A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

With the widespread use of artificial intelligence (AI) systems and applications in our everyday lives, accounting for fairness has gained significant importance in designing and engineering of such systems. AI systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that these decisions do not reflect discriminatory behavior toward certain groups or populations. More recently some work has been developed in traditional machine learning and deep learning that address such challenges in different subdomains. With the commercialization of these systems, researchers are becoming more aware of the biases that these applications can contain and are attempting to address them. In this survey, we investigated different real-world applications that have shown biases in various ways, and we listed different sources of biases that can affect AI applications. We then created a taxonomy for fairness definitions that machine learning researchers have defined to avoid the existing bias in AI systems. In addition to that, we examined different domains and subdomains in AI showing what researchers have observed with regard to unfair outcomes in the state-of-the-art methods and ways they have tried to address them. There are still many future directions and solutions that can be taken to mitigate the problem of bias in AI systems. We are hoping that this survey will motivate researchers to tackle these issues in the near future by observing existing work in their respective fields.

CCS Concepts: • **Computing methodologies** → *Artificial intelligence*

Additional Key Words and Phrases: Fairness and bias in artificial intelligence, machine learning, deep learning, natural language processing, representation learning

1 INTRODUCTION

Machine learning algorithms have penetrated every aspect of our lives. Algorithms make movie recommendations, suggest products to buy, and who to date. They are increasingly used in high-stakes scenarios such as loans [109] and hiring decisions [19, 39]. There are clear benefits to algorithmic decision-making; unlike people, machines do not become tired or bored [45, 115], and

they can take into account orders of magnitude more factors than people can. However, like people, algorithms are vulnerable to biases that render their decisions "unfair" [6, 117]. In the context of decision-making, fairness is the *absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics*. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people. A canonical example comes from a tool used by courts in the United States to make pretrial detention and release decisions. The software, **Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)**, measures the risk of a person to recommit another crime. Judges use COMPAS to decide whether to release an offender or to keep him or her in prison. An investigation into the software found a bias against African-Americans[1]: COMPAS is more likely to have higher false positive rates for African-American offenders than Caucasian offenders in falsely predicting them to be at a higher risk of recommitting a crime or recidivism. Similar findings have been made in other areas, such as an AI system that judges beauty pageant winners but was biased against darker-skinned contestants,[2] or facial recognition software in digital cameras that overpredicts Asians as blinking.[3] These biased predictions stem from the hidden or neglected biases in data or algorithms.

In this survey, we identify two potential sources of unfairness in machine learning outcomes—those that arise from biases in the data and those that arise from the algorithms. We review research investigating how biases in data skew what is learned by machine learning algorithms and nuances in the way the algorithms themselves work to prevent them from making fair decisions—even when the data is unbiased. Furthermore, we observe that biased algorithmic outcomes might impact user experience, thus generating a feedback loop between data, algorithms, and users that can perpetuate and even amplify existing sources of bias.

We begin the review with several highly visible real-world cases of where unfair machine learning algorithms have led to suboptimal and discriminatory outcomes in Section 2. In Section 3, we describe the different types and sources of biases that occur within the data-algorithms-users loop mentioned above. Next, in Section 4, we present the different ways that the concept of fairness has been operationalized and studied in the literature. We discuss the ways in which these two concepts are coupled. Last, we will focus on different families of machine learning approaches, how fairness manifests differently in each one, and the current state-of-the-art for tackling them in Section 5, followed by potential areas of future work in each of the domains in Section 6.

2 REAL-WORLD EXAMPLES OF ALGORITHMIC UNFAIRNESS

With the popularity of AI and machine learning over the past decades, and their prolific spread in different applications, safety and fairness constraints have become a significant issue for researchers and engineers. Machine learning is used in courts to assess the probability that a defendant recommits a crime. It is used in different medical fields, in childhood welfare systems [35], and autonomous vehicles. All of these applications have a direct effect in our lives and can harm our society if not designed and engineered correctly, that is, with considerations to fairness. [119] has a list of the applications and the ways these AI systems affect our daily lives with their inherent biases, such as the existence of bias in AI chatbots, employment matching, flight routing, and automated legal aid for immigration algorithms, and search and advertising placement algorithms. Reference [67] discusses examples of how bias in the real world can creep into AI and robotic systems, such as bias in face recognition applications, voice recognition, and search engines. Therefore, it is

---

[1]https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[2]https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.
[3]http://content.time.com/time/business/article/0,8599,1954643,00.html.

important for researchers and engineers to be concerned about the downstream applications and their potential harmful effects when modeling an algorithm or a system.

2.1 Systems that Demonstrate Discrimination

COMPAS is an exemplar of a discriminatory system. In addition to this, discriminatory behavior was also evident in an algorithm that would deliver advertisements promoting jobs in **Science, Technology, Engineering, and Math (STEM)** fields [88]. This advertisement was designed to deliver advertisements in a gender-neutral way. However, less women compared to men saw the advertisement due to gender-imbalance that would result in younger women being considered as a valuable subgroup and more expensive to show advertisements to. This optimization algorithm would deliver ads in a discriminatory way although its original and pure intention was to be gender-neutral. Bias in facial recognition systems [124] and recommender systems [136] have also been largely studied and evaluated and in many cases shown to be discriminative towards certain populations and subgroups. To be able to address the bias issue in these applications, it is important for us to know where these biases are coming from and what we can do to prevent them.

We have enumerated the bias in COMPAS, which is a widely used commercial risk assessment software. In addition to its bias, it also contains performance issues when compared to humans. When compared to non-expert human judgment in a study, it was discovered to be not any better than a normal human [46]. It is also interesting to note that although COMPAS uses 137 features, only 7 of those were presented to the people in the study. Reference [46] further argues that COMPAS is not any better than a simple logistic regression model when making decisions. We should think responsibly and recognize that the application of these tools and their subsequent decisions affect peoples' lives; therefore, considering fairness constraints is a crucial task while designing and engineering these types of sensitive tools. In another similar study, while investigating sources of group unfairness (unfairness across different groups is defined later), the authors in Reference [141] compared SAVRY, a tool used in risk assessment frameworks that includes human intervention in its process, with automatic machine learning methods to see which one is more accurate and more fair. Conducting these types of studies should be done more frequently, but prior to releasing the tools to avoid doing harm.

2.2 Assessment Tools

An interesting direction that researchers have taken is introducing tools that can assess the amount of fairness in a tool or system. For example, Aequitas [132] is a toolkit that lets users to test models with regards to several bias and fairness metrics for different population subgroups. Aequitas produces reports from the obtained data that helps data scientists, machine learning researchers, and policymakers to make conscious decisions and avoid harm and damage toward certain populations. **AI Fairness 360 (AIF360)** is another toolkit developed by IBM to help moving fairness research algorithms into an industrial setting and to create a benchmark for fairness algorithms to get evaluated and an environment for fairness researchers to share their ideas [11]. These types of toolkits can be helpful for learners, researchers, and people working in the industry to move towards developing fair machine learning application away from discriminatory behavior.

3 BIAS IN DATA, ALGORITHMS, AND USER EXPERIENCES

Most AI systems and algorithms are data driven and require data upon which to be trained. Thus, data is tightly coupled to the functionality of these algorithms and systems. In the cases where the underlying training data contains biases, the algorithms trained on them will learn these biases
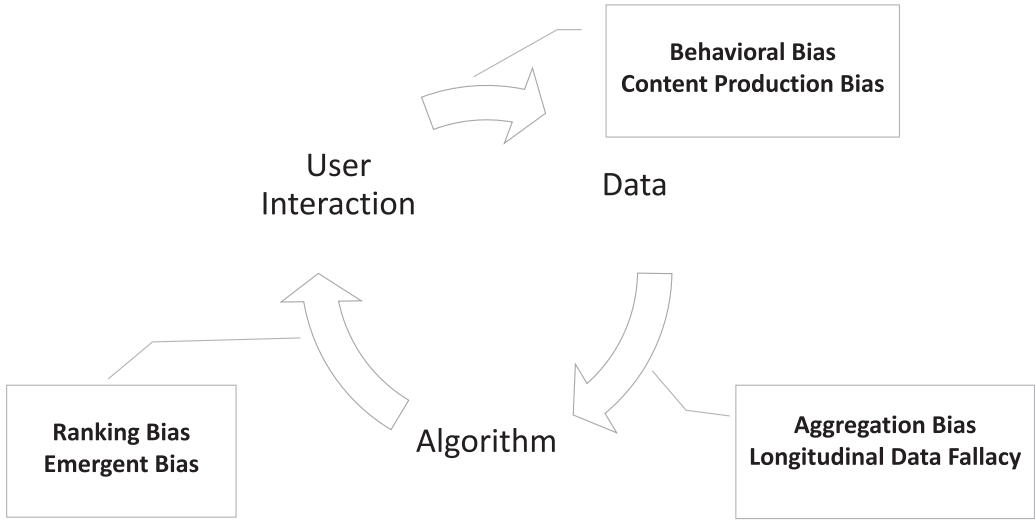
Fig. 1. Examples of bias definitions placed in the data, algorithm, and user interaction feedback loop.

and reflect them into their predictions. As a result, existing biases in data can affect the algorithms using the data, producing biased outcomes. Algorithms can even amplify and perpetuate existing biases in the data. In addition, algorithms themselves can display biased behavior due to certain design choices, even if the data itself is not biased. The outcomes of these biased algorithms can then be fed into real-world systems and affect users' decisions, which will result in more biased data for training future algorithms. For example, imagine a web search engine that puts specific results at the top of its list. Users tend to interact most with the top results and pay little attention to those further down the list [92]. The interactions of users with items will then be collected by the web search engine, and the data will be used to make future decisions on how information should be presented based on popularity and user interest. As a result, results at the top will become more and more popular, not because of the nature of the result but due to the biased interaction and placement of results by these algorithms [92]. The loop capturing this feedback between biases in data, algorithms, and user interaction is illustrated in Figure 1. We use this loop to categorize definitions of bias in the section below.

3.1 Types of Bias

Bias can exist in many shapes and forms, some of which can lead to unfairness in different downstream learning tasks. In Reference [140], authors talk about sources of bias in machine learning with their categorizations and descriptions to motivate future solutions to each of the sources of bias introduced in the paper. In Reference [116], the authors prepare a complete list of different types of biases with their corresponding definitions that exist in different cycles from data origins to its collection and its processing. Here, we will reiterate the most important sources of bias introduced in these two papers and also add in some work from other existing research papers. Additionally, we will introduce a different categorization of these definitions in the paper according to the data, algorithm, and user interaction loop.

   3.1.1 Data to Algorithm. In this section, we talk about biases in data, which, when used by ML training algorithms, might result in biased algorithmic outcomes.

(1) **Measurement Bias.** *Measurement, or reporting, bias arises from how we choose, utilize, and measure particular features* [140]. An example of this type of bias was observed in the recidivism risk prediction tool COMPAS, where prior arrests and friend/family arrests were used as proxy variables to measure level of "riskiness" or "crime"—which on its own can be viewed as mismeasured proxies. This is partly due to the fact that minority communities are controlled and policed more frequently, so they have higher arrest rates. However, one should not conclude that because people coming from minority groups have higher arrest rates, therefore they are more dangerous, as there is a difference in how these groups are assessed and controlled [140].

(2) **Omitted Variable Bias.** *Omitted variable bias[4] occurs when one or more important variables are left out of the model* [38, 110, 127]. An example for this case would be when someone designs a model to predict, with relatively high accuracy, the annual percentage rate at which customers will stop subscribing to a service, but soon observes that the majority of users are canceling their subscription without receiving any warning from the designed model. Now imagine that the reason for canceling the subscriptions is appearance of a new strong competitor in the market that offers the same solution, but for half the price. The appearance of the competitor was something that the model was not ready for; therefore, it is considered to be an omitted variable.

