Bias in Machine Learning - What is it Good for?

Thomas Hellström and **Virginia Dignum** and **Suna Bensch** [1,2]

Abstract. In public media as well as in scientific publications, the term *bias* is used in conjunction with machine learning in many different contexts, and with many different meanings. This paper proposes a taxonomy of these different meanings, terminology, and definitions by surveying the, primarily scientific, literature on machine learning. In some cases, we suggest extensions and modifications to promote a clear terminology and completeness. The survey is followed by an analysis and discussion on how different types of biases are connected and depend on each other. We conclude that there is a complex relation between bias occurring in the machine learning pipeline that leads to a model, and the eventual bias of the model (which is typically related to social discrimination). The former bias may or may not influence the latter, in a sometimes bad, and sometime good way.

1 INTRODUCTION

Media, as well as scientific publications, frequently report on 'Bias in Machine Learning', and how systems based on AI or machine learning are 'sexist' [3] or 'discriminatory' [4] [10,37]. In the field of machine learning, the term bias has an established historical meaning that, at least on the surface, totally differs from how the term is used in typical news reporting. Furthermore, even within machine learning, the term is used in very many different contexts and with very many different meanings. Definitions are not always given, and if they are, the relation to other usages of the word is not always clear. Furthermore, definitions sometimes overlap or contradict each other [8].

The main contribution of this paper is a proposed taxonomy of the various meanings of the term bias in conjunction with machine learning. When needed, we suggest extensions and modifications to promote a clear terminology and completeness. We argue that this is more than a matter of definitions of terms. Terminology shapes how we identify and approach problems, and furthermore how we communicate with others. This is particularly important in multidisciplinary work, such as application-oriented machine learning.

The taxonomy is based on a survey of published research in several areas, and is followed by a discussion on how different types of biases are connected and depend on each other.

Since humans are involved in both the creation of bias, and in the application of, potentially biased, systems, the presented work is related to several of the AI-HLEG recommendations for building Human-Centered AI systems.

Machine learning is a wide research field with several distinct approaches. In this paper we focus on *inductive learning*, which is a corner stone in machine learning. Even with this specific focus, the amount of relevant research is vast, and the aim of the survey is not to provide an overview of all published work, but rather to cover the wide range of different usages of the term bias.

This paper is organized as follows. Section 2 briefly summarizes related earlier work. In Section 3 we survey various sources of bias, as it appears in the different steps in the machine learning process. Section 4 contains a survey of various ways of defining bias in the model that is the outcome of the machine learning process. In Section 5 we provide a taxonomy of bias, and discuss the different types of found biases and how they relate to each other. Section 6 concludes the paper.

2 RELATED WORK

A number of reviews, with varying focuses related to bias have been published recently. Barocas and Selbst [3] give as good overview of various kinds of biases in data generation and preparation for machine learning. Loftus et al. [31] review a number of both non causal and causal notions on fairness, which is closely related to bias. Suresh and Guttag [45] identify a number of sources of bias in the machine learning pipeline. Olteanu et al. [35] investigate bias and usage of data from a social science perspective. Our analysis is complementary to the work cited above, by focusing on bias in conjunction with machine learning, and by examining a wider range of biases. We also provide a novel analysis and discussion on the connections and dependencies between the different types of biases.

3 SOURCES OF BIAS

Our survey of sources of bias is organized in sections corresponding to the major steps in the machine learning process (see Figure 1). In Section 3.1 we describe bias as a major concept in the learning step. In Section 3.2 we focus on our biased world, which is the source of information for the learning process. Section 3.3 describes the plethora of biases related terms used in the data generation process.

3.1 Bias in learning

In inductive learning, the aim is to use a data set $\{(x_i, y_i)\}_{i=1}^N$ to find a function $f^*(x)$ such that $f^*(x_i)$ approximates y_i in a good way. Each x_i is a vector of *features*, while y is denoted the *output*, *target variable* or *label*. An often-discussed case is when machine learning

[1] Umeå University, Sweden, email: {thomas.hellstrom,virginia.dignum, suna.bensch}@umu.se

[2] This research is partly funded by the Swedish funding agency VINNOVA, project number 2019-01930.

