bias issues will eventually be undesirably conflated.

5.2 Underlying Causes of Confusion

In this section, we investigate various factors that may contribute to the confusion discussed in the previous section. Specifically, we examine the historical context, the preconception about bias, and the methodologies adopted to address different biases, to provide insights on how and why such confusion has persisted in the literature.

5.2.1 Historical Context

We first examine the historical origins of bias issues. In Fig. 2, we summarize the enrichment of the concept "bias" in machine learning from the perspective of Type I Bias and Type II Bias and highlight key milestones throughout its history. Originally, "bias" is defined as unfair favoritism or prejudice towards one thing, person, or group over another (DiTomaso, 2015). Specifically, bias issues are especially evident in real-world decision-making processes, such as advertising, financial creditworthiness, employment, education, and criminal justice (Ruggeri et al., 2023; Edmond & Martire, 2019). To promote fairness, certain sensitive attributes (*e.g.*, sex, age, and race) are by law defined as protected attributes that cannot be discriminated against in the decision-making process (Corbett-Davies & Goel, 2018). In this initial stage, decisions are primarily made by humans. Thus, the main bias issue is whether human decision-making depends on protected attributes, which aligns with Type II Bias in our definitions.

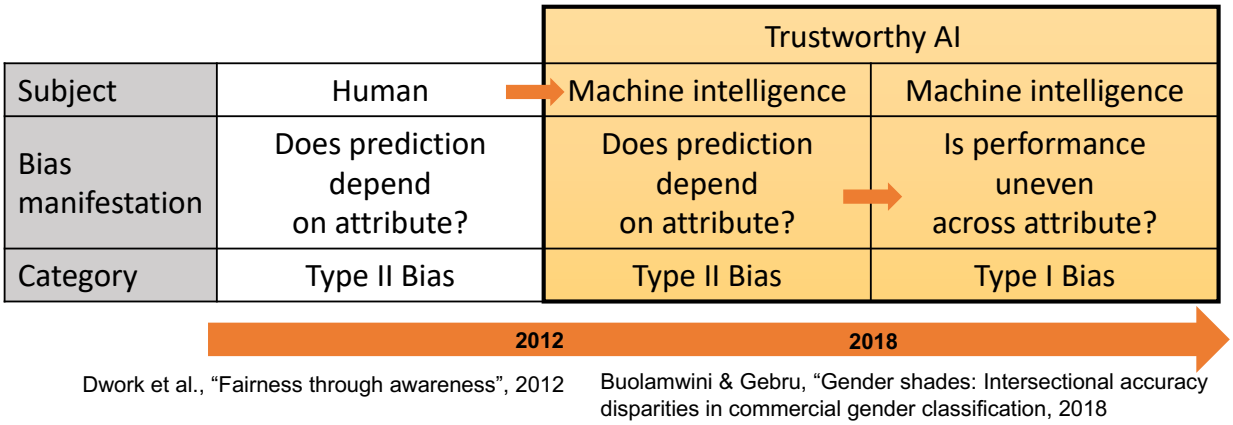


Figure 2: The enrichment of the concept "bias" in machine intelligence with important milestones. Initially, "bias" implied that human decision-making depends on protected attributes (Type II Bias). As machine intelligence began aiding human decision-making processes, the subject of "bias" broadened from humans to algorithms. Along with the continued advances of machine intelligence, a new aspect of bias issues, performance disparity across demographic groups (Type I Bias), further enriched the meaning of "bias". Currently, addressing both Type I Bias and Type II Bias is essential for ensuring Trustworthy AI.

Following the emergence of neural networks, machine learning models start to assist in human decision-making processes (Bastani et al., 2021; Dankwa-Mullan et al., 2019). This evolution also leads to an expansion of the subject in the discussion regarding bias issues, from human decision-making to algorithmic decision-making (Starke et al., 2022). With this change, numerous works begin to explore if algorithmic decision-making depends on protected attributes (*i.e.*, demographic parity) (Dwork et al., 2012; Kusner et al., 2017), which also aligns with Type II Bias. Meanwhile, along with the advancement of neural networks, their performance becomes a crucial evaluation criterion. Consequently, it brings significant attention to a new aspect of bias issues: performance disparity across demographic groups (Buolamwini & Gebru, 2018; Quan et al., 2023), which aligns with Type I Bias in our definitions. Furthermore, new fairness criteria such as equalized odds and equal opportunity (Hardt et al., 2016), which address disparities in true positive rates and false positive rates across demographic groups, are adopted from demographic parity.

We conjecture that the confusion arises because the term "bias" in neural networks has been endowed with multiple important meanings over time without well-defined distinctions. This ambiguity leads individuals to interpret different types of predominant biases from the same term. Specifically, some individuals associate the primary bias with performance disparity due to the critical role of model performance in model evaluation. Conversely, other individuals prioritize prediction disparity since it is the prevalent bias deeply embedded in real-world scenarios. Consequently, denoting these two different but predominant biases with the single term "bias" results in misunderstandings in the broader literature.

5.2.2 Preconception about Bias

The preconception of researchers about bias, stemming from their specific relevant fields, also contributes to the confusion. Specifically, bias issues encompass a wide range of relevant fields, some of which are associated with Type I Bias and others with Type II Bias. For instance, Type I Bias involves long-tail distribution (Cao et al., 2020), catastrophic forgetting (Kirkpatrick et al., 2017), domain adaptation (Li et al., 2014), and various biometric tasks (Xiao et al., 2023; Hutiri & Ding, 2022). In contrast, Type II Bias involves shortcut learning (Geirhos et al., 2020), simplicity bias (Teney et al., 2022), invariant representation learning (Creager et al., 2019), out-of-distribution challenges (Shen et al., 2021). In this sense, researchers from diverse fields hold their own preconceived notions of bias based on their field-specific knowledge. For instance, in several biometric tasks (*e.g.*, face recognition, face detection, face verification) With identity as target and sex as an attribute, uneven performance across sex (the manifestation of Type I Bias) is naturally regarded as bias since the primary focus of biometric systems is on model performance (Robinson et al., 2020).

However, the dependence between model prediction and attribute (the manifestation of Type II Bias) might not be considered as bias since there naturally only exists non-overlapping targets across attribute (Wang et al., 2019b). For instance, an individual can be categorized as either male or female but not both, thereby resulting in a natural association between identity prediction and specific sex. Furthermore, due to the absence of clear distinctions regarding bias issues, research groups from different fields may not share a unified perspective on bias and may interpret it differently. However, they use similar bias-related terms in their papers and present them in the same venues, which potentially causes confusion regarding bias issues.

5.2.3 Similar Methodologies

The existing confusion also arises from the overlap in methodologies used to address Type I Bias and Type II Bias. For instance, to mitigate Type I Bias, several studies (Morales et al., 2020; Gong et al., 2020; Dhar et al., 2021) enhance the performance for minority groups by preventing the model from encoding the information of protected attributes. Similarly, to tackle Type II Bias, some methods (Ragonesi et al., 2021; Zhu et al., 2021; Kim et al., 2019a) aim to develop representations that are invariant to the protected attribute by minimizing mutual information between the learned representation and the protected attribute. Both of these methods can be categorized into invariant representation learning (Arjovsky et al., 2019). Furthermore, domain adaptation is also utilized for both Type I Bias (Kan et al., 2015; Guo et al., 2020) and Type II Bias (Rosenfeld et al., 2022). These similarities in methodologies obscure the distinction between Type I Bias and Type II Bias, thereby inducing confusion.

6 Experimental Discussion

In this section, we empirically investigate the distinction between Type I Bias and Type II Bias. Specifically, we conduct experiments on two synthetic datasets and two well-known real-world datasets: Adult Income Dataset (Dua & Graff, 2017) and CelebA Dataset (Liu et al., 2015). First, we use synthetic data to demonstrate that Type I Bias and Type II Bias are unrelated, *i.e.*, one can exist without the presence of the other bias. Next, we utilize the Adult dataset to further illustrate the difference between Type I Bias and Type II Bias in real-world scenarios. Last, we employ the CelebA dataset to evaluate the effectiveness of multiple representative bias assessment metrics in assessing Type I Bias and Type II Bias. All experimental results are obtained by averaging the results over 10 trials. While our experimental validation focuses on image and tabular data, the proposed taxonomy and definitions extend naturally to language tasks. For example, Type I Bias in text classification can manifest as uneven performance across demographic subgroups, such as sentiment classifiers that perform better on reviews written in Standard American English than on African American Vernacular English, even when ground truth labels are consistent (Tatman, 2017). Besides, Type II Bias in text tasks often arises when model predictions (*e.g.*, toxicity, hate speech, or occupation classification) are influenced by identity-indicative features such as names, pronouns, or phrases associated with a protected group—even when these features are irrelevant to the task (Gallegos et al., 2024).

6.1 Unrelated Occurrence

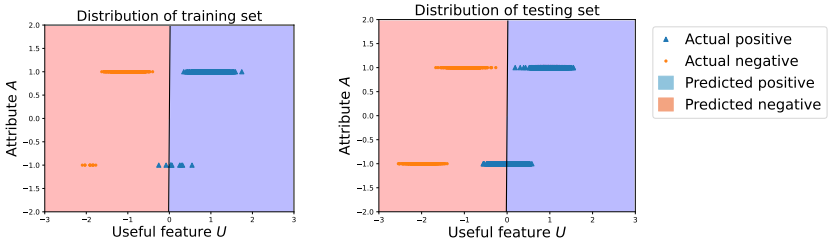
In this section, we leverage synthetic data to simulate two scenarios: the first scenario showcases the presence of Type I Bias without Type II Bias, while the second scenario showcases the presence of Type II Bias without Type I Bias.

Setup. We construct the synthetic dataset containing instances (x, y), where x denotes a two-dimensional input consisting of the useful feature u and the binary attribute a, and y denotes the target label. Next, we apply a classifier $C : \mathcal{X} \to \mathcal{Y}$ to consume the input x and produce the prediction $\hat{y} = C(x) = C(u, a) \in \mathcal{Y}$. The classifier is a single fully connected layer (FC) followed by the binary cross-entropy loss. To evaluate Type I Bias, we measure the difference in accuracy. To assess Type II Bias, we utilize the Calders-Verwer discrimination score (Calders & Verwer, 2010) defined as $|P(\hat{Y} = y|A = 1) - P(\hat{Y} = y|A = -1)|$.

6.1.1 Type I Bias Exists without Type II Bias

To induce Type I Bias, we construct a training set which is imbalanced across attribute A, where the subset with $A = -1$ represents the minority group, as shown in Fig. 3. Additionally, we vary the optimal classification boundary across A since one widely accepted cause of Type I Bias is that the model trained on the sufficient samples in majority groups might not effectively generalize to minority groups (Wang & Russakovsky, 2023). And, we construct a testing set which is balanced across values of the attribute A. Details of the dataset construction are presented in Appendix 2.2.

Analysis. In Fig. 3, we observe that the learned classification boundary is vertical at $X = 0$, which is primarily determined by dominant samples in the majority group. The vertical boundary suggests that the model does not use attribute A for classification. Furthermore, as highlighted in Tab. 7, given that $P(\hat{Y} = y|A = 1) = P(\hat{Y} = y|A = -1)\ \forall\ y \in \{0, 1\}$, model prediction $\hat{Y}$ is independent with attribute A, *i.e.*, Type II Bias does not exist. However, it is noteworthy that there is a significant performance disparity between the majority and minority groups, which confirms the existence of Type I Bias.



(a) Training set. (b) Testing set.

Figure 3: Distribution of training and testing sets regarding synthetic data. The vertical classification boundary (labeled as the black line) reveals that the classifier does not utilize A for classification. However, there are more wrong predictions in the group of $A = -1$ than in the group of $A = 1$, which violates performance parity.

Table 7: Type I Bias exists without Type II Bias since there exists accuracy disparity across A while $\hat{Y}$ and A are independent.

	Accuracy	$P(\hat{Y} = 0\|A)$	$P(\hat{Y} = 1\|A)$
$A = 1$	100.00	66.7%	33.3%
$A = -1$	65.33	66.7%	33.3%
$\|\Delta\|$	34.67	0	0

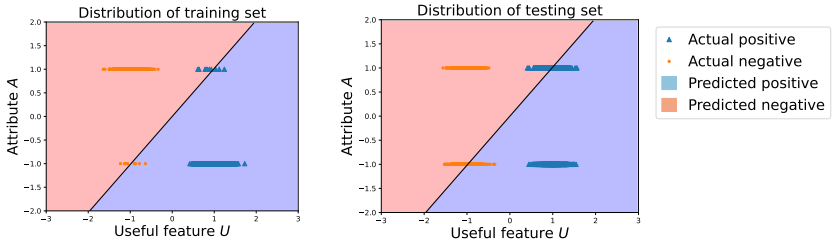
6.1.2 Type II Bias Exists without Type I Bias

To induce Type II Bias, we construct a training set in which the combinations $(A = 1, Y = 0)$ and $(A = -1, Y = 1)$ occur more frequently than other combinations, as illustrated in Fig. 4. This setup is motivated by the widely accepted understanding that a spurious association between the target label Y and the attribute A in the training data is a key contributor to Type II Bias (Nam et al., 2020; Zhu et al., 2021; Tartaglione et al., 2021). For evaluation, we construct a testing set that is balanced across both Y and A, ensuring no association between the two variables. Details of the dataset construction are presented in Appendix 2.2.

Analysis. In Fig. 4, we observe that the learned classification boundary is not vertical, which suggests that the classifier relies on A for decision-making. Furthermore, as highlighted in Tab. 8, given that $P(\hat{Y} = y|A = 1) \neq P(\hat{Y} = y|A = -1)\ \forall\ y \in \{0, 1\}$, model prediction $\hat{Y}$ is not independent with attribute A, *i.e.*, Type II Bias exists. However, for Type I Bias, it is noteworthy that there is no significant performance disparity between the majority and minority groups.