(3) **Representation Bias.** *Representation bias arises from how we sample from a population during data collection process* [140]. Non-representative samples lack the diversity of the population, with missing subgroups and other anomalies. Lack of geographical diversity in datasets like ImageNet (as shown in Figures 3 and 4) results in demonstrable bias towards Western cultures.

(4) **Aggregation Bias.** *Aggregation bias (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population.* An example of this type of bias can be seen in clinical aid tools. Consider diabetes patients who have apparent morbidity differences across ethnicities and genders. Specifically, HbA1c levels, which are widely used to diagnose and monitor diabetes, differ in complex ways across genders and ethnicities. Therefore, a model that ignores individual differences will likely not be well-suited for all ethnic and gender groups in the population [140]. This is true even when they are represented equally in the training data. Any general assumptions about subgroups within the population can result in aggregation bias.

    (a) **Simpson's Paradox.** Simpson's paradox is a type of aggregation bias that arises in the analysis of heterogeneous data [18]. The paradox arises when an association observed in aggregated data disappears or reverses when the same data is disaggregated into its underlying subgroups (Figure 2(a)). One of the better-known examples of the type of paradox arose during the gender bias lawsuit in university admissions against UC Berkeley [16]. After analyzing graduate school admissions data, it seemed like there was bias toward women, a smaller fraction of whom were being admitted to graduate programs compared to their male counterparts. However, when admissions data was separated and analyzed over the departments, women applicants had equality and in some cases even a small advantage over men. The paradox happened as women tended to apply to departments with lower admission rates for both genders. Simpson's paradox has been observed in a variety of domains, including biology [37], psychology [81], astronomy [105], and computational social science [91].

    (b) **Modifiable Areal Unit Problem** is a statistical bias in geospatial analysis, which arises when modeling data at different levels of spatial aggregation [56]. This bias results in different trends learned when data is aggregated at different spatial scales.
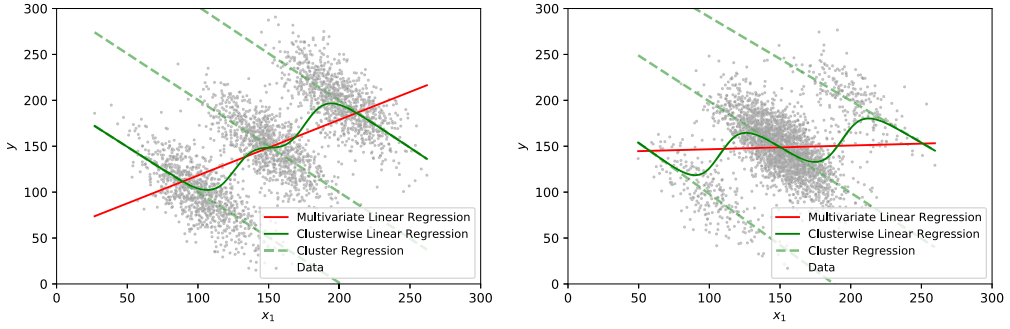
Fig. 2. Illustration of biases in data. The red line shows the regression (MLR) for the entire population, while dashed green lines are regressions for each subgroup, and the solid green line is the unbiased regression. (a) When all subgroups are of equal size, then MLR shows a positive relationship between the outcome and the independent variable. (b) Regression shows almost no relationship in less balanced data. The relationships between variables within each subgroup, however, remain the same. (Credit: Nazanin Alipourfard.)
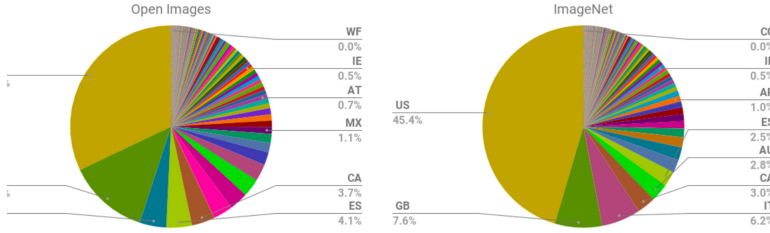
Fig. 3. Fraction of each country, represented by their two-letter ISO codes, in Open Images and ImageNet image datasets. In both datasets, US and Great Britain represent the top locations (from Reference [138], © Shreya Shankar).
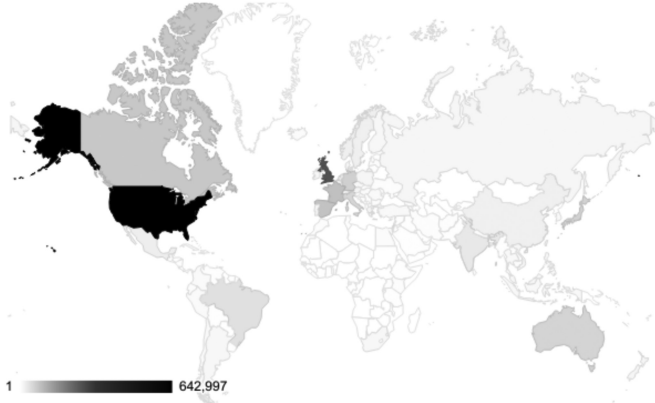
Fig. 4. Geographic distribution of countries in the Open Images dataset. In their sample, almost one third of the data was US-based, and 60% of the data was from the six most represented countries across North America and Europe (from Reference [138], © Shreya Shankar).

(5) **Sampling Bias.** *Sampling bias is similar to representation bias, and it arises due to non-random sampling of subgroups.* As a consequence of sampling bias, the trends estimated for one population may not generalize to data collected from a new population. For the intuition,

consider the example in Figure 2. The left plot represents data collected during a study from three subgroups, which were uniformly sampled (Figure 2(a)). Suppose the next time the study was conducted, one of the subgroups was sampled more frequently than the rest (Figure 2(b)). The positive trend found by the regression model in the first study almost completely disappears (solid red line in plot on the right), although the subgroup trends (dashed green lines) are unaffected.

(6) **Longitudinal Data Fallacy.** Researchers analyzing temporal data must use *longitudinal analysis* to track cohorts over time to learn their behavior. Instead, temporal data is often modeled using cross-sectional analysis, which combines diverse cohorts at a single time point. The heterogeneous cohorts can bias cross-sectional analysis, leading to different conclusions than longitudinal analysis. As an example, analysis of bulk Reddit data [10] revealed that comment length decreased over time on average. However, bulk data represented a cross-sectional snapshot of the population, which in reality contained different cohorts who joined Reddit in different years. When data was disaggregated by cohorts, the comment length within each cohort was found to increase over time.

(7) **Linking Bias.** *Linking bias arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users* [116]. In Reference [101] authors show how social networks can be biased toward low-degree nodes when only considering the links in the network and not considering the content and behavior of users in the network. Reference [149] also shows that user interactions are significantly different from social link patterns that are based on features, such as method of interaction or time. The differences and biases in the networks can be a result of many factors, such as network sampling, as shown in References [59, 107], which can change the network measures and cause different types of problems.

3.1.2 Algorithm to User. Algorithms modulate user behavior. Any biases in algorithms might introduce biases in user behavior. In this section, we talk about biases that are as a result of algorithmic outcomes and affect user behavior as a consequence.

(1) **Algorithmic Bias.** *Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm* [9]. The algorithmic design choices, such as use of certain optimization functions, regularizations, choices in applying regression models on the data as a whole or considering subgroups, and the general use of statistically biased estimators in algorithms [44], can all contribute to biased algorithmic decisions that can bias the outcome of the algorithms.

(2) **User Interaction Bias.** *User Interaction bias is a type of bias that can not only be observant on the Web but also get triggered from two sources—the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction* [9]. This type of bias can be influenced by other types and subtypes, such as presentation and ranking biases.

    (a) **Presentation Bias.** *Presentation bias is a result of how information is presented* [9]. For example, on the Web users can only click on content that they see, so the seen content gets clicks, while everything else gets no click. And it could be the case that the user does not see all the information on the Web [9].

    (b) **Ranking Bias.** *The idea that top-ranked results are the most relevant and important will result in attraction of more clicks than others.* This bias affects search engines [9] and crowdsourcing applications [92].

(3) **Popularity Bias**. *Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation—for example, by fake reviews or social bots* [113]. As an

instance, this type of bias can be seen in search engines [71, 113] or recommendation systems where popular objects would be presented more to the public. But this presentation may not be a result of good quality; instead, it may be due to other biased factors.

(4) **Emergent Bias.** *Emergent bias occurs as a result of use and interaction with real users. This bias arises as a result of change in population, cultural values, or societal knowledge usually some time after the completion of design* [53]. This type of bias is more likely to be observed in user interfaces, since interfaces tend to reflect the capacities, characteristics, and habits of prospective users by design [53]. This type of bias can itself be divided into more subtypes, as discussed in detail in Reference [53].

(5) **Evaluation Bias.** *Evaluation bias happens during model evaluation* [140]. This includes the use of inappropriate and disproportionate benchmarks for evaluation of applications such as Adience and IJB-A benchmarks. These benchmarks are used in the evaluation of facial recognition systems that were biased toward skin color and gender [24] and can serve as examples for this type of bias [140].

3.1.3 User to Data. Many data sources used for training ML models are user-generated. Any inherent biases in users might be reflected in the data they generate. Furthermore, when user behavior is affected/modulated by an algorithm, any biases present in those algorithm might introduce bias in the data generation process. Here, we list several important types of such biases.

(1) **Historical Bias.** *Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection* [140]. An example of this type of bias can be found in a 2018 image search result where searching for women CEOs ultimately resulted in fewer female CEO images due to the fact that only 5% of Fortune 500 CEOs were women—which would cause the search results to be biased towards male CEOs [140]. These search results were of course reflecting the reality, but whether or not the search algorithms should reflect this reality is an issue worth considering.

(2) **Population Bias.** *Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population of the platform from the original target population* [116]. Population bias creates non-representative data. An example of this type of bias can arise from different user demographics on different social platforms, such as women being more likely to use Pinterest, Facebook, Instagram, while men being more active in online forums like Reddit or Twitter. More such examples and statistics related to social media use among young adults according to gender, race, ethnicity, and parental educational background can be found in Reference [64].

(3) **Self-selection Bias.** *Self-selection bias[4] is a subtype of the selection or sampling bias in which subjects of the research select themselves.* An example of this type of bias can be observed in an opinion poll to measure enthusiasm for a political candidate, where the most enthusiastic supporters are more likely to complete the poll.

(4) **Social Bias.** *Social bias happens when others' actions affect our judgment* [9]. An example of this type of bias can be a case where we want to rate or review an item with a low score, but when influenced by other high ratings, we change our scoring thinking that perhaps we are being too harsh [9, 147].

(5) **Behavioral Bias.** *Behavioral bias arises from different user behavior across platforms, contexts, or different datasets* [116]. An example of this type of bias can be observed in Reference [104], where authors show how differences in emoji representations among platforms can result in

---

[4]https://data36.com/statistical-bias-types-explained/.

different reactions and behavior from people and sometimes even leading to communication errors.

(6) **Temporal Bias.** *Temporal bias arises from differences in populations and behaviors over time* [116]. An example can be observed in Twitter where people talking about a particular topic start using a hashtag at some point to capture attention, then continue the discussion about the event without using the hashtag [116, 142].

(7) **Content Production Bias.** *Content Production bias arises from structural, lexical, semantic, and syntactic differences in the contents generated by users* [116]. An example of this type of bias can be seen in Reference [114] where the differences in use of language across different gender and age groups is discussed. The differences in use of language can also be seen across and within countries and populations.

