

# 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 AI systems and applications in our everyday lives, it is important to take fairness issues into consideration while designing and engineering these types of systems. Such systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that the decisions do not reflect discriminatory behavior toward certain groups or populations. We have recently seen work in machine learning, natural language processing, and deep learning that addresses such challenges in different subdomains. With the commercialization of these systems, researchers are becoming aware of the biases that these applications can contain and have attempted 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 in order 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 how 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**; **Philosophical/theoretical foundations of 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 algorithms have penetrated every aspect of our daily life. Algorithms make movie recommendations, suggest products to buy, and are increasingly used in high-stakes decisions in loan applications [93], dating and hiring [18, 35]. There are clear benefits to algorithmic decision-making; unlike people, machines do not become tired or bored [40, 98], and 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” [5, 100]. In the context of decision-making, fairness is the *absence of any prejudice or favoritism toward an individual or a 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 parole 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:<sup>1</sup> COMPAS is more likely to assign a higher risk score to African-American offenders than to Caucasians with the same profile. Similar findings

---

<sup>1</sup><https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

have been made in other areas, such as an AI system that judges beauty pageant winners but was biased against darker-skinned contestants,<sup>2</sup> or facial recognition software in digital cameras that overpredicts Asians as blinking.<sup>3</sup> 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 arising from biases in the data and those arising 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 that prevent them from making fair decisions—even when the data is unbiased.

We begin the review with several highly visible real-world cases of where unfair machine learning algorithms have led to suboptimal and discriminatory outcomes. We then describe the many types of biases that occur in data and present the different ways that the concept of fairness has been operationalized and studied in 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, followed by potential areas of future work in each of the domains.

## 2 REAL-WORLD EXAMPLES OF ALGORITHMIC UNFAIRNESS

With the popularity of AI and machine learning over the past decades, and their epidemic spread in different applications, safety and fairness constraints have become a huge issue for researchers and engineers. Machine learning is used in courts to assess the probability that a defendant recommit a crime. It is used in different medical fields, in childhood welfare systems [32], 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, with considerations to fairness. [101] 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. [58] also discusses some 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 important for researchers and engineers to be concerned about the downstream applications and their potential harmful effects when modeling an algorithm or a system. A well-known example is COMPAS, which is a widely used commercial risk assessment software that was compared to normal human judgment in a study and was later discovered to be not any better than a normal human [41]. It is also interesting to note that although COMPAS uses 137 features, only 7 of those were presented to the people in the study. In [41], authors also argued that COMPAS is not any better than a simple logistic regression model when making decisions. We should think responsibly, and recognize that these tools are used in courts and are actually making decisions which 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 [119] compared SAVRY, a tool used in risk assessment frameworks that includes human intervention in its process, with automatic machine learning methods in order to see which one is more accurate and fair. Conducting these types of studies should be done more frequently, but prior to releasing the tools in order to avoid doing harm.

Another interesting direction that researchers have taken is introducing tools that can assess the amount of fairness in a tool or system. For example, Aequitas [110] is a toolkit that lets users to test

<sup>2</sup><https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people>

<sup>3</sup><http://content.time.com/time/business/article/0,8599,1954643,00.html>

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 at IBM in order 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 [10]. 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. In addition to COMPAS, discriminatory behavior was also evident in an algorithm that would deliver advertisements promoting jobs in Science, Technology, Engineering, and Math (STEM) fields [74]. 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 which 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 [106] and recommender systems [114] have also been largely studied and evaluated and in many cases shown to be discriminative towards certain populations and subgroups. In order to be able to address the bias issue in these application, it is important for us to know where these biases are coming from and what we can do to prevent them.

### 3 BIAS IN DATA

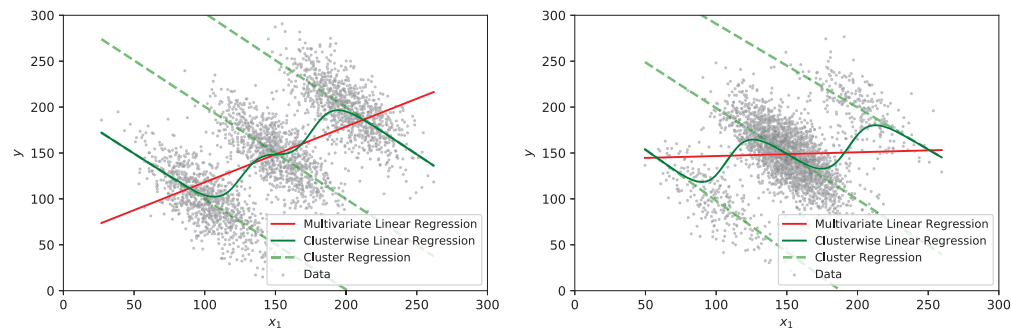


Fig. 1. Illustration of biases in data. Red line shows the regression (MLR) for the entire population, while dashed green lines are regressions for each subgroup, and the solid green line is the unbiased regression. (a) When all subgroups are of equal size, then MLR shows a positive relationship between 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)

Data, especially big data, is often heterogeneous, generated by subgroups with their own characteristics and behaviors. The heterogeneities, some of which are described below, can bias the data. A model learned on biased data may lead to unfair and inaccurate predictions. To illustrate how biases in data affect machine learning, consider a hypothetical nutrition study which measured how the outcome, body mass index (BMI), changes as a function of daily pasta calorie intake (Figure 1). Regression analysis (solid red line) shows a positive relationship in the population between these variables. The positive trend suggests that increased pasta consumption is associated with higher BMI. However, unbeknown to researchers, the study population was heterogeneous, composed of subgroups that vary in their fitness level—people who did not exercise, people with normal activity levels, and athletes. When data is disaggregated by fitness level, the trends within each subgroup are

negative (dashed green lines), leading to the conclusion that increased pasta consumption is, in fact, associated with a lower BMI. Recommendations for pasta consumption that come from the naive analysis are opposite to those coming from a more careful analysis that accounts for the differences between subgroups.