6.2 Different Manifestations in the Real World

In this section, we utilize the Adult Income Dataset (Dua & Graff, 2017) to illustrate different manifestations of Type I Bias and Type II Bias in real-world scenarios. Adult Dataset is a census dataset where the target is whether a person earns a higher income (over 50K USD per year), and the protected attribute is sex. As shown in Tab. 9, the dataset is partitioned into four quarters based on the combination of target labels and protected attribute labels, given that both are binary in nature. The statistics illustrate that the Adult dataset is well-suited for investigating both Type I Bias and Type II Bias. Specifically, the dataset exhibits an uneven distribution across sex, with a larger number of female individuals (16,192) compared to male individuals (32,650), which could induce Type I Bias. Furthermore, the dataset also exhibits a substantial



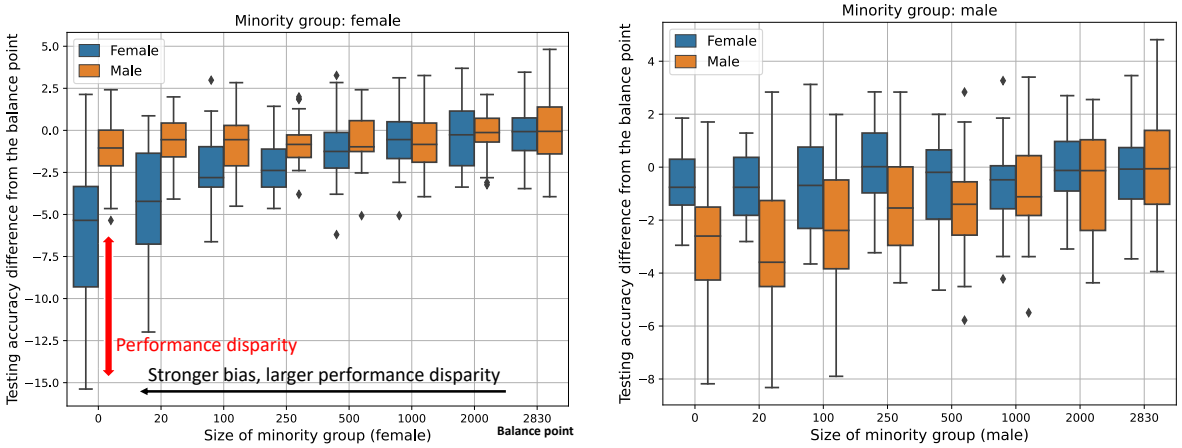
(a) Training set. (b) Testing set.

Figure 4: Distribution of training and testing sets regarding synthetic data. The non-vertical classification boundary (labeled as the black line) reveals that the classifier utilizes A for classification. However, the number of wrong predictions is approximately the same across A, thereby fulfilling performance parity.

Table 8: Type II Bias exists since $\hat{Y}$ and A are not independent while there is no accuracy disparity across A.

	Accuracy	$P(\hat{Y}=0\|A)$	$P(\hat{Y}=1\|A)$
$A=1$	85.98	64.1%	35.9%
$A=-1$	85.97	35.4%	64.6%
$\|\Delta\|$	≈ 0	28.7%	28.7%

disparity in the number of samples with higher income between females (1,769) and males (9,918), which could induce Type II Bias. Besides, we consider the magnitude of disparity as a reflection of bias strength. In practice, we acknowledge that the significance of a given disparity is inherently context-dependent. So, it is advisable in practical deployments to apply domain-specific thresholds when interpreting bias metrics.



(a) Female is the minority group. (b) Male is the minority group.

Figure 5: Illustration of Type I Bias on Adult, which manifests as uneven performance between the minority and majority groups. As Type I Bias becomes stronger (the minority size decreases), the accuracy for the minority group diminishes while the accuracy for the majority group remains unchanged, thereby enlarging the performance disparity across the minority and majority groups.

Setup. We perform data pre-processing on the input census data. Specifically, we transform the categorical features using one-hot encoding and normalize the numerical features into a Gaussian distribution with zero mean and unit variance. Consequently, each input sample is transformed into a 108-dimensional vector. For the training model, we employ a three-layer multilayer perceptron (MLP) followed by the binary cross-entropy loss as the baseline classifier.

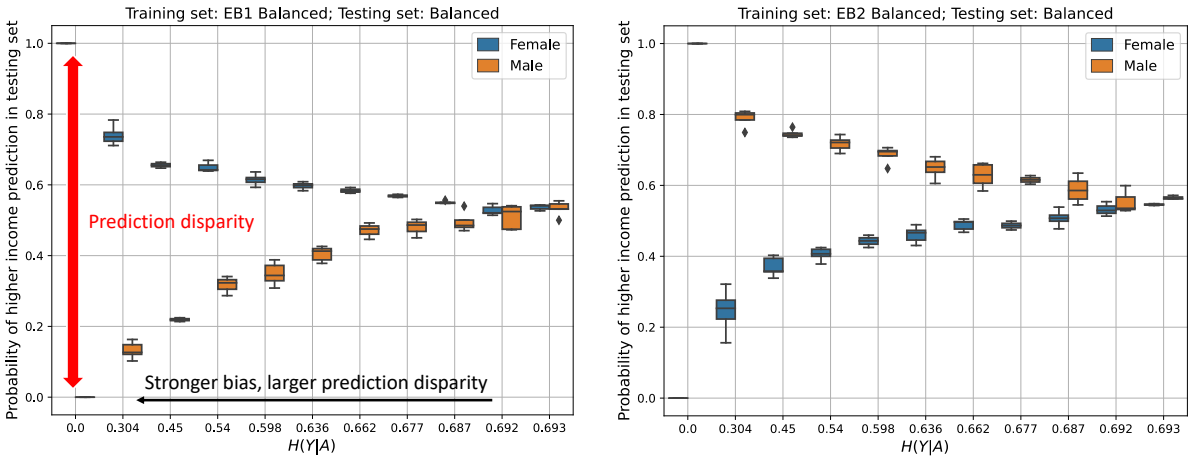
6.2.1 Type I Bias

To investigate Type I Bias, we construct several imbalanced training sets and control the bias strength by modifying the degree of imbalance in the training set. Specifically, we initially construct a balanced training

Table 9: Statistics of the Adult dataset. The number of females is greater than the number of males, which could induce Type I Bias. Furthermore, the number of samples with higher income and samples with lower income are different across sex categories, which could induce Type II Bias.

	Higher income	Lower income	Total
Female	1,769	14,423	16,192
Male	9,918	22,732	32,650
Total	11,687	37,155	48,842

set across both target Y and attribute A using 80% of the entire dataset and a balanced testing set with the remaining samples. We then manually adjust the size of the minority group in the training set while maintaining the size of the majority group to control bias strength. Additionally, we construct two distinct groups of training sets, with either females or males as the minority group. For instance, considering the setting where the female is a minority group and the minority size is 100, the training set would consist of 50 higher-income females and 50 lower-income females, in addition to all males from the balanced training set. We conduct experiments under different minority sizes and present the testing performance versus the size of the minority group in Fig. 5.



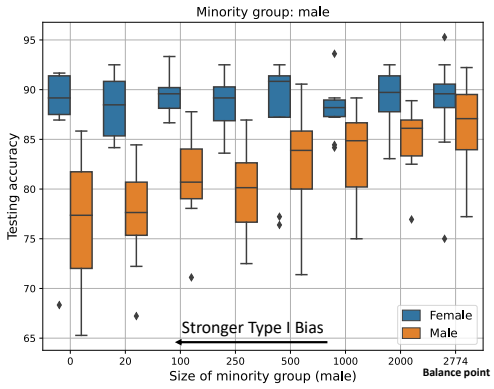
(a) Trained on EB1 Balanced consisting of females with higher income and males with lower income.

(b) Trained on EB2 Balanced consisting of females with lower income and males with higher income.

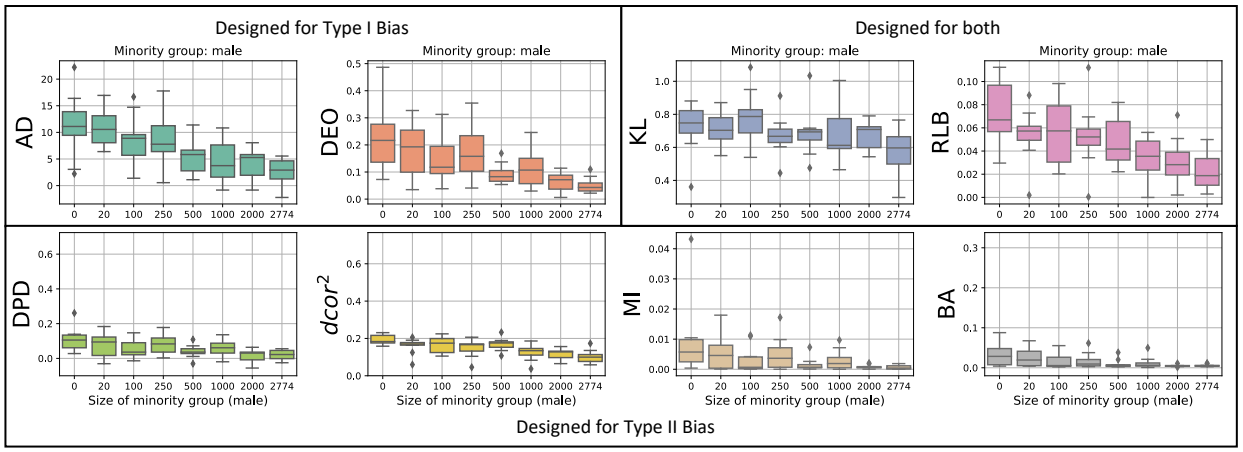
Figure 6: Illustration of Type II Bias on Adult, which manifests as the dependence between model prediction and attribute. As Type II Bias intensifies ($H(Y|A)$ decreases, rendering the attribute more predictable of the target), the prediction probability in outputting a specific prediction diverges between females and males, *i.e.*, decision-making increasingly relies on the attribute.

Analysis. Notably, we notice a non-zero accuracy disparity between females ($85.15\%_{\pm 1.52}$) and males ($78.38\%_{\pm 1.90}$) at the balance point where the training set is evenly distributed across both target Y and attribute A. We conjecture that this disparity is mainly because certain groups are inherently more difficult to classify than other groups (Klare et al., 2012). To facilitate a clearer analysis of Type I Bias, we use the accuracy difference from the testing accuracy at the balance point to represent the testing performance. This difference in testing accuracy, denoted as Acc_{diff}, is calculated by subtracting the testing accuracy at the balance point from the absolute accuracy at a given bias strength, *i.e.*, $Acc_{\text{diff}} = Acc_{\text{abs}} - Acc_{\text{balance}}$. In Fig. 5, we observe that the performance disparity exists across the minority group and the majority group. The accuracy for the minority group tends to decrease as its size diminishes (bias strength increases), especially when there are very limited samples from the minority group. Furthermore, in Fig. 5a, we observe that stronger bias results in larger performance fluctuations (bigger spread in the boxplot), which highlights the lack of robustness under such conditions. In summary, the manifestation of Type I Bias in

real-world scenarios is uneven performance across demographic groups. One plausible cause is the imbalance in data representation across these groups in the training set. For instance, some demographic groups may be underrepresented due to long-tail distribution (Cao et al., 2020), resulting in a skewed distribution of samples across different demographic groups. Consequently, while data-driven models are more accurately trained on demographic groups with sufficient samples, they may not be as effective for underrepresented groups, which leads to poor prediction accuracy and unfairness towards these groups.



(a) Testing accuracy.



(b) Evaluation with various bias assessment metrics.

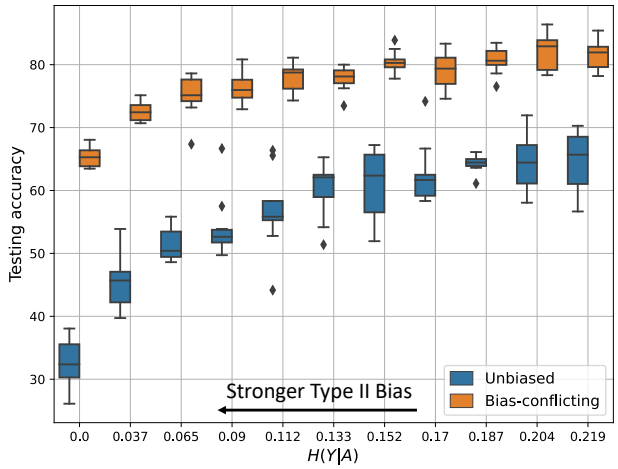
Figure 7: Investigation of Type I Bias on CelebA with males as the minority group. As bias strength diminishes (the size of the minority group enlarges), the accuracy of the minority group enhances, leading to a reduction in the accuracy disparity between females and males, and the bias assessed by metrics tailored to evaluate Type I Bias is also mitigated.