Existing work tries to categorize these bias definitions into groups, such as definitions falling solely under data or user interaction. However, due to the existence of the feedback loop phenomenon [36], these definitions are intertwined, and we need a categorization that closely models this situation. This feedback loop is not only existent between the data and the algorithm, but also between the algorithms and user interaction [29]. Inspired by these papers, we modeled categorization of bias definitions, as shown in Figure 1, and grouped these definitions on the arrows of the loop where we thought they were most effective. We emphasize the fact again that these definitions are intertwined, and one should consider how they affect each other in this cycle and address them accordingly.

3.2 Data Bias Examples

There are multiple ways that discriminatory bias can seep into data. For instance, using unbalanced data can create biases against underrepresented groups. Reference [166] analyzes some examples of the biases that can exist in the data and algorithms and offers some recommendations and suggestions toward mitigating these issues.

3.2.1 Examples of Bias in Machine Learning Data. In Reference [24], the authors show that datasets such as IJB-A and Adience are imbalanced and contain mainly light-skinned subjects— 79.6% in IJB-A and 86.2% in Adience. This can bias the analysis towards dark-skinned groups who are underrepresented in the data. In another instance, the way we use and analyze our data can create bias when we do not consider different subgroups in the data. In Reference [24], the authors also show that considering only male-female groups is not enough, but there is also a need to use race to further subdivide the gender groups into light-skinned females, light-skinned males, dark-skinned males, and dark-skinned females. It is only in this case that we can clearly observe the bias towards dark-skinned females, as previously dark-skinned males would compromise for dark-skinned females and would hide the underlying bias towards this subgroup. Popular machine-learning datasets that serve as a base for most of the developed algorithms and tools can also be biased—which can be harmful to the downstream applications that are based on these datasets. For instance, ImageNet [131] and Open Images [86] are two widely used datasets in machine learning. In Reference [138], researchers showed that these datasets suffer from representation bias and advocate for the need to incorporate geographic diversity and inclusion while creating such datasets.

3.2.2 Examples of Data Bias in Medical Applications. These data biases can be more dangerous in other sensitive applications. For example, in medical domains there are many instances in which the data studied and used are skewed toward certain populations—which can have dangerous consequences for the underrepresented communities. Reference [97] showed how exclusion of African-Americans resulted in their misclassification in clinical studies, so they became advocates

for sequencing the genomes of diverse populations in the data to prevent harm to underrepresented populations. Authors in Reference [139] studied the 23andMe genotype dataset and found that out of 2,399 individuals, who have openly shared their genotypes in public repositories, 2,098 (87%) are European, while only 58 (2%) are Asian and 50 (2%) African. Other such studies were conducted in Reference [54], which states that UK Biobank, a large and widely used genetic dataset, may not represent the sampling population. Researchers found evidence of a "healthy volunteer" selection bias. Reference [146] has other examples of studies on existing biases in the data used in the medical domain. Reference [153] also looks at machine learning algorithms and data utilized in medical fields and writes about how artificial intelligence in health care has not impacted all patients equally.

3.3 Discrimination

Similar to bias, discrimination is also a source of unfairness. Discrimination can be considered as a source for unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally, while bias can be considered as a source for unfairness that is due to the data collection, sampling, and measurement. Although bias can also be seen as a source of unfairness that is due to human prejudice and stereotyping, in the algorithmic fairness literature it is more intuitive to categorize them as such according to the existing research in these areas. In this survey, we mainly focus on concepts that are relevant to algorithmic fairness issues. References [98, 129, 148] contain more broad information on discrimination theory that involve more multidisciplinary concepts from legal theory, economics, and social sciences, which can be referenced by the interested readers.

3.3.1 Explainable Discrimination. Differences in treatment and outcomes among different groups can be justified and explained via some attributes in some cases. In situations where these differences are justified and explained, it is not considered to be illegal discrimination and hence called explainable [77]. For instance, authors in Reference [77] state that in the UCI Adult dataset [7], a widely used dataset in the fairness domain, males on average have a higher annual income than females. However, this is because on average females work fewer hours than males per week. Work hours per week is an attribute that can be used to explain low income that needs to be considered. If we make decisions, without considering working hours, such that males and females end up averaging the same income, then we will lead to reverse discrimination, since we would cause male employees to get lower salary than females. Therefore, explainable discrimination is acceptable and legal, as it can be explained through other attributes like working hours. In Reference [77], authors present a methodology to quantify the explainable and illegal discrimination in data. They argue that methods that do not take the explainable part of the discrimination into account may result in non-desirable outcomes, so they introduce a *reverse* discrimination, which is equally harmful and undesirable. They explain how to quantify and measure discrimination in data or a classifier's decisions, which directly considers illegal and explainable discrimination.

3.3.2 Unexplainable Discrimination. In contrast to explainable discrimination, there is unexplainable discrimination in which the discrimination toward a group is unjustified and therefore considered illegal. Authors in Reference [77] also present local techniques for removing only the illegal or unexplainable discrimination, allowing only for explainable differences in decisions. These are preprocessing techniques that change the training data such that it contains no unexplainable discrimination. We expect classifiers trained on this preprocessed data to not capture illegal or unexplainable discrimination. Unexplainable discrimination consists of *direct* and *indirect* discrimination.

(1) **Direct Discrimination.** Direct discrimination happens when protected attributes of individuals explicitly result in non-favorable outcomes toward them [160]. Typically, there are some traits identified by law on which it is illegal to discriminate against, and it is usually these traits that are considered to be "protected" or "sensitive" attributes in computer science literature. A list of some of these protected attributes is provided in Table 3 as specified in the **Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA)** [30].

(2) **Indirect Discrimination.** In indirect discrimination, individuals appear to be treated based on seemingly neutral and non-protected attributes; however, protected groups, or individuals, still get to be treated unjustly as a result of implicit effects from their protected attributes; e.g., the residential zip code of a person can be used in decision-making processes such as loan applications. However, this can still lead to racial discrimination, such as redlining, as despite the fact that zip code appears to be a non-sensitive attribute, it may correlate with race because of the population of residential areas [126, 160].

3.3.3 Sources of Discrimination.

(1) **Systemic Discrimination.** Systemic discrimination refers to policies, customs, or behaviors that are a part of the culture or structure of an organization that may perpetuate discrimination against certain subgroups of the population [40]. Reference [128] found that employers overwhelmingly preferred competent candidates that were culturally similar to them and shared similar experiences and hobbies. If the decision-makers happen to belong overwhelmingly to certain subgroups, then this may result in discrimination against competent candidates that do not belong to these subgroups.

(2) **Statistical Discrimination.** Statistical discrimination is a phenomenon where decision-makers use average group statistics to judge an individual belonging to that group. It usually occurs when the decision-makers (e.g., employers, or law enforcement officers) use an individual's obvious, recognizable characteristics as a proxy for either hidden or more-difficult-to-determine characteristics, which may actually be relevant to the outcome [120].

4 ALGORITHMIC FAIRNESS

Fighting against bias and discrimination has a long history in philosophy and psychology and, recently, in machine learning. However, to be able to fight against discrimination and achieve fairness, one should first define fairness. Philosophy and psychology have tried to define the concept of fairness long before computer science. The fact that no universal definition of fairness exists shows the difficulty of solving this problem [134]. Different preferences and outlooks in different cultures lend a preference to different ways of looking at fairness, which makes it harder to come up with just a single definition that is acceptable to everyone in a situation. Indeed, even in computer science, where most of the work on proposing new fairness constraints for algorithms has come from the West, and a lot of these papers use the same datasets and problems to show how their constraints perform, there is still no clear agreement on which constraints are the most appropriate for those problems. Broadly, fairness is the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making [135]. Even though fairness is an incredibly desirable quality in society, it can be surprisingly difficult to achieve in practice. With these challenges in mind, many fairness definitions are proposed to address different algorithmic bias and discrimination issues discussed in the previous section.

4.1 Definitions of Fairness

In Reference [17], authors studied fairness definitions in political philosophy and tried to tie them to machine learning. Authors in Reference [70] studied the 50-year history of fairness definitions

in the areas of education and machine learning. In Reference [145], authors listed and explained some of the definitions used for fairness in algorithmic classification problems. In Reference [135], authors studied the general public's perception of some of these fairness definitions in computer science literature. Here, we will reiterate and provide some of the most widely used definitions, along with their explanations inspired from Reference [145].

Definition 1 (Equalized Odds). The definition of equalized odds, provided by Reference [63], states that "A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y. $P(\hat{Y}=1|A=0,Y=y) = P(\hat{Y}=1|A=1,Y=y)$, $y \in \{0,1\}$." This means that the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected group members [145]. In other words, the equalized odds definition states that the protected and unprotected groups should have equal rates for true positives and false positives.

Definition 2 (Equal Opportunity). "A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if $P(\hat{Y}=1|A=0,Y=1) = P(\hat{Y}=1|A=1,Y=1)$" [63]. This means that the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected (female and male) group members [145]. In other words, the equal opportunity definition states that the protected and unprotected groups should have equal true positive rates.

Definition 3 (Demographic Parity). Also known as statistical parity. "A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$" [48, 87]. The likelihood of a positive outcome [145] should be the same regardless of whether the person is in the protected (e.g., female) group.

Definition 4 (Fairness through Awareness). "An algorithm is fair if it gives similar predictions to similar individuals" [48, 87]. In other words, any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome.

Definition 5 (Fairness through Unawareness). "An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process" [61, 87].

Definition 6 (Treatment Equality). "Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories" [15].

Definition 7 (Test Fairness). "A score S = S(x) is test fair (well-calibrated) if it reflects the same likelihood of recidivism irrespective of the individual's group membership, R. That is, if for all values of s, $P(Y=1|S=s,R=b)=P(Y=1|S=s,R=w)$" [34]. In other words, the test fairness definition states that for any predicted probability score S, people in both protected and unprotected groups must have equal probability of correctly belonging to the positive class [145].

Definition 8 (Counterfactual Fairness). "Predictor $\hat{Y}$ is counterfactually fair if under any context X =x and A=a, $P(\hat{Y}_{A \leftarrow a}(U)=y|X=x,A=a)=P(\hat{Y}_{A \leftarrow a'}(U)=y|X=x,A=a)$, (for all y and for any value a' attainable by A" [87]. The counterfactual fairness definition is based on the "intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group."

Definition 9 (Fairness in Relational Domains). "A notion of fairness that is able to capture the relational structure in a domain—not only by taking attributes of individuals into consideration but by taking into account the social, organizational, and other connections between individuals" [50].

Table 1. Categorizing Different Fairness Notions into Group, Subgroup, and Individual Types

Name	Reference	Group	Subgroup	Individual
Demographic parity	[48, 87]	✓		
Conditional statistical parity	[41]	✓		
Equalized odds	[63]	✓		
Equal opportunity	[63]	✓		
Treatment equality	[15]	✓		
Test fairness	[34]	✓		
Subgroup fairness	[79, 80]		✓	
Fairness through unawareness	[61, 87]			✓
Fairness through awareness	[48]			✓
Counterfactual fairness	[87]			✓

Definition 10 (Conditional Statistical Parity). For a set of legitimate factors L, predictor $\hat{Y}$ satisfies conditional statistical parity if $P(\hat{Y}|L=1,A=0) = P(\hat{Y}|L=1,A=1)$ [41]. Conditional statistical parity states that people in both protected and unprotected (female and male) groups should have equal probability of being assigned to a positive outcome given a set of legitimate factors L [145].

Fairness definitions fall under different types as shown in Table 1 as follows:

(1) **Individual Fairness.** Give similar predictions to similar individuals [48, 87].
(2) **Group Fairness.** Treat different groups equally [48, 87].
(3) **Subgroup Fairness**. Subgroup fairness intends to obtain the best properties of the group and individual notions of fairness. It is different than these notions but uses them to obtain better outcomes. It picks a group fairness constraint like equalizing false positive and asks whether this constraint holds over a large collection of subgroups [79, 80].