### 3.1 Types of Bias

Bias in data can exist in many shapes and forms, some of which can lead to unfairness in different downstream learning tasks. In [118], authors talk about sources of bias in machine learning with their categorizations and descriptions in order to motivate future solutions to each of the sources of bias introduced in the paper. In [99], 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 some of the most general and important sources of bias introduced in these two papers and also add in some work from other existing research papers. Additionally, we will also introduce a different categorization and grouping of these definitions later in the paper.

- (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* [118]. 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 woman—which would cause the search results to be biased towards male CEOs [118]. 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) **Representation Bias.** *Representation bias happens from the way we define and sample from a population* [118]. Lacking geographical diversity in datasets like ImageNet (as shown in Figures 3 and 4) is an example for this type of bias [118]. This demonstrates a bias towards Western countries.
- (3) **Measurement Bias.** *Measurement bias happens from the way we choose, utilize, and measure a particular feature* [118]. 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 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 [118].
- (4) **Evaluation Bias.** *Evaluation bias happens during model evaluation* [118]. 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 [23], and can serve as examples for this type of bias [118].
- (5) **Aggregation Bias.** *Aggregation bias happens when false conclusions are drawn for a subgroup based on observing other different subgroups or generally when false assumptions about a population affect the model’s outcome and definition* [118]. An example of this type of bias can be seen in clinical aid tools. Consider diabetes patients who have apparent differences across ethnicities and genders, or more specifically, HbA1c levels that are widely used in diagnosis and monitoring of diabetes are different in complicated ways across genders and ethnicities. Therefore, because of these factors and their different meanings and importance across different sub-groups and populations, a single model would most probably not be best-suited for all

groups in a population [118]. This is true even when they are represented equally in the training data. Any general assumptions about different populations can result in aggregation bias.

- (6) **Population Bias.** *Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset or platform from the original target population* [99]. 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 of 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 [56].
- (7) **Simpson’s Paradox.** Simpson’s paradox [17] can bias the analysis of heterogeneous data that is composed of subgroups or individuals with different behaviors. *According to Simpson’s paradox, a trend, association, or characteristic observed in underlying subgroups may be quite different from association or characteristic observed when these subgroups are aggregated.* One of the better-known examples of this type of paradox arose during the gender bias lawsuit in university admissions against UC Berkeley [15]. 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 [34], psychology [68], astronomy [89], and computational social science [77].
- (8) **Longitudinal Data Fallacy.** *Observational studies often treat cross-sectional data as if it were longitudinal, which may create biases due to Simpson’s paradox.* As an example, analysis of bulk Reddit data [9] 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.
- (9) **Sampling Bias.** *Sampling bias 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 again the example in Figure 1. Suppose the next time the study is conducted, one of the subgroups is sampled more frequently than the rest. 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.
- (10) **Behavioral Bias.** *Behavioral bias arises from different user behavior across platforms, contexts, or different datasets* [99]. An example of this type of bias can be observed in [88], where authors show how differences in emoji representations among platforms can result in different reactions and behavior from people and sometimes even leading to communication errors.
- (11) **Content Production Bias.** *Content Production bias arises from structural, lexical, semantic, and syntactic differences in the contents generated by users* [99]. An example of this type of bias can be seen in [97] 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.
- (12) **Linking Bias.** *Linking bias arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users* [99]. In [85] 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. [126] 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 [51, 91], which can change the network measures and cause different types of problems.

- (13) **Temporal Bias.** *Temporal bias arises from differences in populations and behaviors over time* [99]. 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 [99, 120].
- (14) **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* [96]. As an instance, this type of bias can be seen in search engines [61, 96] 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.
- (15) **Algorithmic Bias.** *Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm* [8].
- (16) **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* [8]. This type of bias can be influenced by other types and subtypes, such as Presentation and Ranking biases.
 

**Presentation Bias.** *Presentation bias is a result of how information is presented* [8]. 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 [8].

**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 [8] and crowdsourcing applications [78].
- (17) **Social Bias.** *Social bias happens when other people’s actions or content coming from them affect our judgment.* [8]. 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 [8, 125].
- (18) **Emergent Bias.** *Emergent bias happens 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* [46]. 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 [46]. This type of bias can itself be divided into more subtypes, as discussed in detail in [46].
- (19) **Self-Selection Bias.** *Self-selection bias<sup>4</sup> 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 situations where survey takers decide that they can appropriately participate in a study themselves. For instance, in a survey study about smart or successful students, some less successful students might think that they are successful to take the survey—which would then bias the outcome of the analysis. In fact, the chances of this situation happening is high, as the more successful students probably would not spend time filling out surveys that would increase the risk of self-selection bias.

<sup>4</sup><https://data36.com/statistical-bias-types-explained/>

- (20) **Omitted Variable Bias.** *Omitted variable bias<sup>4</sup> occurs when one or more important variables are left out of the model.* 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 which 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.
- (21) **Cause-Effect Bias.** *Cause-effect bias<sup>4</sup> can happen as a result of the fallacy that correlation implies causation.* An example of this type of bias can be observed in a situation where a data analyst in a company wants to analyze how successful a new loyalty program is. The analyst sees that customers who signed up for the loyalty program are spending more money in the company's e-commerce store than those who did not. It is going to be problematic if the analyst immediately jumps to the conclusion that the loyalty program is successful, since it might be the case that only more committed or loyal customers, who might have planned to spend more money anyway, are interested in the loyalty program in the first place. This type of bias can have serious consequences due to its nature and the roles it can play in sensitive decision-making policies.
- (22) **Observer Bias.** *Observer bias<sup>4</sup> happens when researchers subconsciously project their expectations onto the research.* This type of bias can happen when researchers (unintentionally) influence participants (during interviews and surveys) or when they cherry pick participants or statistics that will favor their research.
- (23) **Funding Bias.** *Funding bias<sup>4</sup> arises when biased results are reported in order to support or satisfy the funding agency or financial supporter of the research study.* As an example, this manifests when employees of a company report biased results in their data and statistics in order to keep the funding agencies or other parties satisfied.