[3] BBC News, Oct. 10, 2018. Accessed July. 29, 2020.

[4] Reuters Technology News, Oct. 10, 2018. Accessed July 29, 2020.

is used to build decision support systems that outputs recommendations, for example on whether to accept or deny loan applications in a bank. The features may in this case be income, property magnitude, family status, credit history, gender, and criminal record for a loan applicant, while the output y is a recommendation, 0 or 1.

Without further restrictions, infinitely many functions perfectly match any given data set, but most of them are typically useless since they simply memorize the given data set but generalize poorly for other data from the same application. Therefore, the search for a good f^* has to be limited to a certain space Ω of functions. For example, the function may be assumed to be linear, which is the assumption in linear regression. While this may be sufficient in some cases, more complex function spaces, such as high-order polynomials, or artificial neural networks are often chosen. Each specific function in Ω is typically defined by a set of numerical weights.

This preference of certain functions over others was denoted *bias* by Tom Mitchell in his paper from 1980 with the title *The Need for Biases in Learning Generalizations* [34], and is a central concept in statistical learning theory. The expression *inductive bias* (also known as *learning bias*) is used to distinguish it from other types of biases.

In general, inductive learning can be expressed as the minimization problem

$$f^* = \underset{f \in \Omega}{\arg\min}\, L(f), \tag{1}$$

where $L(f)$ is a $cost function$ quantifying how well f matches the data. The most common loss function is defined as

$$L(f) = \sum_{i=1}^{N} (f(x_i) - y_i)^2. \tag{2}$$

The process of solving the minimization problem in Eq. 1 is called *training*, and uses a *training set* $\{(x_i, y_i)\}_{i=1}^{N}$ to search through Ω for the function f that minimizes the loss function in Eq. 2.

All machine learning techniques for inductive learning (for example neural networks, support vector machines, and K-nearest neighbor), need some kind of inductive bias to work, and the choice of Ω is often a critical design parameter. Having too low inductive bias (Ω too big) may lead to overfit, causing noise in data to affect the choice of f^*. This leads to bad generalization, meaning that f^* does not approximate new data that was not used during learning. On the other hand, having too high inductive bias (Ω too small) may lead to underfit, meaning that f^* approximates both the used data and new data in an equally poor way.

The learning step includes other types of bias than the inductive bias described above. Many machine learning algorithms, in particular within deep learning, contain a large number of *hyper-parameters* that are not learned during training but have to be chosen by the user [4]. For neural networks, the choice of number of hidden nodes and layers and type of activation functions are strictly part of the definition of Ω, but are often seen as hyper-parameters [51]. Other hyper-parameters are related to the way the optimization problem in Eq. 1 is solved. Besides the choice of algorithm (for example back propagation, Levenberg-Marquardt, or Gauss-Newton), learning rate, batch size, number of epochs, and stopping criteria are all important choices that affect which function f^* is finally chosen. The shape of the loss function $L(f)$ is another hyper-parameter that not necessarily have to be as in Eq. 2. Examples of other possible choices are *Mean Absolute Error, Hinge Loss*, and *Cross-entropy Loss*. The choice of meta-parameters usually have a large impact on the resulting model. We propose to denote this particular type of bias as *hyper-parameter bias*.

The learning step involves more possible sources of bias. One example is denoted *uncertainty bias* [19], and has to do with the probability values that are often computed together with each produced classification in a machine learning algorithm[5]. The probability represents uncertainty, and typically has to be above a set threshold for a classification to be considered. For example, a decision support system for bank loan applications may reject an application although it is classified as 'approve', because the probability is below the threshold. This threshold is usually manually set, and may create a bias against underrepresented demographic groups, since less data normally leads to higher uncertainty.

3.2 A Biased World

The word 'bias' has an established normative meaning in legal language, where it refers to 'judgement based on preconceived notions or prejudices, as opposed to the impartial evaluation of facts' [8]. The world around us is often described as biased in this sense, and since most machine learning techniques simply mimic large amounts of observations of the world, it should come as no surprise that the resulting systems also express the same bias.