It is important to note that, according to Reference [83], it is impossible to satisfy some of the fairness constraints at once except in highly constrained special cases. In Reference [83], the authors show the inherent incompatibility of two conditions: calibration and balancing the positive and negative classes. These cannot be satisfied simultaneously with each other unless under certain constraints; therefore, it is important to take the context and application in which fairness definitions need to be used into consideration and use them accordingly [137]. Another important aspect to consider is time and temporal analysis of the impacts that these definitions may have on individuals or groups. In Reference [94] authors show that current fairness definitions are not always helpful and do not promote improvement for sensitive groups—and can actually be harmful when analyzed over time in some cases. They also show that measurement errors can also act in favor of these fairness definitions; therefore, they show how temporal modeling and measurement are important in evaluation of fairness criteria and introduce a new range of tradeoffs and challenges toward this direction. It is also important to pay attention to the sources of bias and their types when trying to solve fairness-related questions.

5 METHODS FOR FAIR MACHINE LEARNING

There have been numerous attempts to address bias in artificial intelligence to achieve fairness; these stem from domains of AI. In this section, we will enumerate different domains of AI and the work that has been produced by each community to combat bias and unfairness in their methods. Table 2 provides an overview of the different areas that we focus upon in this survey.

Table 2. List of Papers Targeting and Talking about Bias and Fairness in Different Areas

Area	Reference(s)
Classification	[25, 49, 57, 63, 69, 73, 75, 78, 85, 102, 118, 143, 150, 151, 155]
Regression	[1, 14]
PCA	[133]
Community detection	[101]
Clustering	[8, 31]
Graph embedding	[22]
Causal inference	[82, 95, 111, 112, 123, 156, 160, 161]
Variational auto encoders	[5, 42, 96, 108]
Adversarial learning	[90, 152]
Word embedding	[20, 58, 165] [23, 162]
Coreference resolution	[130, 164]
Language model	[21]
Sentence embedding	[99]
Machine translation	[52]
Semantic role labeling	[163]
Named Entity Recognition	[100]

While this section is largely domain-specific, it can be useful to take a cross-domain view. Generally, methods that target biases in the algorithms fall under three categories:

(1) **Pre-processing.** Pre-processing techniques try to transform the data so the underlying discrimination is removed [43]. If the algorithm is allowed to modify the training data, then pre-processing can be used [11].

(2) **In-processing.** In-processing techniques try to modify and change state-of-the-art learning algorithms to remove discrimination during the model training process [43]. If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used during the training of a model— either by incorporating changes into the objective function or imposing a constraint [11, 14].

(3) **Post-processing.** Post-processing is performed after training by accessing a holdout set that was not involved during the training of the model [43]. If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used in which the labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase [11, 14].

Examples of some existing work and their categorization into these types are shown in Table 4. These methods are not just limited to general machine learning techniques, but because of AI's popularity, they have expanded to different domains such as natural language processing and deep learning. From learning fair representations [42, 96, 108] to learning fair word embeddings [20, 58, 165], debiasing methods have been proposed in different AI applications and domains. Most of these methods try to avoid unethical interference of sensitive or protected attributes into the decision-making process, while others target exclusion bias by trying to include users from sensitive groups. In addition, some works try to satisfy one or more of the fairness notions in their methods, such as **disparate learning processes (DLPs)** that try to satisfy notions of treatment disparity and impact disparity by allowing the protected attributes during the training phase but avoiding them during prediction time [93]. A list of protected or sensitive attributes is provided in Table 3. They point out what attributes should not affect the outcome of the decision in housing

Table 3. A List of the Protected Attributes as Specified in the Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA), from Reference [30]

Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

loan or credit card decision-making [30] according to the law. Some of the existing work tries to treat sensitive attributes as noise to disregard their effect on decision-making, while some causal methods use causal graphs and disregard some paths in the causal graph that result in sensitive attributes affecting the outcome of the decision. Different bias-mitigating methods and techniques are discussed below for different domains—each targeting a different problem in different areas of machine learning in detail. This can expand the horizon of the reader on where and how bias can affect the system and try to help researchers carefully look at various new problems concerning potential places where discrimination and bias can affect the outcome of a system.

5.1 Unbiasing Data

Every dataset is the result of several design decisions made by the data curator. Those decisions have consequences for the fairness of the resulting dataset, which in turn affects the resulting algorithms. To mitigate the effects of bias in data, some general methods have been proposed that advocate having good practices while using data, such as having datasheets that would act like a supporting document for the data reporting the dataset creation method, its characteristics, motivations, and its skews [13, 55]. Reference [12] proposes a similar approach for the NLP applications. A similar suggestion has been proposed for models in Reference [106]. Authors in Reference [66] also propose having labels, just like nutrition labels on food, to better categorize each data for each task. In addition to these general techniques, some work has targeted more specific types of biases. For example, Reference [81] has proposed methods to test for cases of Simpson's paradox in the data, and References [3, 4] proposed methods to discover Simpson's paradoxes in data automatically. Causal models and graphs were also used in some work to detect direct discrimination in the data along with its prevention technique that modifies the data such that the predictions would be absent from direct discrimination [159]. Reference [62] also worked on preventing discrimination in data mining, targeting direct, indirect, and simultaneous effects. Other pre-processing approaches, such as messaging [74], preferential sampling [75, 76], disparate impact removal [51], also aim to remove biases from the data.

5.2 Fair Machine Learning

To address this issue, a variety of methods have been proposed that satisfy some of the fairness definitions or other new definitions depending on the application.

Table 4. Algorithms Categorized into Their Appropriate Groups Based on Being Pre-processing,
In-processing, or Post-processing

Algorithm	Reference	Pre-processing	In-processing	Post-processing
Community detection	[101]	✓		
Word embedding	[23]	✓		
Optimized pre-processing	[27]	✓		
Data pre-processing	[76]	✓		
Classification	[155]		✓	
Regression	[14]		✓	
Classification	[78]		✓	
Classification	[151]		✓	
Adversarial learning	[90]		✓	
Classification	[63]			✓
Word embedding	[20]			✓
Classification	[121]			✓

5.2.1 Fair Classification. Since classification is a canonical task in machine learning and is widely used in different areas that can be in direct contact with humans, it is important that these types of methods be fair and be absent from biases that can harm some populations. Therefore, certain methods have been proposed [57, 78, 85, 102] that satisfy certain definitions of fairness in classification. For instance, in Reference [143] authors try to satisfy subgroup fairness in classification, equality of opportunity, and equalized odds in Reference [63], both disparate treatment and disparate impact in References [2, 155], and equalized odds in Reference [150]. Other methods try to not only satisfy some fairness constraints but to also be stable toward change in the test set [69]. The authors in Reference [151] propose a general framework for learning fair classifiers. This framework can be used for formulating fairness-aware classification with fairness guarantees. In another work [25], authors propose three different modifications to the existing Naive Bayes classifier for discrimination-free classification. Reference [118] takes a new approach into fair classification by imposing fairness constraints into a **Multitask learning (MTL)** framework. In addition to imposing fairness during training, this approach can benefit the minority groups by focusing on maximizing the average accuracy of each group as opposed to maximizing the accuracy as a whole without attention to accuracy across different groups. In a similar work [49], authors propose a decoupled classification system where a separate classifier is learned for each group. They use transfer learning to reduce the issue of having less data for minority groups. In Reference [73] authors propose to achieve fair classification by mitigating the dependence of the classification outcome on the sensitive attributes by utilizing the Wasserstein distance measure. In Reference [75] authors propose the **Preferential Sampling (PS)** method to create a discrimination-free training dataset. They then learn a classifier on this discrimination free dataset to have a classifier with no discrimination.

5.2.2 Fair Regression. Reference [14] proposes a fair regression method along with evaluating it with a measure introduced as the **"price of fairness" (POF)** to measure accuracy-fairness tradeoffs. They introduce three fairness penalties as follows:

Individual Fairness: The definition for individual fairness, as stated in Reference [14], "for every cross pair $(x, y) \in S_1$, $(x', y') \in S_2$, a model w is penalized for how differently it treats x and x' (weighted by a function of $|y - y'|$) where S_1 and S_2 are different groups from the sampled

population." Formally, this is operationalized as

$$f_1(w, S) = \frac{1}{n_1 n_2} \sum_{\substack{(x_i, y_i) \in S_1 \\ (x_j, y_j) \in S_2}} d(y_i, y_j)(w.x_i - w.x_j)^2.$$

Group Fairness: "On average, the two groups' instances should have similar labels (weighted by the nearness of the labels of the instances)" [14]:

$$f_2(w, S) = \left(\frac{1}{n_1 n_2} \sum_{\substack{(x_i, y_i) \in S_1 \\ (x_j, y_j) \in S_2}} d(y_i, y_j)(w.x_i - w.x_j) \right)^2.$$

Hybrid Fairness: "Hybrid fairness requires both positive and both negatively labeled cross pairs to be treated similarly in an average over the two groups" [14]:

$$f_3(w, S) = \left(\sum_{\substack{(x_i, y_i) \in S_1 \\ (x_j, y_j) \in S_2 \\ y_i = y_j = 1}} \frac{d(y_i, y_j)(w.x_i - w.x_j)}{n_{1,1} n_{2,1}} \right)^2 + \left(\sum_{\substack{(x_i, y_i) \in S_1 \\ (x_j, y_j) \in S_2 \\ y_i = y_j = -1}} \frac{d(y_i, y_j)(w.x_i - w.x_j)}{n_{1,-1} n_{2,-1}} \right)^2.$$

In addition to the previous work, Reference [1] considers the fair regression problem formulation with regards to two notions of fairness statistical (demographic) parity and bounded group loss. Reference [2] uses decision trees to satisfy disparate impact and treatment in regression tasks in addition to classification.

5.2.3 Structured Prediction. In Reference [163], authors studied the semantic role-labeling models and a famous dataset, imSitu, and realized that only 33% of agent roles in cooking images are man, and the rest of 67% cooking images have woman as agents in the imSitu training set. They also noticed that in addition to the existing bias in the dataset, the model would amplify the bias such that after training a model[5] on the dataset, bias is magnified for "man," filling only 16% of cooking images. Under these observations, the authors of Reference [163] show that structured prediction models have the risk of leveraging social bias. Therefore, they propose a calibration algorithm called **RBA (reducing bias amplification)**; RBA is a technique for debiasing models by calibrating prediction in structured prediction. The idea behind RBA is to ensure that the model predictions follow the same distribution in the training data. They study two cases: multi-label object and visual semantic role labeling classification. They show how these methods amplify the existing bias in data.

5.2.4 Fair PCA. In Reference [133] authors show that vanilla PCA can exaggerate the error in reconstruction in one group of people over a different group of equal size, so they propose a fair method to create representations with similar richness for different populations—not to make them indistinguishable or to hide dependence on a sensitive or protected attribute. They show that vanilla PCA on the **labeled faces in the wild (LFW)** dataset [68] has a lower reconstruction error rate for men than for women faces, even if the sampling is done with an equal weight for both genders. They intend to introduce a dimensionality reduction technique that maintains similar fidelity for different groups and populations in the dataset. Therefore, they introduce Fair PCA and define a fair dimensionality reduction algorithm. Their definition of Fair PCA (as an

[5]Specifically, a Conditional Random Field (CRF).

optimization function) is as follows, in which A and B denote two subgroups, U_A and U_B denote matrices whose rows correspond to rows of U that contain members of subgroups A and B given m data points in R^n:

$$min_{U \in R^{m \times n}, rank(U) \leq d} \ max \left\{ \frac{1}{|A|} loss(A, U_A), \frac{1}{|B|} loss(B, U_B) \right\}.$$

And their proposed algorithm is a two-step process listed below:

(1) Relax the Fair PCA objective to a **semidefinite program (SDP)** and solve it.
(2) Solve a linear program that would reduce the rank of the solution.