Existing work tried 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, which is a situation in which the trained machine learning model makes decisions that produce outcomes, and these very outcomes affect future data that will be collected for subsequent training rounds or models [33], these definitions are intertwined, and we need a categorization which closely models this situation. This feedback loop is not only existent between the data and the algorithm, but also some work analyzed the existence of this loop between the algorithms and user interaction [28]. Getting inspiration from these papers, we also modeled categorization of bias definitions, as shown in Figure 2, 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 try to 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. In [23], the authors show that datasets like 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 skin 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 [23], 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-skin females, light-skin males, dark-skin males, and dark-skin females.

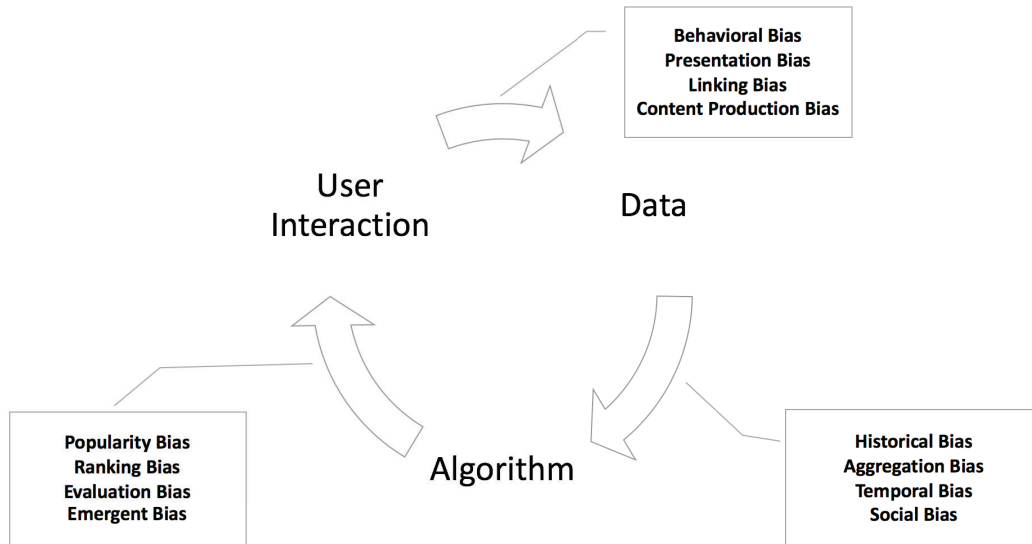


Fig. 2. Bias definitions in the data, algorithm, and user interaction feedback loop are placed on their most appropriate arrows.

It's only in this case that we can clearly observe the bias towards dark-skin females, as previously dark-skin males would compromise for dark-skin females and would hide the underlying bias towards this subgroup. [142] also analyzes some examples of the biases that can exist in the data and algorithms and offer some recommendations and suggestions toward mitigating these issues. 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. [83] 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 [117] 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 [47] 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. [124] has other examples of studies on existing biases in the data used in the medical domain. [130] also looks at machine-learning models and data utilized in medical fields, and writes about how artificial intelligence in health care has not impacted all patients equally. This issue not only exists in medical fields, but 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 and Open Images are two widely used datasets in machine-learning. In [116], researchers studied and showed that these datasets suffer from representation bias and advocate for the need to incorporate geo-diversity and inclusion while creating such datasets.

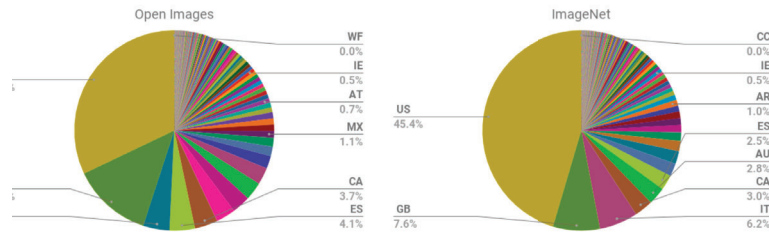


Fig. 3. Fraction of each country, represented by their two-letter ISO codes, in Open Images and ImageNet image datasets. In both datasets, US and Great Britain represent the top locations, from [116].

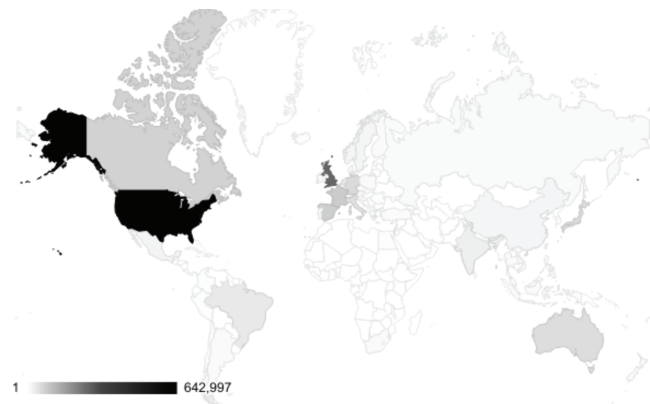


Fig. 4. Geographical distribution representation for countries in the Open Images data set. In their sample, almost one third of the data was US-based, and 60% of the data was from the six most represented countries across North America and Europe, from [116].

## 4 ALGORITHMIC FAIRNESS

Fighting against bias and discrimination has a long history in philosophy and psychology, and recently in machine-learning. However, in order to be able to fight against discrimination and achieve fairness, one should first define the notion of fairness. Philosophy and psychology have tried to define the concept of fairness long before computer science started exploring it. The fact that no universal definition of fairness exists shows the difficulty of solving this problem [112]. 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.

### 4.1 Types of Discrimination

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 [113]. Even though fairness is an incredibly desirable quality in society, it can be surprisingly difficult to achieve in practice. In

order to understand how we can have so many definitions of fairness, it is also crucial to understand the different kinds of discrimination that may occur.