6.2.2 Type II Bias

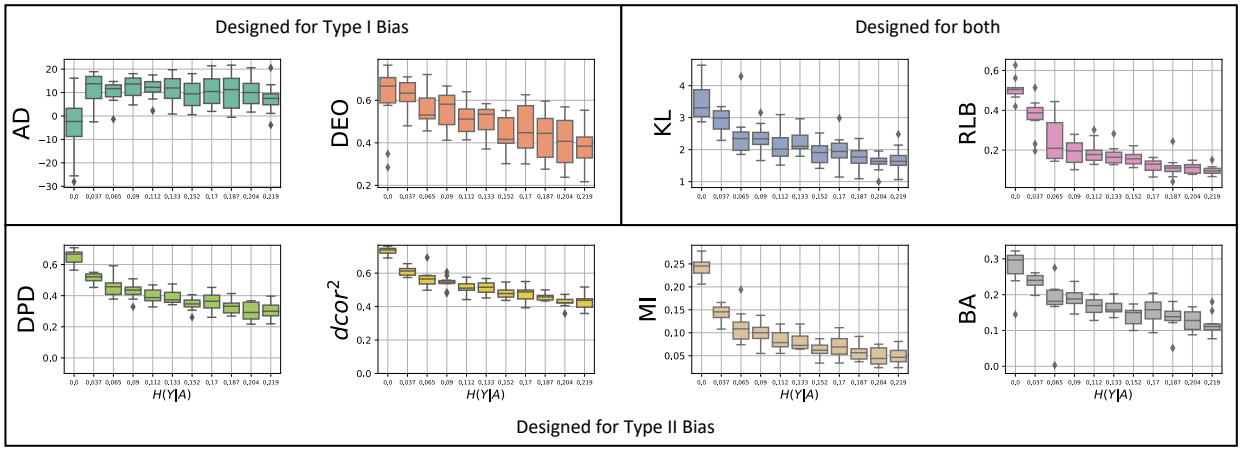
To investigate Type II Bias, we construct the training set where the target Y is associated with the attribute A and control the bias strength by adjusting the strength of the association between Y and A in the training set. Specifically, we initially construct two balanced training datasets consisting of 3538 records, each associating either females or males with higher income: (1) Extreme Bias 1 Balanced (EB1 Balanced) only contains females with higher income and males with lower income, and (2) Extreme Bias 2 Balanced (EB2 Balanced) only contains males with higher income and females with lower income. Subsequently, we adjust the percentage of bias-conflicting samples (samples with the opposite bias present in the training set) while ensuring a consistent number of biased samples. This strategy enables us to construct multiple training sets, each with a distinct conditional entropy $H(Y|A)$ (*i.e.*, the smaller $H(Y|A)$, the more predictive the attribute A is of the target Y, and the stronger the bias). Additionally, we construct a balanced testing

set (Balanced) consisting of 7076 records, ensuring an even distribution of all combinations of target and attribute labels. Note that all these datasets are designed to be balanced across attributes to mitigate the effect of Type I Bias.

Analysis. In Fig. 6, we observe that there is a significant prediction disparity between females and males. Furthermore, this disparity becomes more pronounced as $H(Y|A)$ diminishes (the bias strength increases). In summary, the manifestation of Type II Bias in real-world scenarios is the dependence on the attribute in decision-making processes. One widely accepted reason is an uneven distribution of *specific target groups* across attributes, distinguishing it from Type I Bias, which emerges from an uneven distribution of samples across attributes. For instance, the collected dataset may contain more negative samples for female individuals and positive samples for male individuals compared to other target-attribute combinations. During training, the model may leverage sex as a shortcut feature to simplify the learning process, rather than learning more comprehensive features. However, such an association between specific targets and attributes does not generally exist in the real world. Consequently, during applying, the trained model may still rely on the attribute, which leads to a higher frequency of positive outcomes for specific individuals and further unfair treatment for these groups.



(a) Testing accuracy.



(b) Evaluation with various bias assessment metrics.

Figure 8: Investigation of Type II Bias on CelebA. The evaluation of bias assessment metrics is conducted on *unbiased* testing set. As bias strength diminishes ($H(Y|A)$ increases, rendering the attribute less predictive of the target), the accuracies of both *unbiased* and *bias-conflicting* enhance, and the bias assessed by metrics tailored to evaluate Type II Bias is also mitigated.

6.2.3 Summary

As shown in Fig. 5, Type I Bias manifests as the performance disparity across A, which is evaluated based on the joint distribution of model prediction $\hat{Y}$ and ground truth Y. Conversely, as shown in Fig. 6, Type II Bias manifests as the prediction disparity across A, which is evaluated solely based on the distribution of model prediction $\hat{Y}$. Thus, Type I Bias and Type II Bias are unrelated phenomena and exhibit different impacts on the fairness of neural networks.

6.3 Evaluation of Various Metrics

In this section, we employ the CelebA dataset (Liu et al., 2015) to investigate several representative bias assessment metrics in assessing Type I Bias and Type II Bias. CelebA dataset is an image dataset of human faces where facial attributes (*e.g.*, blond hair) are the prediction target Y and sex is the attribute A. As illustrated in Tab. 10, the CelebA dataset is divided into four parts based on all possible combinations of the binary target and protected attribute labels. The statistics suggest the presence of both Type I and Type II biases in the CelebA dataset.

Table 10: Statistics of the CelebA dataset. The number of females is greater than the number of males, which could induce Type I Bias. Furthermore, the number of samples with blond hair and samples without blond hair are different across sex categories, which could induce Type II Bias.

	Blond hair	Non-blond hair	Total
Female	28,234	89,931	118,165
Male	1,749	82,685	84,434
Total	29,983	172,616	202,599

Setup. To construct training and testing sets, we follow the setup of Adult explained above. In the case of Type I Bias, we construct several training sets with varying bias strength by modifying the size of the minority group in the training set. For testing, we construct a testing set that is balanced across both target and attribute. In the case of Type II Bias, we construct training sets where facial attributes are associated with a particular sex. Specifically, we construct an extreme bias version of the training set consisting of 89754 images with $H(Y|A) = 0$, denoted *TrainEx*, where the bias-conflicting samples (samples exhibiting the opposite bias in the training set) are removed from the original training set. Furthermore, we control bias strength by adjusting the proportion of bias-conflicting samples while maintaining the number of biased samples (samples exhibiting the same bias observed in the training set). For testing, we construct two testing sets: (1) *Unbiased*, consisting of 720 images which contain an even number of samples across all combinations of target and attribute, and (2) *Bias-conflicting*, consisting of 360 images where all biased samples are excluded from *Unbiased* testing set (only bias-conflicting samples remain). In both studies, we consider *blond hair* as the prediction target. For the training model, we utilize ResNet18 (He et al., 2016) followed by the binary cross-entropy loss as the baseline classifier without any debiasing techniques. For bias assessment, we employ a comprehensive list of representative metrics including accuracy disparity (AP) (Quan et al., 2023), difference in equality of opportunity (DEO) (Morales et al., 2020), KL-divergence between score distributions (KL) (Chen & Wu, 2020), representation-level bias (RLB) (Li & Abd-Almageed, 2021), demographic parity distance (DPD) (Creager et al., 2019), distance correlation ($dcor^2$) (Székely et al., 2007), mutual information (MI) (Li & Abd-Almageed, 2023), and bias amplification (BA) (Zhao et al., 2017; Wang & Russakovsky, 2021).

Analysis. In the case of Type I Bias, as shown in Fig. 7a, there exists a noticeable performance disparity across sex. As the size of the minority group increases (bias strength diminishes), the performance of the minority group improves, and the performance gap between the minority and majority groups is mitigated. Notably, the performance gap is nonzero even at the balance point, with females achieving higher accuracy than males. We hypothesize that this is because blond hair is more visually prominent in females with long hair. Consequently, even if the dataset is balanced across sex, males may still be relatively underrepresented, *i.e.*, male images are still insufficient for the model to learn a robust representation of males. In the case of

Type II Bias, as shown in Fig. 8a, the testing accuracy of both *Unbiased* and *Bias-conflicting* testing sets rises as $H(Y|A)$ increases (bias strength diminishes).

For the evaluation of various bias assessment metrics, in Figs. 7b and 8b, we observe a noticeable decline in the metrics tailored for a specific type of bias as the corresponding bias strength diminishes. It is noteworthy that the mean of accuracy disparity (AD) approaches zero in the extreme bias case of Type II Bias, where $H(Y|A) = 0$ (the leftmost point). This can be attributed to the fact that, in such extreme bias situations, the target label is bijectively mapped to the attribute label in the training set. Consequently, the trained model may output arbitrary predictions for both sex in the testing set, which leads to an accuracy disparity that is nearly zero.

7 Path to Follow

In this section, we present a more comprehensive comparison between Type I Bias and Type II Bias based on our investigation of 415 papers. Our comparison encompasses multiple aspects, including the underlying causes, debiasing methods, evaluation protocol, prevalent datasets, and future directions. Most notably, for each type of bias, we summarize debiasing methods in Tab. 11, bias assessment metrics in Tab. 12, and prevalent datasets in Tabs. 13 and 14.

We use classification as a case study due to the abundance of existing work in this area. Nevertheless, the proposed definitions are generalizable to tasks beyond classification. For instance, in regression tasks, mean squared error can be used to assess Type I Bias, while standard measures of attribute dependence (*e.g.*, mutual information) can be applied to evaluate Type II Bias. We hope the comparison can alleviate the cognitive burden from the prevailing confusion between these two types of biases and serve as a roadmap for new researchers to follow.

Table 11: The summary of debiasing methods.

Category	Pre-processing	In-processing	Post-processing
Type I Bias	Balanced dataset collection Buolamwini & Gebru (2018); Karkkainen & Joo (2021) Synthetic dataset generation Balakrishnan et al. (2021); Li & Abd-Almageed (2023) Strategic sampling or reweighting Wang & Deng (2020)	Domain adaptation Wang et al. (2019b); Guo et al. (2020); Kan et al. (2015) Attribute removal Gong et al. (2020); Dhar et al. (2021)	Calibrated equalized odds Pleiss et al. (2017)
Type II Bias	Universal dataset collection Li et al. (2023a) Synthetic dataset generation Ramaswamy et al. (2021); Sattigeri et al. (2019) Domain randomization Tobin et al. (2017)	Mutual information minimization Kim et al. (2019a); Ragonesi et al. (2021); Zhu et al. (2021) Domain-invariant learning Sagawa* et al. (2020); Ahmed et al. (2021); Creager et al. (2021) Adversarial training Nam et al. (2020); Alvi et al. (2018); Zhang et al. (2018)	Ensemble domain-independent training Wang et al. (2020b)

Table 12: The summary of bias assessment metrics.

Category	Metrics
Type I Bias	Difference in performance evaluated by various criteria (*e.g.*, accuracy disparity (AD) Kim et al. (2019b); Quan et al. (2023); Zafar et al. (2017a); Zhao et al. (2019a)) Difference in equality of opportunity (DEO) Morales et al. (2020); Quadrianto et al. (2019); Sattigeri et al. (2019); Lokhande et al. (2020); Ramaswamy et al. (2021) Equal error rate (EER) Mirjalili et al. (2019)
Type II Bias	Demographic parity distance (DPD) Creager et al. (2019); Kim et al. (2019b); Sattigeri et al. (2019) Distance correlation ($dcor^2$) Székely et al. (2007); Adeli et al. (2021) Mutual information (MI) Li & Abd-Almageed (2023) Bias amplification (BA) Wang et al. (2020b); Ramaswamy et al. (2021), Directional BA Wang & Russakovsky (2021); Ramaswamy et al. (2021), Multi-attribute BA Zhao et al. (2023c) Disparity impact Zafar et al. (2017b); Bellamy et al. (2019) Representation bias Li et al. (2018); Li & Vasconcelos (2019) Logit-level loss Xie et al. (2017); Jaiswal et al. (2018)
Both	KL-divergence between score distributions (KL) Chen & Wu (2020); Ramaswamy et al. (2021) Representation-level bias (RLB) Li & Abd-Almageed (2021)

7.1 Type I Bias

7.1.1 Underlying Causes

Data imbalance across different demographic groups in the training set is commonly accepted as the possible cause for Type I Bias (Cherepanova et al., 2023; Röösli et al., 2022). Specifically, real-world data often exhibits the long-tail distribution where some demographic groups yield fewer samples than other groups (Cao et al., 2020). Consequently, given the data-driven nature of neural networks, models may be effectively trained in groups with sufficient samples but undertrained in groups only with limited samples, hence resulting in performance disparity across different groups and lower performance for minority groups. On the other hand, recent work suggests that Type I Bias can manifest even when the training set is balanced across

demographic groups (Wang & Deng, 2020). This challenges the conventional understanding of the causes of Type I Bias but promotes the discussion of other possible causes. For instance, Type I Bias may be induced by the underrepresentation of specific demographic groups (Wang & Russakovsky, 2023) or the intrinsic challenges associated with recognizing and classifying specific demographic groups (Klare et al., 2012).

7.1.2 Debiasing Methods

Addressing Type I Bias essentially involves optimizing the model to enhance its performance for minority groups while maintaining its performance for majority groups. The strategies can be broadly classified into three main categories based on the stage when the debiasing intervention is applied relative to the model training phase: pre-processing, in-processing, and post-processing. First, pre-processing methods intervene before the training phase. They are primarily designed based on the cause of Type I Bias (the imbalanced distribution across demographic groups in the training set). For instance, the straightforward approach is to construct a balanced real dataset for training (Karkkainen & Joo, 2021) or supplement minority groups with sufficient synthetic training samples (Li & Abd-Almageed, 2023). Another approach in this category involves strategically resampling to increase the occurrence of samples from minority groups or reweighting to assign higher importance to samples from underrepresented groups (Wang & Deng, 2020). Second, in-processing methods are integrated during the model training phase. Most notably, domain adaptation techniques (Kan et al., 2015; Guo et al., 2020) adapt well-learned representations from the majority group to the minority group, and attribute removal methods leverage adversarial learning (Gong et al., 2020; Dhar et al., 2021) to remove demographic information from learned representations. Lastly, post-processing methods apply debiasing techniques after the training process. One common technique is to calibrate the model predictions, ensuring that they adhere to specific fairness criteria (*e.g.*, equalized odds) (Pleiss et al., 2017).