5.2.5 Community Detection/Graph Embedding/Clustering. Inequalities in online communities and social networks can also potentially be another place where bias and discrimination can affect the populations. For example, in online communities, users with a fewer number of friends or followers face a disadvantage of being heard in online social media [101]. In addition, existing methods, such as community detection methods, can amplify this bias by ignoring these low-connected users in the network or by wrongfully assigning them to the irrelevant and small communities. In Reference [101] authors show how this type of bias exists and is perpetuated by the existing community detection methods. They propose a new attributed community detection method, called CLAN, to mitigate the harm toward disadvantaged groups in online social communities. CLAN is a two-step process that considers the network structure alongside node attributes to address exclusion bias, as indicated below:

(1) Detect communities using modularity values (Step 1-unsupervised using only network structure).
(2) Train a classifier to classify users in the minor groups, putting them into one of the major groups using held-out node attributes (Step 2-supervised using other node attributes).

Fair methods in domains similar to community detection are also proposed, such as graph embedding [22] and clustering [8, 31].

5.2.6 Causal Approach to Fairness. Causal models can ascertain causal relationships between variables. Using causal graphs, one can represent these causal relationships between variables (nodes of the graph) through the edges of the graph. These models can be used to remove unwanted causal dependence of outcomes on sensitive attributes such as gender or race in designing systems or policies [95]. Many researchers have used causal models and graphs to solve fairness-related concerns in machine learning. In References [33, 95], authors discuss in detail the subject of causality and its importance while designing fair algorithms. There has been much research on discrimination discovery and removal that uses causal models and graphs to make decisions that are irrespective of sensitive attributes of groups or individuals. For instance, in Reference [160] authors propose a causal-based framework that detects direct and indirect discrimination in the data along with their removal techniques. Reference [161] is an extension to the previous work. Reference [156] gives a nice overview of most of the previous work done in this area by the authors, along with discussing system-, group-, and individual-level discrimination and solving each using their previous methods, in addition to targeting direct and indirect discrimination. By expanding on the previous work and generalizing it, authors in Reference [112] propose a similar pathway approach for fair inference using causal graphs; this would restrict certain problematic and discriminative pathways in the causal graph flexibly given any set of constraints. This holds when the path-specific effects can be identified from the observed distribution. In Reference [32] authors introduce the path-specific counterfactual fairness definition, which is an extension to

counterfactual fairness definition [87], and propose a method to achieve it further extending the work in Reference [112]. In Reference [111] authors extended a formalization of algorithmic fairness from their previous work to the setting of learning optimal policies that are subject to constraints based on definitions of fairness. They describe several strategies for learning optimal policies by modifying some of the existing strategies, such as Q-learning, value search, and G-estimation, based on some fairness considerations. In Reference [158] authors only target discrimination discovery and no removal by finding instances similar to another instance and observing if a change in the protected attribute will change the outcome of the decision. If so, then they declare the existence of discrimination. In Reference [82], authors define the following two notions of discrimination—unresolved discrimination and proxy discrimination—as follows:

Unresolved Discrimination: "A variable V in a causal graph exhibits unresolved discrimination if there exists a directed path from A to V that is not blocked by a resolving variable, and V itself is non-resolving" [82].

Proxy Discrimination: "A variable V in a causal graph exhibits potential proxy discrimination, if there exists a directed path from A to V that is blocked by a proxy variable and V itself is not a proxy" [82]. They proposed methods to prevent and avoid them. They also show that no observational criterion can determine whether a predictor exhibits unresolved discrimination; therefore, a causal reasoning framework needs to be incorporated.

In Reference [123], instead of using the usual risk difference $RD = p_1 - p_2$, authors propose a causal risk difference $RD^c = p_1 - p_2^c$ for causal discrimination discovery. They define p_2^c to be:

$$p_2^c = \frac{\sum_{\mathbf{s} \in S, dec(\mathbf{s}) = \ominus} w(\mathbf{s})}{\sum_{\mathbf{s} \in S} w(\mathbf{s})}.$$

RD^c not close to zero means that there is a bias in decision value due to group membership (causal discrimination) or to covariates that have not been accounted for in the analysis (omitted variable bias). This RD^c then becomes their causal discrimination measure for discrimination discovery. Reference [157] is another work of this type that uses causal networks for discrimination discovery.

5.3 Fair Representation Learning

5.3.1 Variational Auto Encoders. Learning fair representations and avoiding the unfair interference of sensitive attributes has been introduced in many different research papers. A well-known example is the Variational Fair Autoencoder, introduced in Reference [96]. Here, they treat the sensitive variable as the nuisance variable, so by removing the information about this variable they will get a fair representation. They use a maximum mean discrepancy regularizer to obtain invariance in the posterior distribution over latent variables. Adding this **maximum mean discrepancy (MMD)** penalty into the lower bound of their VAE architecture satisfies their proposed model for having the Variational Fair Autoencoder. Similar work, but not targeting fairness specifically, has been introduced in Reference [72]. In Reference [5] authors also propose a debiased VAE architecture called DB-VAE, which learns sensitive latent variables that can bias the model (e.g., skin tone, gender) and propose an algorithm on top of this DB-VAE using these latent variables to debias systems like facial detection systems. In Reference [108] authors model their representation-learning task as an optimization objective that would minimize the loss of the mutual information between the encoding and the sensitive variable. The relaxed version of this assumption is shown in Equation (1). They use this to learn fair representation and show that adversarial training is unnecessary and in some cases even counter-productive. In Equation (1), c is the sensitive variable

and z the encoding of x.

$$\min_q \mathcal{L}(q, x) + \lambda I(z, c) \tag{1}$$

In Reference [42], authors introduce flexibly fair representation learning by disentanglement that disentangles information from multiple sensitive attributes. Their flexible and fair variational autoencoder is not only flexible with respect to downstream task labels, but also flexible with respect to sensitive attributes. They address the demographic parity notion of fairness, which can target multiple sensitive attributes or any subset combination of them.

5.3.2 Adversarial Learning. In Reference [90] authors present a framework to mitigate bias in models learned from data with stereotypical associations. They propose a model in which they are trying to maximize accuracy of the predictor on y, and at the same time minimize the ability of the adversary to predict the protected or sensitive variable (stereotyping variable z). The model consists of two parts—the predictor and the adversary—as shown in Figure 6. In their model, the predictor is trained to predict Y given X. With the help of a gradient-based approach like stochastic gradient descent, the model tries to learn the weights W by minimizing some loss function LP($\hat{y}$, y). The output layer is passed to an adversary, which is another network. This network tries to predict Z. The adversary may have different inputs depending on the fairness definition needing to be achieved. For instance, to satisfy **Demographic Parity**, the adversary would try to predict the protected variable Z using only the predicted label $\hat{Y}$ passed as an input to it, while preventing the adversary from learning this is the goal of the predictor. Similarly, to achieve **Equality of Odds**, the adversary would get the true label Y in addition to the predicted label $\hat{Y}$. To satisfy **Equality of Opportunity** for a given class y, they would only select instances for the adversary where Y=y. Reference [152] takes an interesting and different direction toward solving fairness issues using adversarial networks by introducing FairGAN, which generates synthetic data that is free from discrimination and is similar to the real data. They use their newly generated synthetic data from FairGAN, which is now debiased, instead of the real data for training and testing. They do not try to remove discrimination from the dataset, unlike many of the existing approaches, but instead generate new datasets similar to the real one that is debiased and preserves good data utility. The architecture of their FairGAN model is shown in Figure 5. FairGAN consists of two components: a generator G_{Dec} that generates the fake data conditioned on the protected attribute $P_G(x, y, s) = P_G(x, y|s)P_G(s)$ where $P_G(s) = P_{data}(s)$, and two discriminators D_1 and D_2. D_1 is trained to differentiate the real data denoted by $P_{data}(x, y, s)$ from the generated fake data denoted by $P_G(x, y, s)$.

In addition to that, for achieving fairness constraints, such as statistical parity, $P_G(x, y|s = 1) = P_G(x, y|s = 0)$, the training of D_2 is such that it emphasizes differentiation of the two types of synthetic (generated by the model) samples $P_G(x, y|s = 1)$ and $P_G(x, y|s = 0)$ indicating if the synthetic samples are from the unprotected or protected groups. Here, s denotes the protected or the sensitive variable, and we adapted the same notation as in Reference [152].

5.4 Fair NLP

5.4.1 Word Embedding. In Reference [20] authors noticed that while using state-of-the-art word embeddings in word analogy tests, "man" would be mapped to "computer programmer" and "woman" would be mapped to "homemaker." This bias toward woman triggered the authors to propose a method to debias word embeddings by proposing a method that respects the embeddings for gender-specific words but debiases embeddings for gender-neutral words by following these steps *(Notice that Step 2 has two different options. Depending on whether you target hard debiasing or soft debiasing, you would use either step 2a or 2b):*
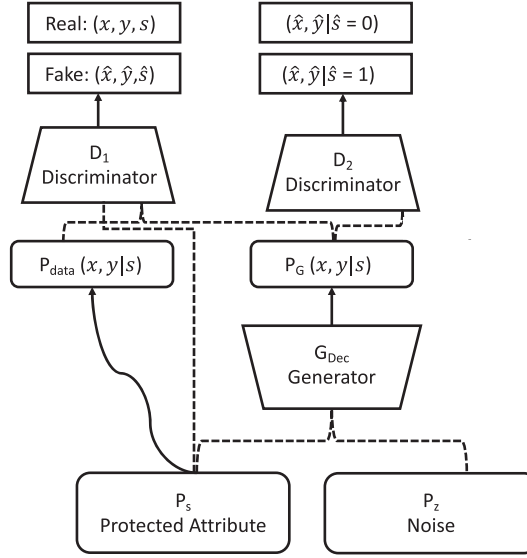
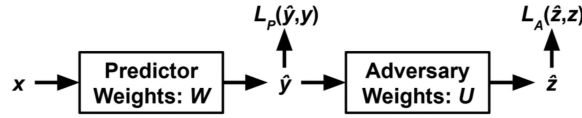Fig. 5. Structure of FairGAN as proposed in Reference [152].

Fig. 6. The architecture of adversarial network proposed in Reference [90] ($^{©}$ Brian Hu Zhang).

(1) **Identify gender subspace.** Identifying a direction of the embedding that captures the bias [20].
(2) **Hard debiasing or soft debiasing:**
    (a) **Hard debiasing (neutralize and equalize).** Neutralize puts away the gender subspace from gender-neutral words and makes sure that all the gender-neutral words are removed and zeroed out in the gender subspace [20]. Equalize makes gender-neutral words to be equidistant from the equality set of gendered words [20].
    (b) **Soft bias correction.** Tries to move as little as possible to retain its similarity to the original embedding as much as possible, while reducing the gender bias. This tradeoff is controlled by a parameter [20].

Following on the footsteps of these authors, other future work attempted to tackle this problem [165] by generating a gender-neutral version of (GloVe called GN-GloVe) that tries to retain gender information in some of the word embedding's learned dimensions, while ensuring that other dimensions are free from this gender effect. This approach primarily relies on GloVe as its base model with gender as the protected attribute. However, a recent paper [58] argues against these debiasing techniques and states that many recent works on debiasing word embeddings have been superficial, that those techniques just hide the bias and do not actually remove it. A recent work [23] took a new direction and proposed a preprocessing method for the discovery of the problematic documents in the training corpus that have biases in them and tried to debias the system by perturbing or removing these documents efficiently from the training corpus. In a very recent work [162], authors target bias in ELMo's contextualized word vectors and attempt to analyze

and mitigate the observed bias in the embeddings. They show that the corpus used for training of ELMo has a significant gender skew, with male entities being nearly three times more common than female entities. This automatically leads to gender bias in these pretrained contextualized embeddings. They propose the following two methods for mitigating the existing bias while using the pretrained embeddings in a downstream task, coreference resolution: (1) train-time data augmentation approach, and (2) test-time neutralization approach.