- (1) **Direct Discrimination.** Direct discrimination happens when protected attributes of individuals explicitly result in non-favorable outcomes toward them [136]. 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 for some of these protected attributes is provided in Table 4 as specified in the Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA) [29].
- (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 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.) [136].
- (3) **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 [36]. For example, a restaurant fulfilling customers' requests that lead to discriminatory placement of employees would be systemic discrimination. [108] found that employers overwhelmingly preferred competent candidates that were culturally similar to them personally, and shared similar experiences and hobbies. If the decision-makers happen to belong overwhelmingly to certain subgroups, this may result in discrimination against competent candidates that do not belong to these subgroups.
- (4) **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, that may actually be relevant to the outcome [102].
- (5) **Explainable Discrimination.** Differences in treatment and outcomes amongst 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 [64]. For instance, authors in [64] state that in the UCI Adult dataset [6], 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 which needs to be considered. If we make decisions, without considering working hours, such that males and females end up averaging the same income, 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 [64], 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.
- (6) **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 [64] also present local techniques for removing only the illegal or unexplainable discrimination, allowing only for explainable differences in decisions. These are pre-processing techniques that change the training data such that it contains no unexplainable discrimination. We expect classifiers trained on this pre-processed data to not capture illegal or unexplainable discrimination.

## 4.2 Definitions of Fairness

In [16], authors studied fairness definitions in political philosophy and tried to tie them to machine-learning. Authors in [60] studied the 50-year history of fairness definitions in the areas of education and machine-learning. In [123], authors listed and explained some of the definitions used for fairness in algorithmic classification problems. In [113], 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 [123].

**Definition 1.** (*Equalized Odds*). *The definition of equalized odds, provided by [55], 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 (male and female) group members [123]. 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  $P(\hat{Y}=1|A=0, Y=1) = P(\hat{Y}=1|A=1, Y=1)$ " [55].* 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 [123]. 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  $P(\hat{Y}=1|A=0) = P(\hat{Y}=1|A=1)$ " [43, 73].* The likelihood of a positive outcome [123] 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" [43, 73].* 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" [53, 73].*

**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" [14].*

**Definition 7.** (*Test Fairness*). *"A score  $S = S(x)$  is testfair (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  $s$ ,  $P(Y=1|S=s, R=b) = P(Y=1|S=s, R=w)$ " [31].* In other words, the test fairness definition states that for any predicted probability score  $S$ , people in both protected and unprotected (female and male) groups must have equal probability of correctly belonging to the positive class [123].

**Definition 8.** (*Counterfactual Fairness*). "Predictor  $\hat{Y}$  is counterfactually fair if under any context  $X = x$  and  $A = a$ ,  $P(\hat{Y}_{A \leftarrow a}(U) = y | X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y | X = x, A = a)$ , (or all  $y$  and for any value  $a'$  attainable by  $A$ " [73]. 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" [44].

**Definition 10.** (*Conditional Statistical Parity*). For a set of legitimate factors  $L$ , predictor  $\hat{Y}$  satisfies conditional statistical parity if  $P(\hat{Y} | L = l, A = 0) = P(\hat{Y} | L = l, A = 1)$  [37]. 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  $L$  [123].

Fairness definitions fall under different types as follows:

- (1) **Individual Fairness.** Give similar predictions to similar individuals [43, 73].
- (2) **Group Fairness.** Treat different groups equally [43, 73].
- (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 in order 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 [66, 67].

Name	Reference	Group	Individual
Demographic parity	[73][43]	✓	
Conditional statistical parity	[37]	✓	
Equalized odds	[55]	✓	
Equal opportunity	[55]	✓	
Fairness through unawareness	[73][53]		✓
Fairness through awareness	[43]		✓
Counterfactual fairness	[73]		✓

Table 1. Categorizing different fairness notions into group vs. individual types.

It is important to note that according to [70], it is impossible to satisfy some of the fairness constraints at once except in highly constrained special cases. In [70], 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 [115]. Another important aspect to consider is time and temporal analysis of the impacts that these definitions may have on individuals or groups. In [80] 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 trade-offs 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 in order 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 domains and sub-domains that we focus upon in this survey.

Domain	Sub-domain	Reference(s)
Data	Simpson’s Paradox	[68]
Machine learning	Classification	[65] [86] [49] [72] [121] [55] [131] [127] [59] [24] [128]
Machine learning	Regression	[13] [1]
Machine learning	PCA	[111]
Machine learning	Community detection	[85]
Machine learning	Clustering	[30] [7]
Machine learning	Graph embedding	[21]
Machine learning	Causal inference	[81] [136] [137] [132] [95] [94] [134] [69] [105] [133]
Natural language processing	Word embedding	[19] [141] [50] [22] [138]
Natural language processing	Coreference resolution	[140] [109]
Natural language processing	Language model	[20]
Natural language processing	Sentence embedding	[84]
Natural language processing	Machine translation	[45]
Natural language processing	Semantic role labeling	[139]
Deep learning / representation learning	Variational auto encoders	[82] [4] [92] [38]
Deep learning / representation learning	Adversarial learning	[76] [129]

Table 2. List of papers targeting and talking about bias and fairness in different areas and sub-areas of machine learning.

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 that the underlying discrimination is removed [39]. If the algorithm is allowed to modify the training data, then pre-processing can be used [10].
- (2) **In-processing.** In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process [39]. 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 [10, 13].

- (3) **Post-processing.** Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model [39]. 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 [10, 13].

Examples of some existing work and their categorization into these types is shown in Table 3. These methods are not just limited to general machine learning techniques, but because of AI’s popularity, they have expanded to different domains in natural language processing and deep learning. From learning fair representations [38, 82, 92] to learning fair word embeddings [19, 50, 141], 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) which 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 [79]. A list of protected or sensitive attributes is provided in Table 4. They point out what attributes should not affect the outcome of the decision in housing loan or credit card decision-making [29] according to the law. Some of the existing work tries to treat sensitive attributes as noise to disregard their effect on the decision-making, while some causal methods try to 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 Bias Mitigation

In order 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 [12, 48]. [11] proposes a similar approach for the NLP applications. A similar suggestion has been proposed for models in [90]. Authors in [57] also propose having labels, just like nutrition labels on food, in order 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, [68] has proposed methods to test for cases of Simpson’s paradox in the data, and [2, 3] 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 [135]. [54] also worked on preventing discrimination in data mining, targeting direct, indirect, and simultaneous effects.