7.1.3 Evaluation Protocol

The effectiveness of methods addressing Type I Bias is evaluated by performance disparity between majority and minority groups. In the case of binary attributes, the disparity is directly gauged by the performance difference between majority and minority groups (Buolamwini & Gebru, 2018; Ramaswamy et al., 2021; Xu et al., 2021c). In the case of non-binary attributes, the disparity is gauged by the standard deviation of performance across all demographic groups (STD) (Amini et al., 2019; Gong et al., 2020; 2021; Qin, 2020). To assess performance, there are a variety of metrics such as error rate (Buolamwini & Gebru, 2018; Sattigeri et al., 2019), loss (Hashimoto et al., 2018), accuracy (Kim et al., 2019b), average precision (AP) (Ramaswamy et al., 2021), positive predictive value (PPV), true positive rate (TPR) (Dhar et al., 2021; Adeli et al., 2021), false positive rate (FPR) (Xu et al., 2021c), average false rate (AFR), mean AFR (M AFR) (Ryu et al., 2017), confusion matrix (Gong et al., 2020), F1 score (Adeli et al., 2021), receiver operating characteristic curve (ROC) (Wang & Deng, 2020; Sattigeri et al., 2019; Mirjalili et al., 2018; Qin, 2020; Yu et al., 2020), area under the ROC (AUC) (Mirjalili et al., 2019; Gong et al., 2020; Adeli et al., 2021). Furthermore, besides these metrics to assess performance disparity, the performance improvement in minority groups compared to the baseline is provided for an intuition of debiasing effectiveness, along with overall performance to illustrate that it is not compromised.

7.1.4 Datasets

Datasets used to investigate Type I Bias mainly exhibit long-tail distributions. Most notably, several benchmark biometric datasets including LFW (Huang et al., 2007), IJB-A (Klare et al., 2015), IJB-C (Maze et al., 2018), and RFW (Wang et al., 2019b), are frequently utilized. A comprehensive list of datasets is presented in Tab. 13.

7.1.5 Future Directions

While many existing works address Type I Bias through balanced sampling or reweighting, more effort should be directed toward designing benchmarks that assess generalization across underrepresented demographic groups, especially in the long-tail regime. Another promising future direction is to delve into the root cause of Type I Bias since the formerly widely accepted cause (data imbalance) has been challenged by the experiment

Table 13: The well-known datasets used to study Type I Bias.

Name	# Subjects	# Samples	Sex (%)		Race (%)				
			Female	Male	European	Asian	Indian	African	Hispanic or Latino
CelebA Liu et al. (2015)	10K	202.5K	58.3	41.7	-	-	-	-	-
MUCT Milborrow et al. (2010)	0.2K	3.7K	50.9	49.1	-	-	-	-	-
RaFD Langner et al. (2010)	67	1.6K	37.3	62.7	-	-	-	-	-
PPB Buolamwini & Gebru (2018)	1.2K	1.2K	44.6	55.4	48.0	-	-	52.0	-
MORPH Ricanek & Tesafaye (2006)	13.6K	55.1K	15.3	84.7	19.2	0.28	-	77.2	3.2
LFW Huang et al. (2007)	5.7K	13K	22.3	77.6	69.9	13.2	2.9	14.0	-
CASIA-Webface Yi et al. (2014)	10K	0.5M	58.9	41.1	84.5	2.6	1.6	11.3	-
VGGFace2 Cao et al. (2018)	8.6K	3.1M	59.3	40.7	74.2	6.0	4.0	15.8	-
MS-Celeb-1M Guo et al. (2016)	90K	5.0M	-	-	76.3	6.6	2.6	14.5	-
IJB-A Maze et al. (2018)	0.5K	5.7K	-	-	66.0	9.8	7.2	17.0	-
IMDB-WIKI Rothe et al. (2018)	20K	500K	41.1	57.1	79.5	2.6	2.3	11.5	4.1
UTK Zhang et al. (2017)	-	20K	Balanced		45.3	14.7	18.4	21.6	-
RFW Wang et al. (2019b)	12K	40K	27.7	72.3	Balanced				-
FairFace Karkkainen & Joo (2021)	-	108K	Balanced		Balanced				

that Type I Bias exists even for a balanced dataset (Wang & Deng, 2020). Furthermore, exploring more effective debiasing methods to achieve even performance across cohorts is always of significant importance, hence, it is a valuable direction.

7.2 Type II Bias

7.2.1 Underlying Causes

The association between prediction targets and attributes in the training set is widely considered the possible cause of Type II Bias (Nam et al., 2020; Zhu et al., 2021; Zhao et al., 2023a). Different from Type I Bias, which originates from an uneven distribution of samples across attributes, Type II Bias arises from an uneven distribution of *specific target groups* across attributes. Specifically, the collected data may encompass a greater number of samples annotated with specific pairs of target labels and attribute labels (*e.g.*, (y^1, a^1) and (y^2, a^2)) than other combinations. Models trained on this dataset may leverage these attributes as shortcut features to simplify the training process rather than acquiring more comprehensive features. Consequently, when applying the trained models in real-world scenarios where the association does not generally exist, they may still rely on these attributes for decision-making and yield predictions that depend on these attributes, thereby resulting in a higher frequency of particular prediction outcomes for particular groups and further unfair treatment for these groups.

7.2.2 Debiasing Methods

Addressing Type II Bias essentially involves acquiring representations that are independent of the attribute while remaining informative for a wide range of downstream tasks (Balunovic et al., 2022). Similar to Type I Bias, the strategies can be classified into three categories: pre-processing, in-processing, and post-processing. First, pre-processing approaches can be further sub-categorized into dataset construction and data preprocessing. Dataset construction mainly encompasses collecting large-scale universal datasets to lessen the likelihood of spurious correlation between the target and the attribute (Li et al., 2023a;b), and generating counterfactual synthetic samples to augment the original biased training set, thereby reducing its inherent bias strength (Sauer & Geiger, 2021; Kim et al., 2021; Goel et al., 2021; Ramaswamy et al., 2021). Data preprocessing mainly encompasses fairness through unawareness (Dwork et al., 2012), which directly eliminates attributes from the input data, and domain randomization (Tobin et al., 2017) to utilize domain knowledge to assign a random value to the attribute label for each sample, thereby rendering it irrelevant to the target prediction. Second, in-processing approaches can be further divided into two subgroups: methods that either explicitly or implicitly minimize the mutual information (MI) between the learned latent features and the specific attribute. Specifically, several methods directly minimize mutual information between the latent representation for the target classification and the protected attributes to learn a representation that is predictive of the target but independent of the attributes (Kim et al., 2019a; Ragonesi et al., 2021; Zhu

et al., 2021). Another group of methods applies adversarial learning with surrogate losses (Nam et al., 2020; Alvi et al., 2018; Zhang et al., 2018) to implicitly reduce the mutual information or utilize domain-invariant learning (Ganin et al., 2016; Zhao et al., 2019b; Albuquerque et al., 2019; Sagawa* et al., 2020; Ahmed et al., 2021; Creager et al., 2021) to minimize classification performance gap across groups by mapping data to a space where distributions are indistinguishable while maintaining task-relevant information. Lastly, for the post-processing method, domain-independent learning (Wang et al., 2020b) learns an ensemble classifier comprising separate classifiers for each demographic group by sharing representations, thereby ensuring that the prediction from the unified model is not biased towards any domain.

7.2.3 Evaluation Protocol

The effectiveness of methods addressing Type II Bias is evaluated by prediction disparity across different groups. In the prevalent evaluation protocol, models are trained on a dataset where the target is associated with the attribute and tested on a held-out dataset where such association is absent (Kim et al., 2019a; Ragonesi et al., 2021; Zhu et al., 2021). Subsequently, the testing accuracy is reported to evaluate the model's capability to reduce the effect of association in the training set (the effectiveness to mitigate Type II Bias) (Wang et al., 2020b). Several studies also present the accuracy of worst-case groups, where the samples yield the opposite of the bias present in the training set (Sagawa* et al., 2020; Liu et al., 2021a; Lee et al., 2021). Furthermore, we summarize other commonly-used bias assessment metrics in Tab. 12. A noteworthy distinction in these bias assessment metrics for Type II Bias compared with Type I Bias is the absence of necessity for ground truth labels. This distinction is attributed to the fact that Type II Bias is defined as the dependence between model prediction and attribute, eliminating the need for ground truth, while evaluating Type I Bias necessitates ground truth to assess model performance.

7.2.4 Datasets

Most notably, several census datasets, including the Adult income dataset (Dua & Graff, 2017), German credit dataset (Dua & Graff, 2017), and COMPAS recidivism dataset (Angwin et al., 2022a), are employed as benchmark datasets to investigate the impact of sensitive/protected attributes in real-world decision-making processes. Additionally, computer vision and natural language processing communities also develop various datasets to investigate Type II Bias, *e.g.*, Colored MNIST (Kim et al., 2019a), CelebA (Liu et al., 2015; Nam et al., 2020), Waterbirds (Sagawa* et al., 2020), and CivilComments-WILDS (Borkan et al., 2019; Koh et al., 2021). A comprehensive list of datasets is summarized in Tab. 14.

Table 14: The well-known datasets used to study Type II Bias.

Name	Modality	Attribute	Target
Adult Dua & Graff (2017)	Tabular	Sex	Income
German Dua & Graff (2017)	Tabular	Sex, age	Credit
COMPAS Angwin et al. (2022a)	Tabular	Race	Recidivism
Colored MNIST Kim et al. (2019a)	Image	Color	Digit
CelebA Liu et al. (2015)	Image	Sex	Facial attributes
IMDB Rothe et al. (2015)	Image	Sex, age	Age, sex
Waterbirds Sagawa* et al. (2020)	Image	Background	Waterbirds or landbirds
CivilComment-WILDS Koh et al. (2021)	Text	Demographic identities	Toxic or non-toxic
WinoBias Zhao et al. (2018)	Text	Gender pronouns	Coreference resolution
Bias in Bios De-Arteaga et al. (2019a)	Text	Gender	Occupation
MS-COCO Lin et al. (2014)	Text & Image	Gender	Object

7.2.5 Future Directions

Type II Bias highlights the importance of learning representations that are causally disentangled from protected or spurious attributes. Future work should focus on developing provably invariant representations using tools from causality, domain invariance, and contrastive learning. Another promising research direction

is to explore the strong bias region (Li et al., 2023a) of Type II Bias, where the target and the attribute are strongly associated in the training set, a scenario that is overlooked by many existing works (Alvi et al., 2018; Kim et al., 2019a). Also, it is important to further explore more challenging scenarios where attribute labels are absent (Wang et al., 2019a; Bahng et al., 2020; Cadene et al., 2019) or unknown biases emerge (Li & Xu, 2021; Zhang et al., 2022b; Creager et al., 2021). Future research should study semi-supervised or unsupervised debiasing methods that can discover hidden biases and adapt accordingly, without relying on full attribute supervision. Moreover, a key challenge in Type II Bias is the lack of reliable diagnostic tools to detect when a model relies on protected attributes. New metrics or interpretability techniques are needed to identify attribute leakage, especially in high-dimensional representations.

7.3 Summary

In this section, we highlight the distinctions between Type I Bias and Type II Bias across multiple aspects and provide further explanations on the comparison in Tab. 1.

- Manifestation. A model exhibiting Type I Bias yields uneven performance across different groups and lower performance in minority groups, whereas a model exhibiting Type II Bias depends on attributes for decision-making and produces specific predictions that are highly associated with specific attributes.
- Disparity. Type I Bias refers to the disparity in prediction performance across attributes, whereas Type II Bias refers to the disparity in prediction outcomes across attributes.
- Causes. Type I Bias stems from insufficient training of underrepresented groups, whereas Type II Bias arises from the association between targets and attributes.
- Dataset inducing bias. An imbalanced distribution of samples across attributes induces Type I Bias, whereas an imbalanced distribution of *specific target groups* across attributes induces Type II Bias.

8 Suggestions

In this section, we first introduce a general framework for distinguishing between Type I and Type II Bias. We then offer practical guidance tailored to different audiences to help prevent confusion between the two types.

8.1 Guiding Framework for Categorizing Biases

To promote clarity in practical applications, we propose the following decision flow for identifying the bias type based on observable phenomena and context:

- **Performance disparities across demographic groups:** If lower accuracy or higher error rates are observed for certain groups (*e.g.*, minority groups), this may correspond to **Type I Bias**.
- **Systematic changes in model output with respect to protected attributes:** If the model's predictions change systematically as protected attributes vary, this may indicate **Type II Bias**.

When still in doubt, consider the used evaluation metric:

- *Does the evaluation require ground truth labels?*
  - Yes → Type I Bias (*e.g.*, fairness metrics such as True Positive Rate (TPR), False Positive Rate (FPR), or accuracy).
  - No → Type II Bias (*e.g.*, evaluating mutual information between the model prediction and a protected attribute).

8.2 Audience-Specific Guidance

In this section, we provide some suggestions for different audiences to help them identify and distinguish between Type I Bias and Type II Bias.

- Researchers:
  - Explicitly and precisely specify the type of bias being addressed to avoid vague terminology. Using well-defined terms such as Type I Bias and Type II Bias ensures clarity.
  - Derive motivation from prior work that addresses the same type of bias. This alignment helps reduce existing confusion in the literature.
  - Avoid introducing new terminology for previously studied biases. If a new term is necessary, clearly distinguish it from existing definitions. Reusing established terms fosters a more unified and comprehensible research community.
- ML Practitioners: When deploying models, evaluate Type I Bias through accuracy disparity and related performance metrics. Assess Type II Bias using measures of attribute dependence such as mutual information analysis.
- Educators: Illustrate the two bias types using concrete examples (*e.g.*, face recognition for Type I Bias and credit lending for Type II Bias). Use guiding questions such as:
  - "Is the model performing worse for Group A?" Yes $\rightarrow$ Type I Bias
  - "Would changing the attribute label (*e.g.*, sex) change the model's prediction?" Yes $\rightarrow$ Type II Bias

9 Conclusion

Through an investigation of 415 papers, we uncover the substantial confusion surrounding two prevalent types of biases within the machine learning community, which amplifies the learning burden for new researchers. Subsequently, we delve into the possible causes of the confusion. Most notably, we observe that researchers from diverse backgrounds hold different preconceptions about bias, leading to a lack of unified terminology for the same type of bias over an extended period. To alleviate the existing confusion and restore clarity in the literature, we present mathematical definitions for these two prevalent types of biases. Furthermore, we unify a comprehensive list of papers under these definitions and distinguish these two types of biases from multiple perspectives. Through this endeavor, we seek to facilitate the discussion on bias-related issues among researchers with diverse backgrounds.