5.4.2 Coreference Resolution. Reference [164] shows that coreference systems have a gender bias. They introduce a benchmark, called WinoBias, focusing on gender bias in coreference resolution. In addition to that, they introduce a data-augmentation technique that removes bias in the existing state-of-the-art coreferencing methods, in combination with using word2vec debiasing techniques. Their general approach is as follows: They first generate auxiliary datasets using a rule-based approach in which they replace all the male entities with female entities and the other way around. Then they train models with a combination of the original and the auxiliary datasets. They use the above solution in combination with word2vec debiasing techniques to generate word embeddings. They also point out sources of gender bias in coreference systems and propose solutions to them. They show that the first source of bias comes from the training data and propose a solution that generates an auxiliary dataset by swapping male and female entities. Another case arises from the resource bias (word embeddings are bias), so the proposed solution is to replace GloVe with a debiased embedding method. Last, another source of bias can come from unbalanced gender lists, and balancing the counts in the lists is a solution they proposed. In another work [130], authors also show the existence of gender bias in three state-of-the-art coreference resolution systems by observing that for many occupations, these systems resolve pronouns in a biased fashion by preferring one gender over the other.

5.4.3 Language Model. In Reference [21] authors introduce a metric for measuring gender bias in a generated text from a language model based on recurrent neural networks that is trained on a text corpus along with measuring the bias in the training text itself. They use Equation (2), where w is any word in the corpus, f is a set of gendered words that belong to the female category, such as she, her, woman, and so on, and m to the male category, and measure the bias using the mean absolute and standard deviation of the proposed metric along with fitting a univariate linear regression model over it and then analyzing the effectiveness of each of those metrics while measuring the bias.

$$bias(w) = log\left(\frac{P(w|f)}{P(w|m)}\right) \qquad (2)$$

In their language model, they also introduce a regularization loss term that would minimize the projection of embeddings trained by the encoder onto the embedding of the gender subspace following the soft debiasing technique introduced in Reference [20]. Finally, they evaluate the effectiveness of their method on reducing gender bias and conclude by stating that to reduce bias, there is a compromise on perplexity. They also point out the effectiveness of word-level bias metrics over the corpus-level metrics.

5.4.4 Sentence Encoder. In Reference [99] authors extend the research in detecting bias in word embedding techniques to that of sentence embedding. They try to generalize bias-measuring techniques, such as using the **Word Embedding Association Test (WEAT)** [26] in the context of sentence encoders by introducing their new sentence encoding bias-measuring techniques, the **Sentence Encoder Association Test (SEAT)**. They used state-of-the-art sentence encoding techniques, such as CBoW, GPT, ELMo, and BERT, and find that although there was varying evidence of human-like bias in sentence encoders using SEAT, more recent methods like BERT are more

immune to biases. That being said, they are not claiming that these models are bias-free, but state that more sophisticated bias discovery techniques may be used in these cases, thereby encouraging more future work in this area.

5.4.5 Machine Translation. In Reference [52] authors noticed that when translating the word "friend" in the following two sentences from English to Spanish, they achieved different results—although in both cases this word should be translated the same way.
"She works in a hospital, my friend is a nurse."
"She works in a hospital, my friend is a doctor."
In both of these sentences, "friend" should be translated to the female version of Spanish friend "amiga," but the results were not reflecting this expectation. For the second sentence, friend was translated to "amigo,"—the male version of friend in Spanish. This is because doctor is more stereotypical to males and nurse to females, and the model picks this bias or stereotype and reflects it in its performance. To solve this, authors in Reference [52] build an approach that leverages the fact that machine translation uses word embeddings. They use the existing debiasing methods in word embedding and apply them in the machine translation pipeline. This not only helped them to mitigate the existing bias in their system, but also boosted the performance of their system by one BLUE score. In Reference [122] authors show that Google's translate system can suffer from gender bias by making sentences taken from the U.S. Bureau of Labor Statistics into a dozen languages that are gender-neutral, including Yoruba, Hungarian, and Chinese, translating them into English, and showing that Google Translate shows favoritism toward males for stereotypical fields such as STEM jobs. In Reference [144] authors annotated and analyzed the Europarl dataset [84], a large political, multilingual dataset used in machine translation, and discovered that with the exception of the youngest age group (20–30), which represents only a very small percentage of the total amount of sentences (0.71%), more male data is available in all age groups. They also looked at the entire dataset and showed that 67.39% of the sentences are produced by male speakers. Furthermore, to mitigate the gender-related issues and to improve morphological agreement in machine translation, they augmented every sentence with a tag on the English source side, identifying the gender of the speaker. This helped the system in most of the cases, but not always, so further work has been suggested for integrating speaker information in other ways.

5.4.6 Named Entity Recognition. In Reference [100], authors investigate a type of existing bias in various **named entity recognition (NER)** systems. In particular, they observed that in a context where an entity should be tagged as a person entity, such as "John is a person" or "John is going to school," more female names as opposed to male names are being tagged as non-person entities or not being tagged at all. To further formalize their observations, authors propose six different evaluation metrics that would measure amount of bias among different genders in NER systems. They curated templated sentences pertaining to human actions and applied these metrics on names from U.S census data incorporated into the templates. The six introduced measures each aim to demonstrate a certain type of bias and serve a specific purpose in showing various results as follows:

- Error Type-1 Unweighted: Through this type of error, authors wanted to recognize the proportion of entities that are tagged as anything other than the person entity in each of the male vs. female demographic groups. This could be the entity not being tagged or tagged as other entities, such as location.

$$\frac{\sum_{n \in N_f} I(n_{type} \neq PERSON)}{|N_f|}$$

- Error Type-1 Weighted: This type of error is similar to its unweighted case except authors considered the frequency or popularity of names so they could penalize if a more popular name is being tagged wrongfully.

$$\frac{\sum_{n \in N_f} freq_f(n_{type} \neq PERSON)}{\sum_{n \in N_f} freq_f(n)},$$

where $freq_f(\cdot)$ indicates the frequency of a name for a particular year in the female census data. Likewise, $freq_m(\cdot)$ indicates the frequency of a name for a particular year in the male census data.

- Error Type-2 Unweighted: This is a type of error in which the entity is tagged as other entities, such as location or city. Notice that this error does not count if the entity is not tagged.

$$\frac{\sum_{n \in N_f} I(n_{type} \notin \{\emptyset, PERSON\})}{|N_f|},$$

where $\emptyset$ indicates that the name is not tagged.

- Error Type-2 Weighted: This error is again similar to its unweighted case except the frequency is taken into consideration.

$$\frac{\sum_{n \in N_f} freq_f(n_{type} \notin \{\emptyset, PERSON\})}{\sum_{n \in N_f} freq_f(n)}$$

- Error Type-3 Unweighted: This is a type of error in which it reports if the entity is not tagged at all. Notice that even if the entity is tagged as a non-person entity this error type would not consider it.

$$\frac{\sum_{n \in N_f} I(n_{type} = \emptyset)}{|N_f|}$$

- Error Type-3 Weighted: Again, this error is similar to its unweighted case with frequency taken into consideration.

$$\frac{\sum_{n \in N_f} freq_f(n_{type} = \emptyset)}{\sum_{n \in N_f} freq_f(n)}$$

Authors also investigate the data that these NER systems are trained on and find that the data is also biased toward female gender by not including as versatile names as there should be to represent female names.

5.5 Comparison of Different Mitigation Algorithms

The field of algorithmic fairness is a relatively new area of research and work still needs to be done for its improvement. With that being said, there are already papers that propose fair AI algorithms and bias mitigation techniques and compare different mitigation algorithms using different benchmark datasets in the fairness domain. For instance, authors in Reference [65] propose a geometric solution to learn fair representations that removes correlation between protected and unprotected features. The proposed approach can control the tradeoff between fairness and accuracy via an adjustable parameter. In this work, authors evaluate the performance of their approach on different benchmark datasets, such as COMPAS, Adult, and German, and compare them against various different approaches for fair learning algorithms considering fairness and accuracy measures [65, 72, 154, 155]. In addition, IBM's **AI Fairness 360 (AIF360)** toolkit [11] has implemented many of the current fair learning algorithms and has demonstrated some of the results as demos

that can be utilized by interested users to compare different methods with regards to different fairness measures.

6 CHALLENGES AND OPPORTUNITIES FOR FAIRNESS RESEARCH

While there have been many definitions of, and approaches to, fairness in the literature, the study in this area is anything but complete. Fairness and algorithmic bias still holds a number of research opportunities. In this section, we provide pointers to outstanding challenges in fairness research and an overview of opportunities for development of understudied problems.

6.1 Challenges

There are several remaining challenges to be addressed in the fairness literature. Among them are:

(1) **Synthesizing a definition of fairness.** Several definitions of what would constitute fairness from a machine learning perspective have been proposed in the literature. These definitions cover a wide range of use cases and as a result are somewhat disparate in their view of fairness. Because of this, it is nearly impossible to understand how one fairness solution would fare under a different definition of fairness. Synthesizing these definitions into one remains an open research problem, since it can make evaluation of these systems more unified and comparable. Having a more unified fairness definition and framework can also help with the incompatibility issue of some current fairness definitions.

(2) **From Equality to Equity.** The definitions presented in the literature mostly focus on *equality*, ensuring that each individual or group is given the same amount of resources, attention, or outcome. However, little attention has been paid to *equity*, which is the concept that each individual or group is given the resources they need to succeed [60]. Operationalizing this definition and studying how it augments or contradicts existing definitions of fairness remains an exciting future direction.

(3) **Searching for Unfairness.** Given a definition of fairness, it should be possible to identify instances of this unfairness in a particular dataset. Inroads toward this problem have been made in the areas of data bias by detecting instances of Simpson's Paradox in arbitrary datasets [3]; however, unfairness may require more consideration due to the variety of definitions and the nuances in detecting each one.

6.2 Opportunities

In this work, we have taxonomized and summarized the current state of research into algorithmic biases and fairness—with a particular focus on machine learning. Even in this area alone, the research is broad. Subareas, from natural language processing, to representation learning, to community detection, have all seen efforts to make their methodologies more fair. Nevertheless, every area has not received the same amount of attention from the research community. Figure 7 provides an overview of what has been done in different areas to address fairness—categorized by the fairness definition type and domain. Some areas (e.g., community detection at the subgroup level) have received no attention in the literature and could be fertile future research areas.