## 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.

Algorithm	Reference	Pre-Processing	In-Processing	Post-Processing
Community detection	[85]	✓		
Word embedding	[22]	✓		
Optimized pre-processing	[26]	✓		
Data pre-processing	[63]	✓		
Classification	[131]		✓	
Regression	[13]		✓	
Classification	[65]		✓	
Classification	[128]		✓	
Adversarial learning	[76]		✓	
Classification	[55]			✓
Word embedding	[19]			✓
Classification	[103]			✓

Table 3. Algorithms categorized into their appropriate groups based on being pre-processing, in-processing, or post-processing.

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

Table 4. A list of the protected attributes as specified in the Fair Housing and Equal Credit Opportunity Acts (FHA and ECOA), from [29].

**5.2.1 Fair Classification.** Since classification is a serious 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 [49, 65, 72, 86] that satisfy certain definitions of fairness. For instance, in [121] authors try to satisfy the subgroup fairness in the classification task, equality of opportunity and equalized odds in [55], both disparate treatment and disparate impact in [131], and equalized odds in [127]. Other methods try to not only satisfy some fairness constraints but to also be stable toward change in the test set [59]. Other works tried to propose three different modifications to the existing Naive Bayes classifier for discrimination-free classification [24]. The authors in [128], propose a general framework for learning fair classifiers. This framework can be used for formulating fairness-aware classification with fairness guarantees.

**5.2.2 Fair Regression.** [13] proposes a fair regression method along with evaluating it with a measure introduced as the "price of fairness" (POF) to measure accuracy-fairness trade-offs. They

introduce three fairness penalties as follows:

**Individual Fairness:** The definition for individual fairness as stated in [13], "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

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

**Group Fairness:** "On average, the two groups' instances should have similar labels (weighted by the nearness of the labels of the instances)" [13].

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

**Hybrid Fairness:** "Hybrid fairness requires both positive and both negatively labeled cross pairs to be treated similarly in an average over the two groups" [13].

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

In addition to the previous work, [1] considers the fair regression problem formulation with regards to two notions of fairness statistical (demographic) parity and bounded group loss.

**5.2.3 Structured Prediction.** In [139], 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<sup>5</sup> on the dataset, bias is magnified for "man", filling only 16% of cooking images. Under these observations, the authors of the paper [139] 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 exemplify the existing bias in data.

**5.2.4 Fair PCA.** In [111] 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 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 which 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 optimization function) is as follows, in

<sup>5</sup>Specifically, a Conditional Random Field (CRF)

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  $R^n$ :

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

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 [85]. 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 [85] 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 [21] and clustering [7, 30].

**5.2.6 Fair Causal Inference.** Many researchers have used causal models and graphs to solve fairness-related concerns in machine learning. In [81], 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 in order to make decisions that are irrespective of sensitive attributes of groups or individuals. For instance, in [136] authors propose a causal-based framework that detects direct and indirect discrimination in the data along with their removal techniques. [137] is an extension to the previous work. [132] 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 [95] 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 [94] 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 [134] 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, they declare the existence of discrimination. In [69], 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" [69].

**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" [69]. 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 [105], 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:

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

$RD^c$  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. [133] 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 [82]. Here, they treat the sensitive variable as the nuisance variable, so that 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 [62]. In [4] 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, etc.) and propose an algorithm on top of this DB-VAE using these latent variables to debias systems like facial detection systems. In [92] 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 in order 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  $x$ .

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

In [38], 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 [76] 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, in order 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  $Y=y$ . [129] 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 which 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  $G_{Dec}$  which 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  $P_G(x, y, s)$ .

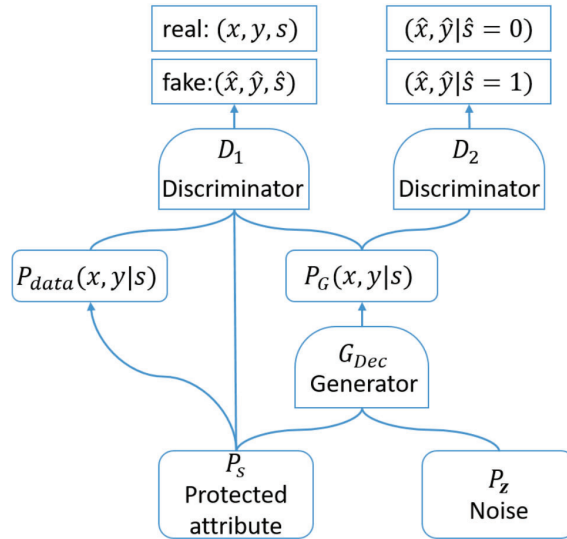


Fig. 5. Structure of FairGAN as proposed in [129].

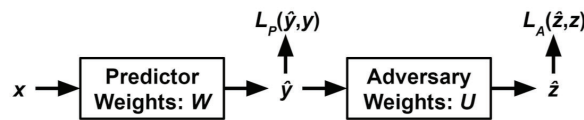


Fig. 6. The architecture of adversarial network proposed in [76].

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 [129].

## 5.4 Fair NLP

**5.4.1 Word Embedding.** In [19] 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)*

- (1) **Identify gender subspace.** Identifying a direction of the embedding that captures the bias [19].
- (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 [19]. Equalize makes gender-neutral words to be equidistant from the equality set of gendered words [19].
  - (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 trade-off is controlled by a parameter [19].

Following on the footsteps of these authors, other future work attempted to tackle this problem [141] 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 [50] 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 don’t actually remove it. A recent work [22] 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 [138], 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.** The [140] paper 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 data set 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 [109], 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 [20] 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, etc., 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.