References

Sravanti Addepalli, Anshul Nasery, Venkatesh Babu Radhakrishnan, Praneeth Netrapalli, and Prateek Jain. Feature reconstruction from outputs can mitigate simplicity bias in neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zH9GcZ3ZGXu.

Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2513–2523, 2021.

Dyah Adila, Shuai Zhang, Boran Han, and Yuyang Wang. Discovering bias in latent space: An unsupervised debiasing approach. *arXiv preprint arXiv:2406.03631*, 2024.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Sushant Agarwal and Amit Deshpande. On the power of randomization in fair classification and representation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1542–1551, 2022.

Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=b9PoimzZFJ.

Sumyeong Ahn, Seongyoon Kim, and Se-Young Yun. Mitigating dataset bias by using per-sample gradient. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=7mgUec-7GMv.

Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. Characterizing the risk of fairwashing. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=9PnKduzf-FT.

Sina Akbari, Ehsan Mokhtarian, AmirEmad Ghassami, and Negar Kiyavash. Recursive causal structure learning in the presence of latent variables and selection bias. *Advances in Neural Information Processing Systems*, 34:10119–10130, 2021.

Nil-Jana Akpinar, Zachary Lipton, and Alexandra Chouldechova. The impact of differential feature underreporting on algorithmic fairness. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1355–1382, 2024.

Ibrahim Alabdulmohsin and Mario Lucic. A near-optimal algorithm for debiasing trained machine learning models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=H5TBqNFPKSJ.

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Peter Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. CLIP the bias: How useful is balancing data in multimodal learning? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FIGXAxr9E4.

Ibrahim M Alabdulmohsin, Jessica Schrouff, and Sanmi Koyejo. A reduction to binary approach for debiasing multiclass datasets. *Advances in Neural Information Processing Systems*, 35:2480–2493, 2022.

Vítor Albiero and Kevin W Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. *arXiv preprint arXiv:2008.06989*, 2020.

Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

Isabela Albuquerque, Jessica Schrouff, David Warde-Farley, Taylan Cemgil, Sven Gowal, and Olivia Wiles. Evaluating model bias requires characterizing its mistakes. *arXiv preprint arXiv:2407.10633*, 2024.

Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, 2022.

Abdulaziz A Almuzaini, Chidansh A Bhatt, David M Pennock, and Vivek K Singh. Abcinml: Anticipatory bias correction in machine learning applications. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1552–1560, 2022.

Jose M Alvarez, Kristen M Scott, Bettina Berendt, and Salvatore Ruggieri. Domain adaptive decision trees: Implications for accuracy and fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 423–433, 2023.

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 289–295, 2019.

Bang An, Zora Che, Mucong Ding, and Furong Huang. Transferring fairness under distribution shifts via fair consistency regularization. *Advances in Neural Information Processing Systems*, 35:32582–32597, 2022.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022a.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022b.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 206–214, 2021.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkgsUJrtDB.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pp. 528–539. PMLR, 2020.

Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of biasin face analysis algorithms. In *Deep Learning-Based Face Analytics*, pp. 327–359. Springer, 2021.

Ari Ball-Burack, Michelle Seng Ah Lee, Jennifer Cobbe, and Jatinder Singh. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 116–128, 2021.

Mislav Balunovic, Anian Ruoss, and Martin Vechev. Fair normalizing flows. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BrFIKuxrZE.

Carlo Alberto Barbano, Benoit Dufumier, Enzo Tartaglione, Marco Grangetto, and Pietro Gori. Unbiased supervised contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Ph5cJSfD2XN.

Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454*, 2021.

Abhipsa Basu, Saswat Subhajyoti Mallick, and Venkatesh Babu Radhakrishnan. Mitigating biases in black-box feature extractors for image classification tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Joachim Baumann, Anikó Hannák, and Christoph Heitz. Enforcing group fairness in algorithmic decision making: Utility maximization under sufficiency. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2315–2326, 2022.

Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. Bias on demand: A modelling framework that generates synthetic data with bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1002–1013, 2023.

Joachim Baumann, Piotr Sapiezynski, Christoph Heitz, and Anikó Hannák. Fairness in online ad delivery. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1418–1432, 2024.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

Yahav Bechavod. Monotone individual fairness. *arXiv preprint arXiv:2403.06812*, 2024.

Ann-Kristin Becker, Oana Dumitrasc, and Klaus Broelemann. Standardized interpretable fairness measures for continuous risk scores. In *Forty-first International Conference on Machine Learning*.

Catarina G Belém, Preethi Seshadri, Yasaman Razeghi, and Sameer Singh. Are models biased on text without gender-related language? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=w1JanwReU6.

Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Zakharchenko, Lucas Rosenblatt, and Julia Stoyanovich. The possibility of fairness: Revisiting the impossibility theorem in practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 400–422, 2023.

Samuel James Bell and Levent Sagun. Simplicity bias leads to amplified performance disparities. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 355–369, 2023.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Kevin Bello and Jean Honorio. Fairness constraints can help exact inference in structured prediction. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11322–11332. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8248a99e81e752cb9b41da3fc43fbe7f-Paper.pdf.

Bilel Benbouzid. Fairness in machine learning from the perspective of sociology of statistics: How machine learning is becoming scientific by turning its back on metrological realism. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 35–43, 2023.

Harry Bendekgey and Erik B. Sudderth. Scalable and stable surrogates for flexible classifiers with fairness constraints. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Eyy4Tb1SY94.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Peter Bevan and Amir Atapour-Abarghouei. Skin deep unlearning: Artefact and instrument debiasing in the context of melanoma classification. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1874–1892. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bevan22a.html.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504, 2023.

Ruta Binkyte, Daniele Gorla, and Catuscia Palamidessi. Babe: Enhancing fairness via estimation of explaining variables. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1917–1925, 2024.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, pp. 149–159. PMLR, 2018.

Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin, and Daniel Ho. Algorithmic fairness and vertical equity: Income fairness with irs tax audit models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1479–1503, 2022.

Jack Blandin and Ian A Kash. Learning fairness from demonstrations via inverse reinforcement learning. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 51–61, 2024.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Avrim Blum, Kevin Stangl, and Ali Vakilian. Multi stage screening: Enforcing fairness and maximizing efficiency in a pre-existing pipeline. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1178–1193, 2022.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pp. 491–500, 2019.

Francois Buet-Golfouse and Islam Utyagulov. Towards fair unsupervised learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1399–1409, 2022.

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

Maarten Buyl and Tijl De Bie. Optimal transport of classifiers to fairness. *Advances in Neural Information Processing Systems*, 35:33728–33740, 2022.

Maarten Buyl, MaryBeth Defrance, and Tijl De Bie. fairret: a framework for differentiable fairness regularization terms. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NnyD0Rjx2B.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 370–378, 2023.

Kathleen Cachel and Elke Rundensteiner. Prefair: Combining partial preferences for fair consensus decision-making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1133–1149, 2024.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21:277–292, 2010.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pp. 13–18. IEEE, 2009.

Alessandra Calvi and Dimitris Kotzinos. Enhancing ai fairness through impact assessment in the european union: a legal and computer science perspective. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1229–1245, 2023.

Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gurbuzbalaban, and Umut Simsekli. Asymmetric heavy tails and implicit bias in gaussian noise injections. In *International Conference on Machine Learning*, pp. 1249–1260. PMLR, 2021.

Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 309–318, 2019.

Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, and Zhen Lei. Domain balancing: Face recognition on long-tailed domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020.

Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111, 2020.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 319–328, 2019.

L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International conference on machine learning*, pp. 1349–1359. PMLR, 2020.

L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1349–1361. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/celis21a.html.

Junyi Chai and Xiaoqian Wang. Self-supervised fair representation learning without demographics. *Advances in Neural Information Processing Systems*, 35:27100–27113, 2022a.

Junyi Chai and Xiaoqian Wang. Fairness with adaptive weights. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2853–2866. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/chai22a.html.

Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35:19152–19164, 2022.

Rwiddhi Chakraborty, Yinong Wang, Jialu Gao, Runkai Zheng, Cheng Zhang, and Fernando De la Torre. Visual data diagnosis and debiasing with concept graphs. *arXiv preprint arXiv:2409.18055*, 2024.

Eunice Chan, Zhining Liu, Ruizhong Qiu, Yuheng Zhang, Ross Maciejewski, and Hanghang Tong. Group fairness via group consensus. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1788–1808, 2024.

Peter W Chang, Leor Fishman, and Seth Neel. Feature importance disparities for data bias investigations. *arXiv preprint arXiv:2303.01704*, 2023.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=MO5PiKHELW.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3): 1–39, 2023a.

Lin Chen, Michal Lukasik, Wittawat Jitkrittum, Chong You, and Sanjiv Kumar. On bias-variance alignment in deep models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=i2Phucne30.

Mingliang Chen and Min Wu. Towards threshold invariant fair classification. In *Conference on Uncertainty in Artificial Intelligence*, pp. 560–569. PMLR, 2020.

Wenlong Chen, Yegor Klochkov, and Yang Liu. Post-hoc bias scoring is optimal for fair classification. In *The Twelfth International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=FM5xfcaR2Y.

Xin Chen, Zexing Xu, Zishuo Zhao, and Yuan Zhou. Personalized pricing with group fairness constraint. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1520–1530, 2023b.

Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. Fairness transferability subject to bounded distribution shift. *Advances in Neural Information Processing Systems*, 35:11266–11278, 2022.

Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14980–14991, 2021.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=N6JECD-PI5w.

Valeriia Cherepanova, Steven Reich, Samuel Dooley, Hossein Souri, John Dickerson, Micah Goldblum, and Tom Goldstein. A deep dive into dataset imbalance and bias in face identification. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 229–247, 2023.

John J Cherian and Emmanuel J Candès. Statistical inference for fairness auditing. *Journal of Machine Learning Research*, 25(149):1–49, 2024.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15088–15099. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ac3870fcad1cfc367825cda0101eee62-Paper.pdf.

Won Ik Cho, Jiwon Kim, Jaeyeong Yang, and Nam Soo Kim. Towards cross-lingual generalization of translation gender bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 449–457, 2021.

Somnath Basu Roy Chowdhury, Nicholas Monath, Ahmad Beirami, Rahul Kidambi, Kumar Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi. Enhancing group fairness in online settings using oblique decision forests. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=E1NxN5QMOE.

Sanghyeok Chu, Dongwan Kim, and Bohyung Han. Learning debiased and disentangled representations for semantic segmentation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=sUFdZqWeMM.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=DNl5s5BXeBn.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7321–7331. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019a.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019b.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Learning to model and ignore dataset bias with mixed capacity ensembles. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3031–3045, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.272. URL https://aclanthology.org/2020.findings-emnlp.272.

Jean-Rémy Conti and Stephan Clémençon. Assessing uncertainty in similarity scoring: Performance & fairness in face recognition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lAhQCHuANV.

Jean-Rémy Conti, Nathan Noiry, Stephan Clemencon, Vincent Despiegel, and Stéphane Gentric. Mitigating gender bias in face recognition using the von mises-fisher mixture model. In *International Conference on Machine Learning*, pp. 4344–4369. PMLR, 2022.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 582–593, 2020.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *International conference on machine learning*, pp. 2185–2195. PMLR, 2020.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

André Cruz and Moritz Hardt. Unprocessing seven years of algorithmic fairness. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jr03SfWsBS.

André Cruz, Catarina G Belém, João Bravo, Pedro Saleiro, and Pedro Bizarro. FairGBM: Gradient boosting with fairness constraints. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=x-mXzBgCX3a.

Pádraig Cunningham and Sarah Jane Delany. Underestimation bias and underfitting in machine learning. In *Trustworthy AI-Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, September 4–5, 2020, Revised Selected Papers 1*, pp. 20–31. Springer, 2021.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534, 2020.

Irene Dankwa-Mullan, Marc Rivo, Marisol Sepulveda, Yoonyoung Park, Jane Snowdon, and Kyu Rhee. Transforming diabetes care through artificial intelligence: the future is here. *Population health management*, 22(3):229–242, 2019.

Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the european conference on computer vision (eccv) workshops*, pp. 0–0, 2018.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019a.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019b.

Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 52–59, 2019.

Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. A critical survey on fairness benefits of explainable ai. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1579–1595, 2024.

Marybeth Defrance and Tijl De Bie. Maximal fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 851–880, 2023.

MaryBeth Defrance, Maarten Buyl, and Tijl De Bie. Abcfair: an adaptable benchmark approach for comparing fairness methods. *arXiv preprint arXiv:2409.16965*, 2024.

Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. FairerCLIP: Debiasing CLIP's zero-shot predictions using functions in RKHSs. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=HXoq9EqR9e.

Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. FairerCLIP: Debiasing CLIP's zero-shot predictions using functions in RKHSs. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=HXoq9EqR9e.

Eoin Delaney, Zihao Fu, Sandra Wachter, Brent Mittelstadt, and Chris Russell. Oxonfair: A flexible toolkit for algorithmic fairness. *arXiv preprint arXiv:2407.13710*, 2024.

Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J Su, and James Zou. FIFA: Making fairness more generalizable in classifiers trained on imbalanced data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zVrw4OH1Lch.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2083–2102, 2022.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 862–872, 2021.

Prithviraj Dhar, Joshua Gleason, Hossein Souri, Carlos D Castillo, and Rama Chellappa. An adversarial learning algorithm for mitigating gender bias in face recognition. *arXiv preprint arXiv:2006.07845*, 2, 2020.

Prithviraj Dhar, Joshua Gleason, Aniket Roy, Carlos D Castillo, and Rama Chellappa. Pass: Protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15087–15096, 2021.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=bYi_2708mKK.

Nancy DiTomaso. Racism and discrimination versus advantage and favoritism: Bias for versus bias against. *Research in Organizational Behavior*, 35:57–77, 2015.

Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1639–1656, 2022.