7 CONCLUSION

In this survey, we introduced problems that can adversely affect AI systems in terms of bias and unfairness. The issues were viewed primarily from two dimensions: data and algorithms. We illustrated problems that demonstrate why fairness is an important issue. We further showed examples of the potential real-world harm that unfairness can have on society—such as applications in judicial systems, face recognition, and promoting algorithms. We then went over the definitions of
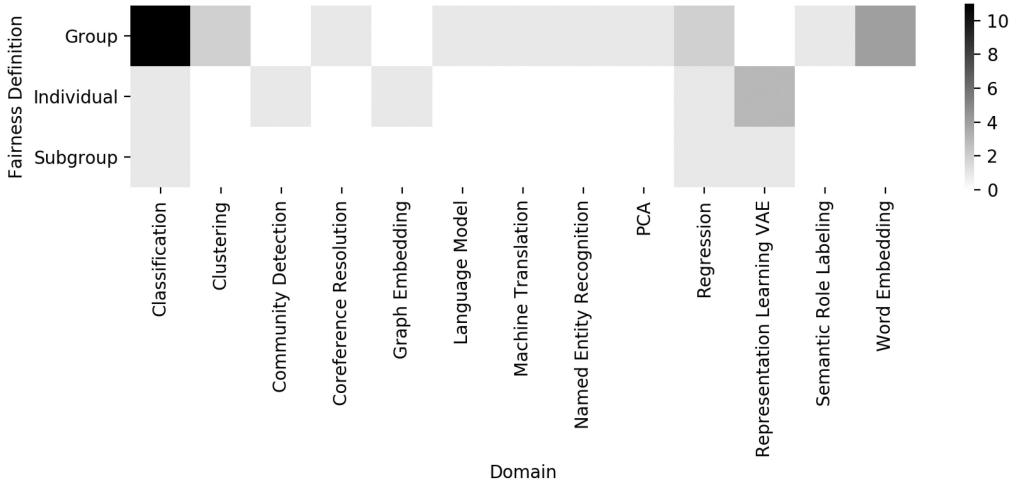
Fig. 7. Heatmap depicting distribution of previous work in fairness, grouped by domain and fairness definition.

fairness and bias that have been proposed by researchers. To further stimulate the interest of readers, we provided some of the work done in different areas in terms of addressing the biases that may affect AI systems and different methods and domains in AI, such as general machine learning, deep learning, and natural language processing. We then further subdivided the fields into a more fine-grained analysis of each subdomain and the work being done to address fairness constraints in each. The hope is to expand the horizons of the readers to think deeply while working on a system or a method to ensure that it has a low likelihood of causing potential harm or bias toward a particular group. With the expansion of AI use in our world, it is important that researchers take this issue seriously and expand their knowledge in this field. In this survey, we categorized and created a taxonomy of what has been done so far to address different issues in different domains regarding the fairness issue. Other possible future work and directions can be taken to address the existing problems and biases in AI that we discussed in the previous sections.

A APPENDIX

A.1 Datasets for Fairness Research

Aside from the existence of bias in datasets, there are datasets (some shown in Table 5) that are specifically used to address bias and fairness issues in machine learning. There are also some datasets that are introduced to target the issues and biases previously observed in older existing datasets. Below, we list some of the widely known datasets that have the characteristics discussed in this survey.

A.1.1 UCI Adult Dataset. UCI Adult dataset, also known as "Census Income" dataset, contains information, extracted from the 1994 census data about people with attributes such as age, occupation, education, race, sex, marital status, native country, hours per week, and so on, indicating whether the income of a person exceeds $50K/yr or not. It can be used in fairness-related studies that want to compare gender or race inequalities based on people's annual incomes, or various other studies [7].

Table 5. Most Widely Used Datasets in the Fairness Domain with Additional Information about Each of the Datasets Including Their Size and Area of Concentration

Dataset Name	Reference	Size	Area
UCI adult dataset	[7]	48,842 income records	Social
German credit dataset	[47]	1,000 credit records	Financial
Pilot parliaments benchmark dataset	[24]	1,270 images	Facial images
WinoBias	[164]	3,160 sentences	Coreference resolution
Communities and crime dataset	[125]	1,994 crime records	Social
COMPAS Dataset	[89]	18,610 crime records	Social
Recidivism in juvenile justice dataset	[28]	4,753 crime records	Social
Diversity in faces dataset	[103]	1 million images	Facial images

A.1.2 German Credit Dataset. The German Credit dataset contains 1,000 credit records containing attributes such as personal status and sex, credit score, credit amount, housing status and so on. It can be used in studies about gender inequalities on credit-related issues [47].

A.1.3 WinoBias. The WinoBias dataset follows the winograd format and has 40 occupations in sentences that are referenced to human pronouns. There are two types of challenge sentences in the dataset requiring linkage of gendered pronouns to either male or female stereotypical occupations. It was used in the coreference resolution study to certify if a system has gender bias or not—in this case, towards stereotypical occupations [164].

A.1.4 Communities and Crime Dataset. The Communities and Crime dataset gathers information from different communities in the United States related to several factors that can highly influence some common crimes such as robberies, murders, or rapes. The data includes crime data obtained from the 1990 US LEMAS survey and the 1995 FBI Unified Crime Report. It also contains socio-economic data from the 1990 US Census.

A.1.5 COMPAS Dataset. The COMPAS dataset contains records for defendants from Broward County indicating their jail and prison times, demographics, criminal histories, and COMPAS risk scores from 2013 to 2014 [89].

A.1.6 Recidivism in Juvenile Justice Dataset. The Recidivism in Juvenile Justice dataset contains all juvenile offenders between ages 12–17 who committed a crime between years 2002 and 2010 and completed a prison sentence in 2010 in Catalonia's juvenile justice system [141].

A.1.7 Pilot Parliaments Benchmark Dataset. The Pilot Parliaments Benchmark dataset, also known as PPB, contains images of 1,270 individuals in the national parliaments from three European (Iceland, Finland, Sweden) and three African (Rwanda, Senegal, South Africa) countries. This benchmark was released to have more gender and race balance, diversity, and representativeness [24].

A.1.8 Diversity in Faces Dataset. The **Diversity in Faces (DiF)** is an image dataset collected for fairness research in face recognition. DiF is a large dataset containing one million annotations for face images. It is also a diverse dataset with diverse facial features, such as different craniofacial distances, skin color, facial symmetry and contrast, age, pose, gender, resolution, along with diverse areas and ratios [103].

ACKNOWLEDGMENTS

REFERENCES

[1] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *Proceedings of the International Conference on Machine Learning*. 120–129.

[2] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1418–1426.

[3] Nazanin Alipourfard, Peter G. Fennell, and Kristina Lerman. 2018. Can you trust the trend? Discovering Simpson's paradoxes in social data. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. ACM, 19–27.

[4] Nazanin Alipourfard, Peter G. Fennell, and Kristina Lerman. 2018. Using Simpson's paradox to discover interesting patterns in behavioral data. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*.

[5] Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, USA, 289–295. https://doi.org/10.1145/3306618.3314243

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2019. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] A. Asuncion and D. J. Newman. 2007. UCI Machine Learning Repository. Retrieved from http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html.

[8] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. 2019. Scalable fair clustering. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 405–413. Retrieved from http://proceedings.mlr.press/v97/backurs19a.html.

[9] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (May 2018), 54–61. DOI:https://doi.org/10.1145/3209581

[10] Samuel Barbosa, Dan Cosley, Amit Sharma, and Roberto M Cesar Jr. 2016. Averaging gone wrong: Using time-aware analyses to better understand behavior. In *Proceedings of the 25th International Conference on World Wide Web*. 829–841.

[11] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic et al. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[12] Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Ling.* 6 (2018), 587–604. DOI:https://doi.org/10.1162/tacl_a_00041

[13] Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. 2019. Towards standardization of data licenses: The Montreal data license.

[14] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. arXiv:cs.LG/1706.02409 (2017).

[15] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2021), 3–44. https://doi.org/10.1177/0049124118782533

[16] Peter J. Bickel, Eugene A. Hammel, and J. William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187, 4175 (1975), 398–404.

[17] R. D. P. Binns. 2018. Fairness in machine learning: Lessons from political philosophy. *J. Mach. Learn. Res.* (2018).

[18] Colin R. Blyth. 1972. On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Assoc.* 67, 338 (1972), 364–366.

[19] Miranda Bogen and Aaron Rieke. 2018. *Help Wanted: An Examination of Hiring Algorithms, Equity and Bias*. Technical Report. Upturn.

[20] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 4349–4357.

[21] Shikha Bordia and Samuel Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 7–15.

[22] Avishek Bose and William Hamilton. 2019. Compositional fairness constraints for graph embeddings. In *Proceedings of the International Conference on Machine Learning*. 715–724.

[23] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 803–811. Retrieved from http://proceedings.mlr.press/v97/brunet19a.html.

[24] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, 77–91. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html.

[25] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining Knowl. Discov.* 21, 2 (2010), 277–292.

[26] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[27] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001. Retrieved from http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf.

[28] Manel Capdevila, Marta Ferrer, and Eulália Luque. 2005. La reincidencia en el delito en la justicia de menores. *Centro de Estudios jurídicos y formación especializada, Generalitat de Catalunya. Documento no publicado* (2005).

[29] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 224–232.

[30] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 339–348.

[31] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. 2019. Proportionally fair clustering. In *Proceedings of the International Conference on Machine Learning*. 1032–1041.

[32] S. Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. 7801–7808.

[33] S. Chiappa and W. S. Isaac. 2019. A causal Bayesian networks viewpoint on fairness. In *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data. Privacy and Identity 2018. IFIP Advances in Information and Communication Technology*, E. Kosta, J. Pierson, D. Slamanig, S. Fischer-Hübner, S. Krenn (Eds.,) Vol. 547. Springer, Cham.

[34] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[35] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, 134–148. Retrieved from http://proceedings.mlr.press/v81/chouldechova18a.html.

[36] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[37] John S. Chuang, Olivier Rivoire, and Stanislas Leibler. 2009. Simpson's paradox in a synthetic microbial system. *Science* 323, 5911 (2009), 272–275.

[38] Kevin A. Clarke. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict Manag. Peace Sci.* 22, 4 (2005), 341–352.

[39] Lee Cohen, Zachary C. Lipton, and Yishay Mansour. 2019. Efficient candidate screening under multiple tests and implications for fairness. arXiv:cs.LG/1905.11361 (2019).

[40] United States. Equal Employment Opportunity Commission. 1987. *EEOC Compliance Manual*. U.S. Equal Employment Opportunity Commission, Washington, D.C..

[41] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.

[42] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *Proceedings of the International Conference on Machine Learning*. 1436–1445.

[43] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data* 5, 2 (2017), 120–134.

[44] David Danks and Alex John London. 2017. Algorithmic bias in autonomous systems. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 4691–4697.

[45] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proc. Nat. Acad. Sci.* 108, 17 (2011), 6889–6892.

[46] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* 4, 1 (2018). DOI:https://doi.org/10.1126/sciadv.aao5580

[47] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml.

[48] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS'12)*. ACM, New York, NY, 214–226. DOI:https://doi.org/10.1145/2090236.2090255

[49] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, 119–133. Retrieved from http://proceedings.mlr.press/v81/dwork18a.html.

[50] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. 2018. Fairness in relational domains. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*. ACM, New York, NY, 108–114. DOI:https://doi.org/10.1145/3278721.3278733

[51] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. Association for Computing Machinery, New York, NY, 259–268. DOI:https://doi.org/10.1145/2783258.2783311

[52] Joel Escudé Font and Marta R. Costa-Jussà. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116* (2019).

[53] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. DOI:https://doi.org/10.1145/230538.230561

[54] Anna Fry, Thomas J. Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E. Allen. 2017. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Amer. J. Epidem.* 186, 9 (06 2017), 1026–1034. DOI:https://doi.org/10.1093/aje/kwx246

[55] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets.

[56] C. E. Gehlke and Katherine Biehl. 1934. Certain effects of grouping upon the size of the correlation coefficient in census tract material. *J. Amer. Statist. Assoc.* 29, 185A (1934), 169–170. DOI:https://doi.org/10.2307/2277827

[57] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[58] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).

[59] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014. Assessing the bias in samples of large online networks. *Soc. Netw.* 38 (2014), 16–27.

[60] Susan T. Gooden. 2015. *Race and Social Equity: A Nervous Area of Government*. Routledge.

[61] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *Proceedings of the NIPS Symposium on Machine Learning and the Law*, Vol. 1. 2.