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

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 [19]. Finally, they evaluate the effectiveness of their method on reducing gender bias and conclude by stating that in order 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 [84] 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 [25]) in the context of sentence encoders by introducing their new sentence encoding bias-measuring techniques, and 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 [45] 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 [45] 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 [104] 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 [122] authors annotated and analyzed the Europarl dataset [71], 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.

## 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.
- (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 [52]. 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 [2]; 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.

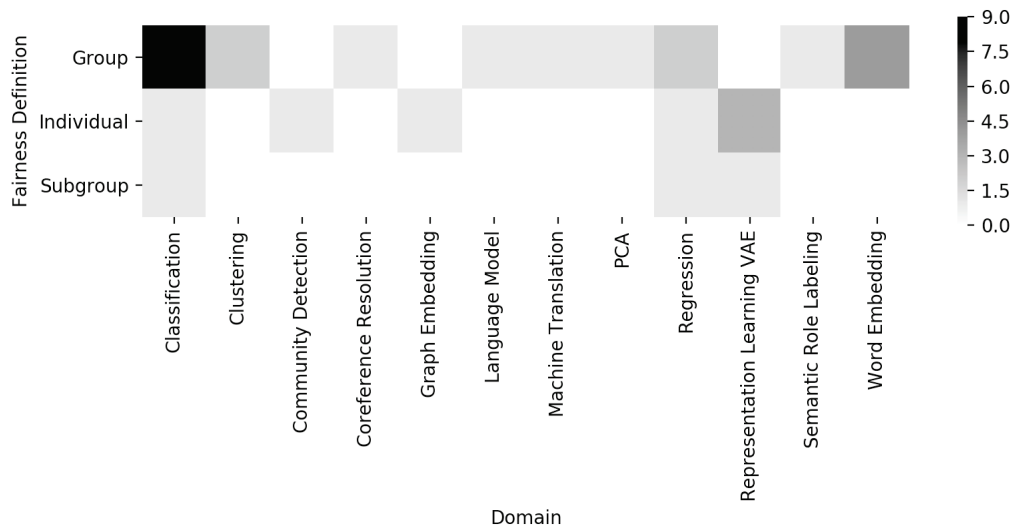


Fig. 7. Heatmap depicting distribution of previous work in fairness, grouped by domain and fairness definition.

## 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 some of the problems and showed why considering fairness is an important issue, along with providing some 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 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.

## 8 ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR0011890019. We would like to thank Mila and IVADO for their

summer school on Bias and Discrimination in AI and all the speakers. Special Thank to Golnoosh Farnadi and Jihane Lamouri for organizing the event.

## 9 APPENDIX

### 9.1 Datasets for Fairness Research

Aside from the existence of bias in datasets, there are datasets 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.

**9.1.1 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 etc., 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 [6].

**9.1.2 German Credit Dataset.** The German Credit dataset contains 1000 credit records containing attributes such as personal status and sex, credit score, credit amount, housing status etc. It can be used in studies about gender inequalities on credit-related issues [42].

**9.1.3 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 [140].

**9.1.4 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.

**9.1.5 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 [75].

**9.1.6 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 [119].

**9.1.7 Pilot Parliaments Benchmark Dataset.** The Pilot Parliaments Benchmark dataset, also known as PPB, contains images of 1270 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 [23].

**9.1.8 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 [87].

Dataset Name	Reference	Size	Area
UCI adult dataset	[6]	48,842 income records	Social
German credit dataset	[42]	1,000 credit records	Financial
Pilot parliaments benchmark dataset	[23]	1,270 images	Facial images
WinoBias	[140]	3,160 sentences	Coreference resolution
Communities and crime dataset	[107]	1,994 crime records	Social
COMPAS Dataset	[75]	18,610 crime records	Social
Recidivism in juvenile justice dataset	[27]	4,753 crime records	Social
Diversity in faces dataset	[87]	1 million images	Facial images

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.

## REFERENCES

- [1] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *International Conference on Machine Learning*. 120–129.
- [2] 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 Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 19–27.
- [3] Nazanin Alipourfard, Peter G Fennell, and Kristina Lerman. 2018. Using Simpson’s Paradox to Discover Interesting Patterns in Behavioral Data. In *Twelfth International AAAI Conference on Web and Social Media*.
- [4] Alexander Amini, Ava Soleimany, Wilko Schwarting, Sangeeta Bhatia, and Daniela Rus. 2019. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. (2019).
- [5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: there’s software used across the country to predict future criminals. And it’s biased against blacks. ProPublica 2016. (2016).
- [6] A. Asuncion and D.J. Newman. 2007. UCI Machine Learning Repository. (2007). <http://www.ics.uci.edu/~lmslearn/>
- [7] 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, Long Beach, California, USA, 405–413. <http://proceedings.mlr.press/v97/backurs19a.html>
- [8] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (May 2018), 54–61. <https://doi.org/10.1145/3209581>
- [9] Samuel Barbosa, Dan Cosley, Amit Sharma, and Roberto M. Cesar-Jr. 2016. Averaging Gone Wrong: Using Time-Aware Analyses to Better Understand Behavior. (April 2016), 829–841.
- [10] Rachel KE 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).
- [11] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. [https://doi.org/10.1162/tacl\\_a\\_00041](https://doi.org/10.1162/tacl_a_00041)
- [12] Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. [n. d.]. TOWARDS STANDARDIZATION OF DATA LICENSES: THE MONTREAL DATA LICENSE. ([n. d.]).
- [13] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. (2017). [arXiv:cs.LG/1706.02409](https://arxiv.org/abs/1706.02409)
- [14] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. [n. d.]. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* ([n. d.]), 0049124118782533.
- [15] 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.
- [16] RDP Binns. 2018. Fairness in machine learning: Lessons from political philosophy. *Journal of Machine Learning Research* (2018).
- [17] Colin R Blyth. 1972. On Simpson’s paradox and the sure-thing principle. *J. Amer. Statist. Assoc.* 67, 338 (1972), 364–366.
- [18] Miranda Bogen and Aaron Rieke. 2018. *Help wanted: an examination of hiring algorithms, equity*. Technical Report. and bias. Technical report, Upturn.