Evan Dong, Aaron Schein, Yixin Wang, and Nikhil Garg. Addressing discretization-induced bias in demographic prediction. *arXiv preprint arXiv:2405.16762*, 2024.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/83cdcec08fbf90370fcf53bdd56604ff-Paper.pdf.

Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2020.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Hassan Awadallah, and Xia Hu. Fairness via representation neutralization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=nHRGW_wETLQ.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pp. 2803–2813. PMLR, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Gary Edmond and Kristy A Martire. Just cognition: scientific research on bias and some implications for legal procedure and decision-making. *The Modern Law Review*, 82(4):633–664, 2019.

Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022.

Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. Debiasing graph neural networks via learning disentangled causal substructure. *Advances in Neural Information Processing Systems*, 35:24934–24946, 2022.

Ronglong Fang and Yuesheng Xu. Addressing spectral bias of deep neural networks by multi-grade deep learning. *arXiv preprint arXiv:2410.16105*, 2024.

Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 489–503, 2021.

Chloë FitzGerald and Samia Hurst. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18(1):1–18, 2017.

Hortense Fong, Vineet Kumar, Anay Mehrotra, and Nisheeth K Vishnoi. Fairness for auc via feature augmentation. *arXiv preprint arXiv:2111.12823*, 2021.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338, 2019.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Prakhar Ganesh, Hongyan Chang, Martin Strobel, and Reza Shokri. On the impact of machine learning randomness on group fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1789–1800, 2023.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Joshua Gardner, Renzhe Yu, Quan Nguyen, Christopher Brooks, and Rene Kizilcec. Cross-institutional transfer learning for educational models: Implications for model performance, fairness, and equity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1664–1684, 2023.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3197–3208. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/20d749bc05f47d2bd3026ce457dcfd8e-Paper.pdf.

Khashayar Gatmiry, Zhiyuan Li, Sashank J Reddi, and Stefanie Jegelka. Simplicity bias via global convergence of sharpness minimization. *arXiv preprint arXiv:2410.16401*, 2024.

Solenne Gaucher, Alexandra Carpentier, and Christophe Giraud. The price of unfairness in linear bandits with biased feedback. *Advances in Neural Information Processing Systems*, 35:18363–18376, 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Azin Ghazimatin, Matthaus Kleindessner, Chris Russell, Ziawasch Abedjan, and Jacek Golebiowski. Measuring fairness of rankings under noisy sensitive information. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2263–2279, 2022.

Stephen Giguere, Blossom Metevier, Yuriy Brun, Philip S. Thomas, Scott Niekum, and Bruno Castro da Silva. Fairness guarantees under demographic shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=wbPObLm6ueA.

Talia B Gillis, Vitaly Meursault, and Berk Ustun. Operationalizing the search for less discriminatory alternatives in fair lending. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 377–387, 2024.

Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=9YlaeLfuhJF.

Paul Gölz, Anson Kahng, and Ariel D Procaccia. Paradoxes in fair machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *European conference on computer vision*, pp. 330–347. Springer, 2020.

Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3414–3424, 2021.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International conference on machine learning*, pp. 2357–2365. PMLR, 2019.

Przemyslaw A Grabowicz, Nicholas Perello, and Aarshee Mishra. Marrying fairness and explainability in supervised learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1905–1916, 2022.

Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Learning unbiased representations via rényi minimization. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pp. 749–764. Springer, 2021.

Vincent Grari, Thibault Laugel, Tatsunori Hashimoto, sylvain lamprier, and Marcin Detyniecki. On the fairness ROAD: Robust optimization for adversarial debiasing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xnhvVtZtLD.

Riccardo Grazzi, Arya Akhavan, John IF Falk, Leonardo Cella, and Massimiliano Pontil. Group meritocratic fairness in linear contextual bandits. *Advances in Neural Information Processing Systems*, 35:24392–24404, 2022.

Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 90–99, 2019.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, pp. 11. Barcelona, Spain, 2016.

Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology, 2019.

Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z Li. Learning meta face recognition in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6163–6172, 2020.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pp. 87–102. Springer, 2016.

Mingming Ha, Xuewen Tao, Wenfang Lin, Qionxu Ma, Wujiang Xu, and Linxun Chen. Fine-grained dynamic framework for bias-variance joint optimization on data missing not at random. *arXiv preprint arXiv:2405.15403*, 2024.

Hu Han, Anil K Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 40 (11):2597–2609, 2017.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. FFB: A fair fairness benchmark for in-processing group fairness methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=TzAJbTClAz.

Xudong Han, Timothy Baldwin, and Trevor Cohn. Everybody needs good neighbours: An unsupervised locality-based method for bias mitigation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=pOnhudsvzR.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 392–402, 2020.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv:1908.10763*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE conf. on computer vision and pattern recognition*, pp. 770–778, 2016.

Xilin He, Jingyu Hu, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Muhammad Haris Khan, and Linlin Shen. Towards combating frequency simplicity-biased learning for domain generalization. *arXiv preprint arXiv:2410.16146*, 2024.

Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 181–190, 2019.

Thomas Henzinger, Mahyar Karimi, Konstantin Kueffner, and Kaushik Mallik. Runtime monitoring of dynamic fairness properties. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 604–614, 2023.

Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. A structural approach to selection bias. *Epidemiology*, pp. 615–625, 2004.

Gaurush Hiranandani, Harikrishna Narasimhan, and Sanmi Koyejo. Fair performance metric elicitation. *Advances in Neural Information Processing Systems*, 33:11083–11095, 2020.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Gender and racial bias in visual question answering datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1280–1292, 2022.

Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021.

Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Bias mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

Ramtin Hosseini, Li Zhang, Bhanu Garg, and Pengtao Xie. Fair and accurate decision making through group-aware learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13254–13269. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hosseini23a.html.

Brian Hsu, Rahul Mazumder, Preetam Nandy, and Kinjal Basu. Pushing the limits of fairness impossibility: Who's the fairest of them all? *Advances in Neural Information Processing Systems*, 35:32749–32761, 2022.

Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 535–545, 2020.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 49–58, 2019.

Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 230–247, 2022.

Inwoo Hwang, Sangjun Lee, Yunhyeok Kwak, Seong Joon Oh, Damien Teney, Jin-Hwa Kim, and Byoung-Tak Zhang. Selecmix: Debiased learning by contradicting-pair sampling. *Advances in Neural Information Processing Systems*, 35:14345–14357, 2022.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

Sofia Jaime and Christoph Kern. Ethnic classifications in algorithmic fairness: Concepts, measures and implications in practice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 237–253, 2024.

Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised Adversarial Invariance. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 5097–5107. Curran Associates, Inc., 2018.

Myeongho Jeon, Daekyung Kim, Woochul Lee, Myungjoo Kang, and Joonseok Lee. A conservative approach for unbiased learning on unknown biases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16752–16760, 2022.

Sérgio Jesus, José Pombal, Duarte Alves, André Cruz, Pedro Saleiro, Rita Ribeiro, João Gama, and Pedro Bizarro. Turning the tables: Biased, imbalanced, dynamic tabular datasets for ml evaluation. *Advances in Neural Information Processing Systems*, 35:33563–33575, 2022.

Disi Ji, Padhraic Smyth, and Mark Steyvers. Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18600–18612. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d83de59e10227072a9c034ce10029c39-Paper.pdf.

Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2021.

Jiayin Jin, Zeru Zhang, Yang Zhou, and Lingfei Wu. Input-agnostic certified group fairness via Gaussian parameter smoothing. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10340–10361. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jin22g.html.

Jinqiu Jin, Haoxuan Li, and Fuli Feng. On the maximal local disparity of fairness-aware classifiers. *arXiv preprint arXiv:2406.03255*, 2024a.

Ruinan Jin, Zikang Xu, Yuan Zhong, Qiongsong Yao, Qi Dou, S Kevin Zhou, and Xiaoxiao Li. Fairmedfm: fairness benchmarking for medical imaging foundation models. *arXiv preprint arXiv:2407.00983*, 2024b.

Nikola Jovanović, Mislav Balunović, Dimitar I. Dimitrov, and Martin Vechev. Fare: Provably fair representation learning with practical certificates. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A unified debiasing approach for vision-language models across modalities and tasks. *Advances in Neural Information Processing Systems*, 37:21034–21058, 2025a.

Sangwon Jung, Sanghyuk Chun, and Taesup Moon. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10348–10357, 2022.

Sangwon Jung, Taeeon Park, Sanghyuk Chun, and Taesup Moon. Re-weighting based group fairness regularization via classwise robust optimization. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=Q-WfHzmiG9m.

Sangwon Jung, Sumin Yu, Sanghyuk Chun, and Taesup Moon. Do counterfactually fair image classifiers satisfy group fairness?–a theoretical and empirical study. *Advances in Neural Information Processing Systems*, 37:56041–56053, 2025b.

Yeonsung Jung, Hajin Shim, June Yong Yang, and Eunho Yang. Fighting fire with fire: Contrastive debiasing without bias-free data via generative bias-transformation. In *International Conference on Machine Learning*, pp. 15435–15450. PMLR, 2023b.

Yeonsung Jung, Jaeyun Song, June Yong Yang, Jin-Hwa Kim, Sung-Yub Kim, and Eunho Yang. A simple remedy for dataset bias via self-influence: A mislabeled sample perspective. *arXiv preprint arXiv:2411.00360*, 2024.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2439–2448. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kallus18a.html.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Meina Kan, Shiguang Shan, and Xilin Chen. Bi-shifting auto-encoder for unsupervised domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3846–3854, 2015.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020.

Jian Kang, Yinglong Xia, Ross Maciejewski, Jiebo Luo, and Hanghang Tong. Deceptive fairness attacks on graphs via meta learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=iS5ADHNg2A.

Mintong Kang, Linyi Li, Maurice Weber, Yang Liu, Ce Zhang, and Bo Li. Certifying some distributional fairness with subpopulation decomposition. *Advances in Neural Information Processing Systems*, 35: 31045–31058, 2022.

Kimmo Karkkainen and Jungseock 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*, pp. 1548–1558, 2021.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 100–109, 2019.

Mohammad Mahdi Khalili, Xueru Zhang, and Mahed Abroshan. Loss balancing for fair supervised learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2630–2639. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kilbertus18a.html.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9012–9020, 2019a.

Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14992–15001, 2021.

Jin-Young Kim and Sung-Bae Cho. Fair representation for safe artificial intelligence via adversarial learning of unbiased information bottleneck. In *SafeAI@ AAAI*, pp. 105–112, 2020.

Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pp. 5264–5274. PMLR, 2020.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019b.

Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. *Advances in Neural Information Processing Systems*, 35:18403–18415, 2022.

Yeongmin Kim, Byeonghu Na, Minsang Park, JoonHo Jang, Dongjun Kim, Wanmo Kang, and Il chul Moon. Training unbiased diffusion models from biased dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=39cPKijBed.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1931–1939, 2015.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Shuming Kong, Yanyan Shen, and Linpeng Huang. Resolving training biases via influence-based data relabeling. In *International Conference on Learning Representations*, 2021.

Youjin Kong. Are "intersectionally fair" ai algorithms really fair to women of color? a philosophical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 485–494, 2022.

Nikola Konstantinov and Christoph H Lampert. Fairness-aware pac learning from corrupted data. *The Journal of Machine Learning Research*, 23(1):7173–7232, 2022.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. Noise-tolerant fair classification. *Advances in neural information processing systems*, 32, 2019.

Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

Mike Laszkiewicz, Imant Daunhawer, Julia E Vogt, Asja Fischer, and Johannes Lederer. Benchmarking the fairness of image upsampling methods. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 489–517, 2024.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International conference on machine learning*, pp. 1078–1088. PMLR, 2020.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

Tosca Lechner, Shai Ben-David, Sushant Agarwal, and Nivasini Ananthakrishnan. Impossibility results for fair representations. *arXiv preprint arXiv:2107.03483*, 2021.

Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021.

Messi HJ Lee, Jacob M Montgomery, and Calvin K Lai. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1321–1340, 2024.

Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, and Anupam Datta. Feature-wise bias amplification. *arXiv preprint arXiv:1812.08999*, 2018.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

Jiazhi Li and Wael Abd-Almageed. Information-theoretic bias assessment of learned representations of pretrained face recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–8. IEEE, 2021.

Jiazhi Li and Wael Abd-Almageed. Cat: Controllable attribute translation for fair facial attribute classification. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pp. 363–381. Springer, 2023.

Jiazhi Li and Wael AbdAlmageed. Ethics and fairness for diabetes artificial intelligence. In Dr. Klonoff, Dr. David Kerr, and Dr. Juan Espinoza (eds.), *Diabetes Digital Health, Telehealth, and Artificial Intelligence*. Elsevier, 2024.

Jiazhi Li, Mahyar Khayatkhoei, Jiageng Zhu, Hanchen Xie, Mohamed E Hussein, and Wael AbdAlmageed. Information-theoretic bounds on the removal of attribute-specific bias from neural networks. *arXiv preprint arXiv:2310.04955*, 2023a.

Jiazhi Li, Mahyar Khayatkhoei, Jiageng Zhu, Hanchen Xie, Mohamed E Hussein, and Wael AbdAlmageed. Sabaf: Removing strong attribute bias from neural networks with adversarial filtering. *arXiv preprint arXiv:2311.07141*, 2023b.

Mingxiao Li, Tingyu Qu, Ruicong Yao, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZSD3MloKe6.

Peizhao Li and Hongfu Liu. Achieving fairness at no utility cost via data reweighing with influence. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12917–12930. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22p.html.

Peizhao Li, Yifei Wang, Han Zhao, Pengyu Hong, and Hongfu Liu. On dyadic fairness: Exploring and mitigating bias in graph connections. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=xgGS6PmzNq6.

Tianlin Li, Qing Guo, Aishan Liu, Mengnan Du, Zhiming Li, and Yang Liu. Fairer: Fairness as decision rationale alignment. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023c.

Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9572–9581, 2019.

Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 513–528, 2018.

Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and cihang xie. Shape-texture debiased neural network training. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=Db4yerZTYkz.