[62] S. Hajian and J. Domingo-Ferrer. 2013. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. Knowl. Data Eng.* 25, 7 (July 2013), 1445–1459. DOI:https://doi.org/10.1109/TKDE.2012.72

[63] Moritz Hardt, Eric Price, Nati Srebro et al. 2016. Equality of opportunity in supervised learning. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 3315–3323.

[64] Eszter Hargittai. 2007. Whose space? Differences among users and non-users of social network sites. *J. Comput.-mediat. Commun.* 13, 1 (10 2007), 276–297. DOI:https://doi.org/10.1111/j.1083-6101.2007.00396.x

[65] Yuzi He, Keith Burghardt, and Kristina Lerman. 2020. A geometric solution to fair representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 279–285.

[66] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).

[67] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Sci. Eng. Ethics* 24, 5 (2018), 1521–1536.

[68] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Marseille, France.* https://hal.inria.fr/inria-00321923.

[69] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and fair classification. In *Proceedings of the International Conference on Machine Learning.* 2879–2890.

[70] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of test (Un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, 49–58.

[71] L. Introna and H. Nissenbaum. 2000. Defining the Web: The politics of search engines. *Computer* 33, 1 (Jan. 2000), 54–62. DOI:https://doi.org/10.1109/2.816269

[72] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2018. Unsupervised Adversarial Invariance. arXiv:cs.LG/1809.10083 (2018).

[73] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence.* PMLR, 862–872.

[74] F. Kamiran and T. Calders. 2009. Classifying without discriminating. In *Proceedings of the 2nd International Conference on Computer, Control and Communication.* 1–6. DOI:https://doi.org/10.1109/IC4.2009.4909197

[75] Faisal Kamiran and Toon Calders. 2010. Classification with no discrimination by preferential sampling. In *Proceedings of the 19th Machine Learning Conference.* Citeseer, 1–6.

[76] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33, 1 (01 Oct. 2012), 1–33. DOI:https://doi.org/10.1007/s10115-011-0463-8

[77] Faisal Kamiran and Indrė Žliobaitė. 2013. *Explainable and Non-explainable Discrimination in Classification.* Springer Berlin, 155–170. DOI:https://doi.org/10.1007/978-3-642-30487-3_8

[78] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 35–50.

[79] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the International Conference on Machine Learning.* 2569–2577.

[80] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, 100–109.

[81] Rogier Kievit, Willem Eduard Frankenhuis, Lourens Waldorp, and Denny Borsboom. 2013. Simpson's paradox in psychological science: A practical guide. *Front. Psychol.* 4 (2013), 513.

[82] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems.* 656–666.

[83] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[84] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*, Vol. 5. 79–86.

[85] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the World Wide Web Conference (WWW'18).* International World Wide Web Conferences Steering Committee, 853–862. DOI:https://doi.org/10.1145/3178876.3186133

[86] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit et al. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. Retrieved from https://github.com/openimages.

[87] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4066–4076. Retrieved from http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf.

[88] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science* 65, 7 (2019), 2966–2981.

[89] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. 2016. Compas analysis. Retrieved from https://github.com/propublica/compas-analysis.

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

[91] Kristina Lerman. 2018. Computational social scientist beware: Simpson's paradox in behavioral data. *J. Comput. Soc. Sci.* 1, 1 (2018), 49–58.

[92] Kristina Lerman and Tad Hogg. 2014. Leveraging position bias to improve peer recommendation. *PLoS One* 9, 6 (2014), e98914. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.009891%4.

[93] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. 2017. Does mitigating ML's disparate impact require disparate treatment? *stat* 1050 (2017), 19.

[94] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning*.

[95] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).

[96] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2016. The variational fair autoencoder. *stat* 1050 (2016), 4.

[97] Arjun K. Manrai, Birgit H. Funke, Heidi L. Rehm, Morten S. Olesen, Bradley A. Maron, Peter Szolovits, David M. Margulies, Joseph Loscalzo, and Isaac S. Kohane. 2016. Genetic misdiagnoses and the potential for health disparities. *New Eng. J. Med.* 375, 7 (2016), 655–665. DOI:https://doi.org/10.1056/NEJMsa1507092

[98] Ray Marshall. 1974. The economics of racial discrimination: A survey. *J. Econ. Lit.* 12, 3 (1974), 849–871.

[99] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561* (2019).

[100] Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Man is to person as woman is to location: Measuring gender bias in named entity recognition. *arXiv preprint arXiv:1910.10872* (2019).

[101] Ninareh Mehrabi, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Debiasing community detection: The importance of lowly-connected nodes. *arXiv preprint arXiv:1903.08136* (2019).

[102] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, 107–118. Retrieved from http://proceedings.mlr.press/v81/menon18a.html.

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

[104] Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "Blissfully Happy" or "Ready to Fight": Varying interpretations of emoji. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*.

[105] I. Minchev, G. Matijevic, D. W. Hogg, G. Guiglion, M. Steinmetz, F. Anders, C. Chiappini, M. Martig, A. Queiroz, and C. Scannapieco. 2019. Yule-Simpson's paradox in galactic archaeology. *arXiv preprint arXiv:1902.01421* (2019).

[106] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'19)*. ACM, New York, NY, 220–229. DOI:https://doi.org/10.1145/3287560.3287596

[107] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*. AAAI Press.

[108] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. 2018. Invariant representations without adversarial training. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 9084–9093.

[109] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P. Mathur. 2002. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *Int. Trans. Oper. Res.* 9, 5 (2002), 583–597.

[110] David B. Mustard. 2003. Reexamining criminal behavior: The importance of omitted variable bias. *Rev. Econ. Statist.* 85, 1 (2003), 205–211.

[111] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2018. Learning optimal fair policies. *arXiv preprint arXiv:1809.02244* (2018).

[112] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[113] Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Filippo Menczer, and Alessandro Flammini. 2017. How algorithmic popularity bias hinders or promotes quality. *arXiv preprint arXiv:1707.00574* (2017).

[114] Dong-Phuong Nguyen, Rilana Gravel, Rudolf Berend Trieschnigg, and Theo Meder. 2013. "How old do you think I am?": A study of language and age in Twitter. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM '13)*. AAAI Press, 439–448.

[115] Anne O'Keeffe and Michael McCarthy. 2010. *The Routledge Handbook of Corpus Linguistics*. Routledge.

[116] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[117] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY.

[118] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. 2019. Taking advantage of multitask learning for fair classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 227–237.

[119] Osonde A. Osoba and William Welser IV. 2017. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Rand Corporation.

[120] Edmund S. Phelps. 1972. The statistical theory of racism and sexism. *Amer. Econ. Rev.* 62, 4 (1972), 659–661.

[121] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5680–5689. Retrieved from http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf.

[122] Marcelo O. R. Prates, Pedro H. Avelar, and Luis C. Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications* (2019), 1–19.

[123] Bilal Qureshi, Faisal Kamiran, Asim Karim, and Salvatore Ruggieri. 2016. Causal discrimination discovery through propensity score analysis. *arXiv preprint arXiv:1608.03735* (2016).

[124] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, USA, 429–435. https://doi.org/10.1145/3306618.3314244

[125] M. Redmond. 2011. Communities and crime unnormalized dataset. *UCI Machine Learning Repository*: Retrieved from http://www.ics.uci.edu/mlearn/MLRepository.html.

[126] Willy E. Rice. 1996. Race, gender, redlining, and the discriminatory access to loans, credit, and insurance: An historical and empirical analysis of consumers who sued lenders and insurers in federal and state courts, 1950–1995. *San Diego L. Rev.* 33 (1996), 583.

[127] Stephanie K. Riegg. 2008. Causal inference and omitted variable bias in financial aid research: Assessing solutions. *Rev. Higher Educ.* 31, 3 (2008), 329–354.

[128] Lauren A. Rivera. 2012. Hiring as cultural matching: The case of elite professional service firms. *Amer. Sociol. Rev.* 77, 6 (2012), 999–1022.

[129] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

[130] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 8–14. DOI:https://doi.org/10.18653/v1/N18-2002

[131] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein et al. 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252.

[132] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[133] Samira Samadi, Uthaipon Tantipongpipat, Jamie Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The price of fair PCA: One extra dimension. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., 10999–11010. Retrieved from http://dl.acm.org/citation.cfm?id=3327546.3327755.

[134] Nripsuta Ani Saxena. 2019. Perceptions of fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. ACM, New York, NY, 537–538. DOI:https://doi.org/10.1145/3306618.3314314

[135] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How do fairness definitions fare?: Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 99–106.

[136] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the International Conference on Machine Learning*. 1670–1679.

[137] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 59–68.

[138] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *stat* 1050 (2017), 22.

[139] Richard Shaw and Manuel Corpas. [n.d.]. Further bias in personal genomics? ([n. d.]).

[140] Harini Suresh and John V. Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).

[141] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why machine learning may lead to unfairness: evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL'19)*. Association for Computing Machinery, New York, NY, USA, 83–92. https://doi.org/10.1145/3322640.3326705

[142] Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*.

[143] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 6373–6382. Retrieved from http://proceedings.mlr.press/v97/ustun19a.html.

[144] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 3003–3008.

[145] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the IEEE/ACM International Workshop on Software Fairness (FairWare'18)*. IEEE, 1–7.

[146] Selwyn Vickers, Mona Fouad, and Moon S. Chen Jr. 2014. Enhancing minority participation in clinical trials (EM-PaCT): Laying the groundwork for improving minority clinical trial accrual. *Cancer* 120 (2014), vi–vii.

[147] Ting Wang and Dashun Wang. 2014. Why Amazon's ratings might mislead you: The story of herding effects. *Big Data* 2, 4 (2014), 196–204.

[148] Steven L. Willborn. 1984. The disparate impact model of discrimination: Theory and limits. *Amer. UL Rev.* 34 (1984), 799.

[149] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. 2009. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European Conference on Computer Systems*. ACM, 205–218.

[150] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081* (2017).

[151] Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. Fairness-aware Classification: Criterion, Convexity, and Bounds. arXiv:cs.LG/1809.04737 (2018).

[152] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. FairGAN: Fairness-aware generative adversarial networks. In *Proceedings of the IEEE International Conference on Big Data (Big Data'18)*. IEEE, 570–575.

[153] Irene Y. Chen, Peter Szolovits, and Marzyeh Ghassemi. 2019. Can AI help reduce disparities in general medical and mental health care? *AMA J. Ethics* 21 (02 2019), E167–179. DOI:https://doi.org/10.1001/amajethics.2019.167

[154] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. 1171–1180.

[155] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259* (2015).

[156] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: A causal modeling-based framework. *Int. J. Data Sci. Analyt.* 4, 1 (01 Aug. 2017), 1–16. DOI:https://doi.org/10.1007/s41060-017-0058-x

[157] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. On discrimination discovery using causal networks. In *Social, Cultural, and Behavioral Modeling*, Kevin S. Xu, David Reitter, Dongwon Lee, and Nathaniel Osgood (Eds.). Springer International Publishing, Cham, 83–93.

[158] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation testing-based discrimination discovery: A causal inference approach. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 2718–2724. Retrieved from http://dl.acm.org/citation.cfm?id=3060832.3061001.

[159] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1335–1344.

[160] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. 3929–3935. DOI:https://doi.org/10.24963/ijcai.2017/549

[161] L. Zhang, Y. Wu, and X. Wu. 2018. Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. *IEEE Trans. Knowl. Data Eng.* (2018), 1–1. DOI:https://doi.org/10.1109/TKDE.2018.2872988

[162] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 629–634.

[163] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[164] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. arXiv:cs.CL/1804.06876 (2018).

[165] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 4847–4853.

[166] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist it's time to make it fair. Nature Publishing Group.