- [19] 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 *Advances in neural information processing systems*. 4349–4357.
- [20] Shikha Bordia and Samuel Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 7–15.
- [21] Avishek Bose and William Hamilton. 2019. Compositional Fairness Constraints for Graph Embeddings. In *International Conference on Machine Learning*. 715–724.
- [22] 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, Long Beach, California, USA, 803–811. <http://proceedings.mlr.press/v97/brunet19a.html>
- [23] 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, New York, NY, USA, 77–91. <http://proceedings.mlr.press/v81/buolamwini18a.html>
- [24] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.
- [25] 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.
- [26] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *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. <http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf>
- [27] Manel Capdevila, Marta Ferrer, and Eulàlia 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).
- [28] Allison JB 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.
- [29] 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.
- [30] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. 2019. Proportionally Fair Clustering. In *International Conference on Machine Learning*. 1032–1041.
- [31] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
- [32] 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, New York, NY, USA, 134–148. <http://proceedings.mlr.press/v81/chouldechova18a.html>
- [33] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).
- [34] John S. Chuang, Olivier Rivoire, and Stanislas Leibler. 2009. Simpson’s Paradox in a Synthetic Microbial System. *Science* 323, 5911 (2009), 272–275. <https://doi.org/10.1126/science.1166739> arXiv:<https://science.sciencemag.org/content/323/5911/272.full.pdf>
- [35] Lee Cohen, Zachary C. Lipton, and Yishay Mansour. 2019. Efficient candidate screening under multiple tests and implications for fairness. (2019). arXiv:cs.LG/1905.11361
- [36] United States. Equal Employment Opportunity Commission. [n. d.]. *EEOC compliance manual*. [Washington, D.C.]: U.S. Equal Employment Opportunity Commission, [1992].
- [37] 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.
- [38] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly Fair Representation Learning by Disentanglement. In *International Conference on Machine Learning*. 1436–1445.

- [39] 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.
- [40] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108, 17 (2011), 6889–6892.
- [41] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018). <https://doi.org/10.1126/sciadv.aao5580> arXiv:<https://advances.sciencemag.org/content/4/1/eaao5580.full.pdf>
- [42] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. (2017). <http://archive.ics.uci.edu/ml>
- [43] 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, USA, 214–226. <https://doi.org/10.1145/2090236.2090255>
- [44] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. 2018. Fairness in Relational Domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’18)*. ACM, New York, NY, USA, 108–114. <https://doi.org/10.1145/3278721.3278733>
- [45] 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).
- [46] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. <https://doi.org/10.1145/230538.230561>
- [47] 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. *American Journal of Epidemiology* 186, 9 (06 2017), 1026–1034. <https://doi.org/10.1093/aje/kwx246> arXiv:<http://oup.prod.sis.lan/aje/article-pdf/186/9/1026/24330720/kwx246.pdf>
- [48] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. [n. d.]. Datasheets for Datasets. ([n. d.]).
- [49] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [50] 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).
- [51] Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014. Assessing the bias in samples of large online networks. *Social Networks* 38 (2014), 16–27.
- [52] Susan T Gooden. 2015. *Race and social equity: A nervous area of government*. Routledge.
- [53] 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 *NIPS Symposium on Machine Learning and the Law*, Vol. 1. 2.
- [54] S. Hajian and J. Domingo-Ferrer. 2013. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 25, 7 (July 2013), 1445–1459. <https://doi.org/10.1109/TKDE.2012.72>
- [55] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
- [56] Eszter Hargittai. 2007. Whose Space? Differences among Users and Non-Users of Social Network Sites. *Journal of Computer-Mediated Communication* 13, 1 (10 2007), 276–297. <https://doi.org/10.1111/j.1083-6101.2007.00396.x> arXiv:<http://oup.prod.sis.lan/jcmc/article-pdf/13/1/276/22317170/jjemcom0276.pdf>
- [57] 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).
- [58] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics* 24, 5 (2018), 1521–1536.
- [59] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and Fair Classification. In *International Conference on Machine Learning*. 2879–2890.
- [60] 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.
- [61] L. Inrona and H. Nissenbaum. 2000. Defining the Web: the politics of search engines. *Computer* 33, 1 (Jan 2000), 54–62. <https://doi.org/10.1109/2.816269>
- [62] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2018. Unsupervised Adversarial Invariance. (2018). arXiv:cs.LG/1809.10083
- [63] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (01 Oct 2012), 1–33. <https://doi.org/10.1007/s10115-011-0463-8>
- [64] Faisal Kamiran and Indrė Žliobaitė. 2013. *Explainable and Non-explainable Discrimination in Classification*. Springer Berlin Heidelberg, Berlin, Heidelberg, 155–170. [https://doi.org/10.1007/978-3-642-30487-3\\_8](https://doi.org/10.1007/978-3-642-30487-3_8)