Yujia Li, Kevin Swersky, and Richard Zemel. Learning unbiased features. *arXiv preprint arXiv:1412.5244*, 2014.

Zhiheng Li and Chenliang Xu. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14970–14979, 2021.

Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pp. 270–288. Springer, 2022.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6565–6576. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liang21a.html.

Tomasz Limisiewicz, David Mareček, and Tomáš Musil. Debiasing algorithm through model adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XIZEFyVGC9.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Chang Liu, Xiang Yu, Yi-Hsuan Tsai, Masoud Faraki, Ramin Moslemi, Manmohan Chandraker, and Yun Fu. Learning to learn across diverse data biases in deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4072–4082, 2022a.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021a.

Jiaheng Liu, Yudong Wu, Yichao Wu, Chuming Li, Xiaolin Hu, Ding Liang, and Mengyu Wang. Dam: discrepancy alignment metric for face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3814–3823, 2021b.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3150–3158. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/liu18c.html.

Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 381–391, 2020.

Meichen Liu, Lei Ding, Dengdeng Yu, Wulong Liu, Linglong Kong, and Bei Jiang. Conformalized fairness via quantile regression. *Advances in Neural Information Processing Systems*, 35:11561–11572, 2022b.

Pangpang Liu and Yichuan Zhao. Empirical likelihood for fair classification. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GACjMj1MS1.

Sheng Liu, Xu Zhang, Nitesh Sekhar, Yue Wu, Prateek Singhal, and Carlos Fernandez-Granda. Avoiding spurious correlations via logit correction. In *The Eleventh International Conference on Learning Representations*, 2023.

Tianci Liu, Haoyu Wang, Feijie Wu, Hengtong Zhang, Pan Li, Lu Su, and Jing Gao. Towards poisoning fair representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YLJs4mKJCF.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. of the IEEE International Conf. on computer vision*, pp. 3730–3738, 2015.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in neural information processing systems*, 32, 2019.

Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning*, pp. 6360–6369. PMLR, 2020.

Michael Lohaus, Matthäus Kleindessner, Krishnaram Kenthapadi, Francesco Locatello, and Chris Russell. Are two heads the same as one? identifying disparate treatment in fair neural networks. *Advances in Neural Information Processing Systems*, 35:16548–16562, 2022.

Michele Loi and Christoph Heitz. Is calibration a fairness requirement? an argument from the point of view of moral philosophy and decision theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2026–2034, 2022.

Vishnu Suresh Lokhande, Aditya Kumar Akash, Sathya N Ravi, and Vikas Singh. Fairalm: Augmented lagrangian method for training fair models with little regret. In *European Conference on Computer Vision*, pp. 365–381. Springer, 2020.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Shenyu Lu, Yipei Wang, and Xiaoqian Wang. Debiasing attention mechanism in transformer without demographics. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jLIUfrAcMQ.

Marco Lünich and Birte Keller. Explainable artificial intelligence for academic performance prediction. an experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1031–1042, 2024.

Zihan Luo, Hong Huang, Yongkang Zhou, Jiping Zhang, Nuo Chen, and Hai Jin. Are your models still fair? fairness attacks on graph neural networks via node injections. *arXiv preprint arXiv:2406.03052*, 2024.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Aa5oPXc_1IV.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. Ensuring fairness beyond the training data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18445–18456. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d6539d3b57159babf6a72e106beb45bd-Paper.pdf.

Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi. Differential privacy has bounded impact on fairness in classification. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Nina Markl. Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 521–534, 2022.

Joshua L Martin and Kelly Elizabeth Wright. Bias in automatic speech recognition: The case of african american language. *Applied Linguistics*, 44(4):613–630, 2023.

Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conf. on Biometrics (ICB)*, pp. 158–165. IEEE, 2018.

Melissa Mccradden, Oluwadara Odusi, Shalmali Joshi, Ismail Akrout, Kagiso Ndlovu, Ben Glocker, Gabriel Maicas, Xiaoxuan Liu, Mjaye Mazwi, Tee Garnett, et al. What's fair is... fair? presenting justefab, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: Justefab. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1505–1519, 2023.

Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. Characterizing bias in classifiers using generative models. *Advances in neural information processing systems*, 32, 2019.

Bryce McLaughlin, Jann Spiess, and Talia Gillis. On the fairness of machine-assisted human decisions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 890–890, 2022.

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

Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8930–8938, 2021b.

Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1699–1710, 2023.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pp. 107–118. PMLR, 2018.

Vishwali Mhasawade, Alexander D'Amour, and Stephen R Pfohl. A causal perspective on label bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1282–1294, 2024.

Jennifer Mickel. Racial/ethnic categories in ai and algorithmic fairness: Why they matter and what they represent. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2484–2494, 2024.

S. Milborrow, J. Morkel, and F. Nicolls. The MUCT Landmarked Face Database. *Pattern Recognition Association of South Africa*, 2010. http://www.milbo.org/muct.

Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–10. IEEE, 2018.

Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. *IEEE Access*, 7:99735–99745, 2019.

Alan Mishler and Edward Kennedy. Fade: Fair double ensemble learning for observable and counterfactual outcomes. *arXiv preprint arXiv:2109.00173*, 2021.

Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 386–400, 2021.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. Dqi: A guide to benchmark evaluation. *arXiv preprint arXiv:2008.03964*, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=t4485RO6O8P.

Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2158–2164, 2020.

Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*, pp. 7066–7075. PMLR, 2020.

Chaithanya Kumar Mummadi, Ranjitha Subramaniam, Robin Hutmacher, Julien Vitay, Volker Fischer, and Jan Hendrik Metzen. Does enhanced shape bias improve neural network robustness to common corruptions? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=yUxUNaj2Sl.

Kamesh Munagala and Govind S Sankar. Individual fairness in graph decomposition. *arXiv preprint arXiv:2406.00213*, 2024.

Shruti Nagpal, Maneet Singh, Richa Singh, and Mayank Vatsa. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022.

Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 466–477, 2021.

Preetam Nandy, Cyrus Diciccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. Achieving fairness via post-processing in web-scale recommender systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 715–725, 2022.

Dang Nguyen, Paymon Haddad, Eric Gan, and Baharan Mirzasoleiman. Changing the training data distribution to reduce simplicity bias improves in-distribution generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2019.

Hongliang Ni, Lei Han, Tong Chen, Shazia Sadiq, and Gianluca Demartini. Fairness without sensitive attributes via knowledge sharing. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1897–1906, 2024.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Guy Ohayon, Michael Elad, and Tomer Michaeli. Perceptual fairness in image restoration. *arXiv preprint arXiv:2405.13805*, 2024.

Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15360–15370. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/af9c0e0c1dee63e5acad8b7ed1a5be96-Paper.pdf.

Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid. Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 973–987, 2022.

Jinlong Pang, Jialu Wang, Zhaowei Zhu, Yuanshun Yao, Chen Qian, and Yang Liu. Fairness without harm: An influence-guided active sampling approach. *Advances in Neural Information Processing Systems*, 37: 61513–61548, 2025.

Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10389–10398, 2022.

Ioannis Pastaltzidis, Nikolaos Dimitriou, Katherine Quezada-Tavarez, Stergios Aidinlis, Thomas Marquenie, Agata Gurzawska, and Dimitrios Tzovaras. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2302–2314, 2022.

Eike Petersen, Melanie Ganz, Sune Holm, and Aasa Feragen. On (assessing) the fairness of risk score models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 817–829, 2023.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Stephen Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam Shah. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1039–1052, 2022.

Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. Fairness and accuracy under domain generalization. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jBEXnEMdNOL.

Drago Plecko and Elias Bareinboim. Mind the gap: A causal perspective on bias amplification in prediction & decision-making. *arXiv preprint arXiv:2405.15446*, 2024.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Robert Lee Poe and Soumia Zohra El Mestari. The conflict between algorithmic fairness and non-discrimination: An analysis of fair automated hiring. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1907–1916, 2024.

Shangshu Qian, Hung Viet Pham, Thibaud Lutellier, Zeou Hu, Jungwon Kim, Lin Tan, Yaoliang Yu, Jiahao Chen, and Sameena Shah. Are my deep learning systems fair? an empirical study of fixed-seed training. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=kLWGdQYsmC5.

Haoyu Qin. Asymmetric rejection loss for fairer face recognition. *arXiv preprint arXiv:2002.03276*, 2020.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. *arXiv preprint arXiv:2306.11074*, 2023.

Novi Quadrianto and Viktoriia Sharmanska. Recycling privileged learning and distribution matching for fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8227–8236, 2019.

Tangkun Quan, Fei Zhu, Quan Liu, and Fanzhang Li. Learning fair representations for accuracy parity. *Engineering Applications of Artificial Intelligence*, 119:105819, 2023.

Ruggero Ragonesi, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. Learning unbiased representations via mutual information backpropagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2729–2738, 2021.

Aida Rahmattalabi, Phebe Vayanos, Anthony Fulginiti, Eric Rice, Bryan Wilder, Amulya Yadav, and Milind Tambe. Exploring algorithmic fairness in robust graph covering problems. *Advances in neural information processing systems*, 32, 2019.

Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9301–9310, 2021.

Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P Gummadi, and Isabel Valera. Don't throw it away! the utility of unlabeled data in fair decision making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1421–1433, 2022.

Miriam Rateike, Isabel Valera, and Patrick Forré. Designing long-term group fair policies in dynamical systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 20–50, 2024.

Tim Räz. Group fairness: Independence revisited. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 129–137, 2021.

Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. A distributional simplicity bias in the learning dynamics of transformers. *arXiv preprint arXiv:2410.19637*, 2024.

K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 341–345, 2006. doi: 10.1109/FGR.2006.78.

Brianna Richardson, Prasanna Sattigeri, Dennis Wei, Karthikeyan Natesan Ramamurthy, Kush Varshney, Amit Dhurandhar, and Juan E Gilbert. Add-remove-or-relabel: Practitioner-friendly bias mitigation via influential fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 736–752, 2023.

Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–1, 2020.

Kit T Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 142–153, 2020.

Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Fr-train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learning*, pp. 8147–8157. PMLR, 2020.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=YNnpaAKeCfx.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Sample selection for fair and robust training. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021b. URL https://openreview.net/forum?id=2Dg2UQyRpQ.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Improving fair training under correlation shifts. *arXiv preprint arXiv:2302.02323*, 2023.

Eliane Röösli, Selen Bozkurt, and Tina Hernandez-Boussard. Peeking into a black box, the fairness and generalizability of a mimic-iii benchmarking model. *Scientific Data*, 9(1):24, 2022.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 10–15, 2015.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018.

Kai Ruggeri, Sarah Ashcroft-Jones, Giampaolo Abate Romero Landini, Narjes Al-Zahli, Natalia Alexander, Mathias Houe Andersen, Katherine Bibilouri, Katharina Busch, Valentina Cafarelli, Jennifer Chen, et al. The persistence of cognitive biases in financial decisions across economic groups. *Scientific Reports*, 13(1): 10329, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2017.

Sivan Sabato and Elad Yom-Tov. Bounding the fairness and accuracy of classifiers from population statistics. In *International conference on machine learning*, pp. 8316–8325. PMLR, 2020.

Bashir Sadeghi, Runyi Yu, and Vishnu Boddeti. On the global optima of kernelized adversarial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7971–7979, 2019.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. Measuring non-expert comprehension of machine learning fairness metrics. In *International Conference on Machine Learning*, pp. 8377–8387. PMLR, 2020.

Tiago Salvador, Stephanie Cairns, Vikram Voleti, Noah Marshall, and Adam M Oberman. Faircal: Fairness calibration for face verification. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nRj0NcmSuxb.

Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair pca: One extra dimension. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/cc4af25fa9d2d5c953496579b75f6f6c-Paper.pdf.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Hf3qXoiNkR.

Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

Prasanna Sattigeri, Soumya Ghosh, Inkit Padhi, Pierre Dognin, and Kush R Varshney. Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting. *Advances in Neural Information Processing Systems*, 35:35894–35906, 2022.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BXewfAYMmJw.

Yash Savani, Colin White, and Naveen Sundar Govindarajulu. Intra-processing methods for debiasing neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2798–2810. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1d8d70dddf147d2d92a634817f01b239-Paper.pdf.

Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems*, 35:19304–19318, 2022.

Jessica Schrouff, Alexis Bellot, Amal Rannen-Triki, Alan Malek, Isabela Albuquerque, Arthur Gretton, Alexander D'Amour, and Silvia Chiappa. Mind the graph when balancing data for fairness or robustness. *arXiv preprint arXiv:2406.17433*, 2024.

Pola Schwöbel and Peter Remmers. The long arc of fairness: Formalisations and ethical discourse. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2179–2188, 2022.

Khotso Selialia, Yasra Chandio, and Fatima M Anwar. Mitigating group bias in federated learning for heterogeneous devices. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1043–1054, 2024.

Ignacio Serna, Aythami Morales, Julian Fierrez, and Nick Obradovich. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, 305: 103682, 2022.

Ali Shahin Shamsabadi, Mohammad Yaghini, Natalie Dullerud, Sierra Wyllie, Ulrich Aïvodji, Aisha Alaagib, Sébastien Gambs, and Nicolas Papernot. Washing the unwashable: On the (im) possibility of fairwashing detection. *Advances in Neural Information Processing Systems*, 35:14170–14182, 2022.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

Amr Sharaf, Hal Daume III, and Renkun Ni. Promoting fairness in learned models by learning to active learn under parity constraints. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2149–2156, 2022.

Mohit Sharma and Amit Deshpande. How far can fairness constraints help recover from biased data? *arXiv preprint arXiv:2312.10396*, 2023.

Maohao Shen, Yuheng Bu, and Gregory W Wornell. On balancing bias and variance in unsupervised multi-source-free domain adaptation. In *International Conference on Machine Learning*, pp. 30976–30991. PMLR, 2023.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hnrB5YHoYu.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Hilson Shrestha, Kathleen Cachel, Mallak Alkhathlan, Elke Rundensteiner, and Lane Harrison. Help or hinder? evaluating the impact of fairness metrics and algorithms in visualizations for consensus ranking. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1685–1698, 2023.

Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, and Christian Gagné. Fair representation learning through implicit path alignment. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20156–20175. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/shui22a.html.

Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles X Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. *Advances in Neural Information Processing Systems*, 35:34121–34135, 2022b.

Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. Getfair: Generalized fairness tuning of classification models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 289–299, 2022.

Jan Simson, Florian Pfisterer, and Christoph Kern. One model many scores: Using multiverse analysis to prevent fairness hacking and evaluate the influence of model design decisions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1305–1320, 2024.

Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3–13, 2021.

Harvineet Singh, Matthäus Kleindessner, Volkan Cevher, Rumi Chunara, and Chris Russell. When do minimax-fair learning and empirical risk minimization coincide? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11070–11078, 2020.

Dylan Slack, Sorelle A Friedler, and Emile Givental. Fairness warnings and fair-maml: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 200–209, 2020.

Edward Small, Kacper Sokol, Daniel Manning, Flora D Salim, and Jeffrey Chan. Equalised odds is not equal individual odds: Post-processing for group and individual fairness. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1559–1578, 2024.

Alexander Soen, Ibrahim M Alabdulmohsin, Sanmi Koyejo, Yishay Mansour, Nyalleng Moorosi, Richard Nock, Ke Sun, and Lexing Xie. Fair wrapping for black-box predictions. *Advances in Neural Information Processing Systems*, 35:21615–21627, 2022.

Alexander Soen, Hisham Husain, and Richard Nock. Fair densities via boosting the sufficient statistics of exponential families, 2023.

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

Seamus Somerstep, Ya'acov Ritov, and Yuekai Sun. Algorithmic fairness in performative policy learning: Escaping the impossibility of group fairness. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 616–630, 2024.

Silpa Vadakkeeveetil Sreelatha, Adarsh Kappiyath, Abhra Chaudhuri, and Anjan Dutta. Denetdm: Debiasing by network depth modulation. *arXiv preprint arXiv:2403.19863*, 2024.

Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2): 20539517221115189, 2022.

Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 701–713, 2021.

Rebecca S Stone, Nishant Ravikumar, Andrew J Bulpitt, and David C Hogg. Epistemic uncertainty-weighted loss for visual bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2898–2905, 2022.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

Vinith Menon Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. When personalization harms performance: Reconsidering the use of group attributes in prediction. In *International Conference on Machine Learning*. PMLR, 2023.

Chris Sweeney and Maryam Najafian. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 359–368, 2020.

Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

Saeid A Taghanaki, Kristy Choi, Amir Hosein Khasahmadi, and Anirudh Goyal. Robust representation learning via perceptual similarity metrics. In *International Conference on Machine Learning*, pp. 10043–10053. PMLR, 2021.

Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32, 2019.

Zeyu Tang, Jialu Wang, Yang Liu, Peter Spirtes, and Kun Zhang. Procedural fairness through decoupling objectionable data generating components. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=cxfPefbu1s.

Rohan Taori and Tatsunori Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *International Conference on Machine Learning*, pp. 33883–33920. PMLR, 2023.

Enzo Tartaglione, Carlo Alberto Barbano, and Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13508–13517, 2021.

Bahar Taskesen, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. A statistical test for probabilistic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 648–665, 2021.

Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pp. 53–59, 2017.

Muhammad Faaiz Taufiq, Jean-Francois Ton, and Yang Liu. Achievable fairness on your data with utility guarantees. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16761–16772, 2022.

Christopher TH Teo, Milad Abdollahzadeh, Xinda Ma, and Ngai-man Cheung. Fairqueue: Rethinking prompt learning for fair text-to-image generation. *arXiv preprint arXiv:2410.18615*, 2024.

Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332–338, 2020a.

Philipp Terhörst, Mai Ly Tran, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Comparison-level mitigation of ethnic bias in face recognition. *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6, 2020b. URL https://api.semanticscholar.org/CorpusID:219547273.

Yu Tian, Min Shi, Yan Luo, Ava Kouhana, Tobias Elze, and Mengyu Wang. Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=qNrJJZAKI3.

Alexandru Tifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. Frappé: A group fairness framework for post-processing everything. In *Forty-first International Conference on Machine Learning*.

Rishabh Tiwari and Pradeep Shenoy. Overcoming simplicity bias in deep networks using a feature sieve. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 34330–34343. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/tiwari23a.html.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Puja Trivedi, Danai Koutra, and Jayaraman J. Thiagarajan. A closer look at model adaptation using feature distortion and simplicity bias. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=wkg_b4-IwTZ.

Nikita Tsoy and Nikola Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. *arXiv preprint arXiv:2405.17299*, 2024.

Mycal Tucker and Julie A. Shah. Prototype based classification from hierarchy to fairness. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21884–21900. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tucker22a.html.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. *arXiv preprint arXiv:2005.00315*, 2020.

Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

Jacobus van der Linden, Mathijs de Weerdt, and Emir Demirović. Fair and optimal decision trees: A dynamic programming approach. *Advances in Neural Information Processing Systems*, 35:38899–38911, 2022.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Mariia Vladimirova, Federico Pavone, and Eustache Diemert. Fairjob: A real-world dataset for fairness in online systems. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 10442–10469. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/142bff4f4c01dd55c4309860ff3a59f1-Paper-Datasets_and_Benchmarks_Track.pdf.

Hieu Vu, Toan Tran, Man-Chung Yue, and Viet Anh Nguyen. Distributionally robust fair principal components via geodesic descents. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9NVd-DMtThY.

Ada Wan. Fairness in representation for multilingual NLP: Insights from controlled experiments on conditional language modeling. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=-llS6TiOew.

Angelina Wang and Olga Russakovsky. Directional bias amplification. In *International Conference on Machine Learning*, pp. 10882–10893. PMLR, 2021.

Angelina Wang and Olga Russakovsky. Overwriting pretrained bias with finetuning data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3957–3968, 2023.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 336–349, 2022a.

Chenyu Wang, Sharut Gupta, Caroline Uhler, and Tommi S. Jaakkola. Removing biases from molecular representations via information maximization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7TOs9gjAg1.

Haohan Wang, Zexue He, Zachary L. Lipton, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=rJEjjoR9K7.

Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 526–536, 2021.

Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding instance-level impact of fairness constraints. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23114–23130. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/wang22ac.html.

Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9322–9331, 2020.

Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proc. of the IEEE/CVF International Conf. on Computer Vision*, pp. 692–702, 2019b.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5190–5203. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/37d097caf1299d9aa79c2c2b843d2d78-Paper.pdf.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5310–5319, 2019c.

Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. A brief review on algorithmic fairness. *Management System Engineering*, 1(1):7, 2022c.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 8919–8928, 2020b.

Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. Algorithmic unfairness through the lens of eu non-discrimination law: Or why the law is not a decision tree. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 805–816, 2023.

Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. The neutrality fallacy: When algorithmic fairness interventions are (not) positive action. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2060–2070, 2024.

Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. Debiased visual question answering from feature and sample perspectives. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Z4ry59PVMq8.

Robert Williamson and Aditya Menon. Fairness risk measures. In *International conference on machine learning*, pp. 6786–6797. PMLR, 2019.

Christopher Winship and Robert D Mare. Models for sample selection bias. *Annual review of sociology*, 18 (1):327–350, 1992.

Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1174–1185, 2023.

Ziwei Wu and Jingrui He. Fairness-aware model-agnostic positive and unlabeled learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1698–1708, 2022.

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: training on synthetic data amplifies bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2113–2147, 2024.

Ruicheng Xian, Lang Yin, and Han Zhao. Fair and Optimal Classification via Post-Processing. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. In the name of fairness: Assessing the bias in clinical record de-identification. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 123–137, 2023.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8cb22bdd0b7ba1ab13d742e22eed8da2-Paper.pdf.

Zikai Xiong, Niccolò Dalmasso, Shubham Sharma, Freddy Lecue, Daniele Magazzeni, Vamsi K Potluru, Tucker Balch, and Manuela Veloso. Fair wasserstein coresets. *arXiv preprint arXiv:2311.05436*, 2023.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, 2018.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1401–1406. IEEE, 2019.

Gezheng Xu, Qi Chen, Charles Ling, Boyu Wang, and Changjian Shui. Intersectional unfairness discovery. *arXiv preprint arXiv:2405.20790*, 2024.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11492–11501. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/xu21b.html.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pp. 11492–11501. PMLR, 2021b.

Jing Xu, Xu Luo, Xinglin Pan, Yanan Li, Wenjie Pei, and Zenglin Xu. Alleviating the sample selection bias in few-shot learning by removing projection to the centroid. *Advances in Neural Information Processing Systems*, 35:21073–21086, 2022a.

Siqi Xu, Lin Liu, and Zhonghua Liu. Deepmed: Semiparametric causal mediation analysis with debiased deep learning. *Advances in Neural Information Processing Systems*, 35:28238–28251, 2022b.

Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 578–586, 2021c.

Yilun Xu, Hao He, Tianxiao Shen, and Tommi S. Jaakkola. Controlling directions orthogonal to a classifier. In *International Conference on Learning Representations*, 2022c. URL https://openreview.net/forum?id=DIjCrlsu6Z.

Yuancheng Xu, Chenghao Deng, Yanchao Sun, Ruijie Zheng, Xiyao Wang, Jieyu Zhao, and Furong Huang. Adapting static fairness to sequential decision-making: Bias mitigation strategies towards equal long-term benefit rate. In *Forty-first International Conference on Machine Learning*.

Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. *arXiv preprint arXiv:2305.16536*, 2023.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet, Timothy J Hazen, and Alessandro Sordoni. Increasing robustness to spurious correlations using forgettable examples. *arXiv preprint arXiv:1911.03861*, 2019.

Tom Yan and Chicheng Zhang. Active fairness auditing. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24929–24962. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/yan22c.html.

Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. Fairness with overlapping groups; a probabilistic perspective. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4067–4078. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/29c0605a3bab4229e46723f89cf59d83-Paper.pdf.

Jingyi Yang, Joel Miller, and Mesrob Ohannessian. Fairness auditing in urban decisions using lp-based data combination. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1817–1825, 2023a.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 547–558, 2020b.

Yifan Yang, Yang Liu, and Parinaz Naghizadeh. Adaptive data debiasing through bounded exploration. *Advances in Neural Information Processing Systems*, 35:1516–1528, 2022.

Zonghan Yang, Xiaoyuan Yi, Peng Li, Yang Liu, and Xing Xie. Unified detoxifying and debiasing in language generation via inference-time adaptive optimization. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=FvevdI0aA_h.

Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/e6384711491713d29bc63fc5eeb5ba4f-Paper.pdf.

Min-Hsuan Yeh, Blossom Metevier, Austin Hoag, and Philip Thomas. Analyzing the relationship between difference and ratio-based fairness metrics. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 518–528, 2024.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Tongxin Yin, Jean-Francois Ton, Ruocheng Guo, Yuanshun Yao, Mingyan Liu, and Yang Liu. Fair classifiers that abstain without harm. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jvveGAbkVx.

Eric Yang Yu, Zhizhen Qin, Min Kyung Lee, and Sicun Gao. Policy optimization with advantage regularization for long-term fairness in decision systems. *arXiv preprint arXiv:2210.12546*, 2022.

Jun Yu, Xinlong Hao, Haonian Xie, and Ye Yu. Fair face recognition using data balancing, enhancement and fusion. In *European Conference on Computer Vision*, pp. 492–505. Springer, 2020.

Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 18–19, 2020.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/zemel13.html.

Boya Zeng, Yida Yin, and Zhuang Liu. Understanding bias in large-scale visual datasets. *arXiv preprint arXiv:2412.01876*, 2024.

Sebastian Zezulka and Konstantin Genin. From the fair distribution of predictions to the fair distribution of social goods: Evaluating the impact of fair machine learning on long-term unemployment. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1984–2006, 2024.

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

Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=woa783QMul.

Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. Fairness reprogramming. *Advances in Neural Information Processing Systems*, 35:34347–34362, 2022a.

Hongjing Zhang and Ian Davidson. Towards fair deep anomaly detection. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 138–148, 2021.

Marilyn Zhang. Affirmative algorithms: Relational equality as algorithmic fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 495–507, 2022.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning*, pp. 26484–26516. PMLR, 2022b.

Xueru Zhang, Mohammad Mahdi Khalili, Kun Jin, Parinaz Naghizadeh, and Mingyan Liu. Fairness interventions as (dis) incentives for strategic manipulation. In *International Conference on Machine Learning*, pp. 26239–26264. PMLR, 2022c.

Yiliang Zhang and Qi Long. Assessing fairness in the presence of missing data. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=myJO35O7Gg.

Zheng Zhang, Wei Song, Qi Liu, Qingyang Mao, Yiyan Wang, Weibo Gao, Zhenya Huang, Shijin Wang, and Enhong Chen. Towards accurate and fair cognitive diagnosis via monotonic data augmentation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818, 2017.

Bowen Zhao, Chen Chen, Qian-Wei Wang, Anfeng He, and Shu-Tao Xia. Combating unknown bias with effective bias-conflicting scoring and gradient alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3561–3569, 2023a.

Dora Zhao, Jerone Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. In *International Conference on Machine Learning*, pp. 42000–42017. PMLR, 2023b.

Dora Zhao, Jerone TA Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. In *International Conference on Machine Learning (ICML)*, 2023c.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1):2527–2552, 2022.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*, 2019a.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pp. 7523–7532. PMLR, 2019b.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkekl0NFPr.

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

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

Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15002–15012, 2021.

Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hanguang Li, and Y. Liu. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:256416558.

Dominik Zietlow, Michal Rolinek, and Georg Martius. Demystifying inductive biases for (beta-) vae based architectures. In *International Conference on Machine Learning*, pp. 12945–12954. PMLR, 2021.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking fairness for medical imaging. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6ve2CkeQe5S.