- [65] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.
- [66] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*. 2569–2577.
- [67] 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.
- [68] Rogier Kievit, Willem Eduard Frankenhuis, Lourens Waldorp, and Denny Borsboom. 2013. Simpson’s paradox in psychological science: a practical guide. *Frontiers in psychology* 4 (2013), 513.
- [69] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.
- [70] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
- [71] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5. 79–86.
- [72] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioulis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In *Proceedings of the 2018 World Wide Web Conference (WWW ’18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 853–862. <https://doi.org/10.1145/3178876.3186133>
- [73] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *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. <http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf>
- [74] Anja Lambrecht and Catherine E Tucker. 2018. Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads (March 9, 2018)* (2018).
- [75] J Larson, S Mattu, L Kirchner, and J Angwin. 2016. Compas analysis. *GitHub*, available at: <https://github.com/propublica/compas-analysis>[Google Scholar] (2016).
- [76] Blake Lemoine, Brian Zhang, and M Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. (2018).
- [77] Kristina Lerman. 2018. Computational social scientist beware: Simpson’s paradox in behavioral data. *Journal of Computational Social Science* 1, 1 (2018), 49–58.
- [78] Kristina Lerman and Tad Hogg. 2014. Leveraging position bias to improve peer recommendation. *PLoS one* 9, 6 (2014), e98914.
- [79] Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. 2017. Does mitigating ML’s disparate impact require disparate treatment? *stat* 1050 (2017), 19.
- [80] 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*.
- [81] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).
- [82] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2016. THE VARIATIONAL FAIR AUTOENCODER. *stat* 1050 (2016), 4.
- [83] 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 England Journal of Medicine* 375, 7 (2016), 655–665. <https://doi.org/10.1056/NEJMs1507092> arXiv:<https://doi.org/10.1056/NEJMs1507092> PMID: 27532831.
- [84] 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).
- [85] 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).
- [86] 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)*, Soelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 107–118. <http://proceedings.mlr.press/v81/menon18a.html>
- [87] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in Faces. *arXiv preprint arXiv:1901.10436* (2019).
- [88] 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 *Tenth International AAAI Conference on Web and Social Media*.

- [89] I Minchev, G Matijevic, DW 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).
- [90] 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, USA, 220–229. <https://doi.org/10.1145/3287560.3287596>
- [91] 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 *7th International AAAI Conference on Weblogs and Social Media, ICWSM 2013*. AAAI press.
- [92] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. 2018. Invariant Representations without Adversarial Training. In *Advances in Neural Information Processing Systems*. 9084–9093.
- [93] 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. *International Transactions in operational research* 9, 5 (2002), 583–597.
- [94] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2018. Learning Optimal Fair Policies. *arXiv preprint arXiv:1809.02244* (2018).
- [95] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [96] 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).
- [97] 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 Seventh International AAAI Conference on Weblogs and Social Media, ICWSM 2013*. AAAI Press, 439–448. [eemcs-eprint-23604](https://arxiv.org/abs/1307.2360).
- [98] Anne O'Keeffe and Michael McCarthy. 2010. *The Routledge handbook of corpus linguistics*. Routledge.
- [99] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).
- [100] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY, USA.
- [101] Osonde A Osoba and William Welser IV. 2017. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.
- [102] Edmund S Phelps. 1972. The statistical theory of racism and sexism. *The american economic review* 62, 4 (1972), 659–661.
- [103] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *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. <http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf>
- [104] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2018. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications* (2018), 1–19.
- [105] Bilal Qureshi, Faisal Kamiran, Asim Karim, and Salvatore Ruggieri. 2016. Causal discrimination discovery through propensity score analysis. *arXiv preprint arXiv:1608.03735* (2016).
- [106] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products.
- [107] M Redmond. 2011. Communities and crime unnormalized data set. *UCI Machine Learning Repository*. In website: <http://www.ics.uci.edu/mllearn/MLRepository.html> (2011).
- [108] Lauren A Rivera. 2012. Hiring as cultural matching: The case of elite professional service firms. *American sociological review* 77, 6 (2012), 999–1022.
- [109] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 8–14. <https://doi.org/10.18653/v1/N18-2002>
- [110] 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).
- [111] Samira Samadi, Uthaiyon 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., USA, 10999–11010. <http://dl.acm.org/citation.cfm?id=3327546.3327755>
- [112] Nripsuta Ani Saxena. 2019. Perceptions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AI/ES '19)*. ACM, New York, NY, USA, 537–538. <https://doi.org/10.1145/3306618.3314314>

- [113] 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 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 99–106.
- [114] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *International Conference on Machine Learning*. 1670–1679.
- [115] 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.
- [116] 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.
- [117] Richard Shaw and Manuel Corpas. [n. d.]. Further bias in personal genomics? ([n. d.]).
- [118] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002* (2019).
- [119] 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. (2019).
- [120] Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*.
- [121] 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, Long Beach, California, USA, 6373–6382. <http://proceedings.mlr.press/v97/ustun19a.html>
- [122] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3003–3008.
- [123] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.
- [124] Selwyn Vickers, Mona Fouad, and Moon S Chen Jr. 2014. Enhancing Minority Participation in Clinical Trials (EMPaCT): laying the groundwork for improving minority clinical trial accrual. *Cancer* 120 (2014), vi–vii.
- [125] 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.
- [126] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN 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*. Acn, 205–218.
- [127] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081* (2017).
- [128] Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. Fairness-aware Classification: Criterion, Convexity, and Bounds. (2018). [arXiv:cs.LG/1809.04737](https://arxiv.org/abs/1809.04737)
- [129] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 570–575.
- [130] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. 2019. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA journal of ethics* 21 (02 2019), E167–179. <https://doi.org/10.1001/amajethics.2019.167>
- [131] 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).
- [132] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* 4, 1 (01 Aug 2017), 1–16. <https://doi.org/10.1007/s41060-017-0058-x>
- [133] 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.
- [134] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation Testing-based Discrimination Discovery: A Causal Inference Approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI’16)*. AAAI Press, 2718–2724. <http://dl.acm.org/citation.cfm?id=3060832.3061001>
- [135] 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.
- [136] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 3929–3935. <https://doi.org/10.24963/ijcai.2017/549>

- [137] L. Zhang, Y. Wu, and X. Wu. 2018. Causal Modeling-Based Discrimination Discovery and Removal: Criteria, Bounds, and Algorithms. *IEEE Transactions on Knowledge and Data Engineering* (2018), 1–1. <https://doi.org/10.1109/TKDE.2018.2872988>
- [138] 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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 629–634.
- [139] 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 2017 Conference on Empirical Methods in Natural Language Processing*.
- [140] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. (2018). [arXiv:cs.CL/1804.06876](https://arxiv.org/abs/1804.06876)
- [141] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4847–4853.
- [142] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist it’s time to make it fair. (2018).