This bias of the world is sometimes denoted *historical bias*. Historical bias has to do with data that in itself has unwanted properties that are regarded as biased: 'Historical bias arises even if the data is perfectly measured and sampled, if the world *as it is* or *was* leads a model to produce outcomes that are not wanted' [45]. The bias of the world obviously has many dimensions, each one describing some unwanted aspect of the world.

Sometimes, the bias in the world is analyzed by looking at correlations between features, and between features and the label. The authors of [52] show that in a certain data set, the label *cooking* co-occurs unproportionally often with *woman*, as compared to *man*. Since most machine learning techniques depend on correlations, such biases may propagate to learned models or classifiers. The authors of [52] show examples of this, and present techniques to detect and quantify bias related to correlations. They define a subset of output variables G that reflect a demographic attribute such as gender or race (e.g. $G = \{man, woman\}$), and o as a variable that is potentially correlated with the elements of G (e.g. $o =$ *cooking*). To identify unwanted correlations, a bias score for o, with respect to a demographic variable $g \in G$, is defined as

$$b(o,g) = \frac{c(o,g)}{\sum\limits_{g' \in G} c(o,g')},$$

where $c(o,g)$ is the number of occurrences of o and g in a corpus. If $b(o,g) > 1/||G||$, then o is positively correlated with g, which indicates that data is biased in this respect. To identify this particular notion of bias, we propose using the term *co-occurrence bias*.

A large body of research investigate bias properties of text, at sentence level, paragraph level, article level, or entire corpora such as Wikipedia news. In some published work, the word 'bias' simply denotes general, usually unwanted, properties of text [25,40]. *Framing bias* refers to how a text expresses a particular opinion on a topic. The connection between framing bias and gender/race bias is investigated in [27], which presents a corpus with sentences expressing negative

[5] For many classification algorithms for example K-nearest neighbor, Artificial Neural Networks, and Naïve Bayes, the output is the probabilities for each class y_i, given the input x: $P(y_i \mid x)$. The predicted class is simply the class with the highest such probability.

bias towards certain races and genders. Another example of text related bias is *epistemological bias*, which refers to the degree of belief expressed in a proposition. For example, the word *claimed* expresses an epistemological bias towards doubts, as compared to *stated*.

As the authors of [24] conclude, text related bias depends not only on individual words, but also on the context in which they appear. The authors use the broader term *language bias* with reference to to the guidelines for Neutral Point of View[6] (NPOV). Aimed for Wikipedia editors writing on controversial topics, NPOV suggests to '(i) avoid stating opinions as facts, (ii) avoid stating seriously contested assertions as facts, (iii) avoid stating facts as opinions, (iv) prefer nonjudgemental language, and (v) indicate the relative prominence of opposing views'. Wagner et al. [47] present and apply several measures for assessing gender bias in Wikipedia. *Coverage bias* is computed by comparing the proportions of notable men and women that are covered by Wikipedia (the somewhat surprising result is that this proportion is higher for women).

3.3 Bias in Data Generation

In machine learning, data generation is responsible for acquiring and processing observations of the real world, and deliver the resulting data for learning. Several sub-steps can be identified, each one with potential bias that will affect the end result. As a first step, the data of interest has to be specified. The specification guides the measurement step, which may be automatic sensor based data acquisition, or manual observations of phenomena of interest. For inductive learning, data is then usually manually labelled. In the following, possible sources of bias in each of these sub-steps will be surveyed.

3.3.1 Specification bias

We propose the term *specification bias* to denote bias in the choices and specifications of what constitutes the input and output in a learning task, i.e.:

- The features in the vectors x_i in Eq. 2, for example 'income', 'property magnitude', 'family status', 'credit history', and 'gender' in a decision support system for bank loan approvals.
- The output y in Eq. 2, for example 'approve' as target variable. Sometimes the choice of target variable involves creation of new concepts, such as 'creditworthiness', which adds extra bias.
- In the case of categorical features and output, discrete classes related to both x and y, for example 'low', 'medium', and 'high'.

These specifications are typically done by the designer of the system, and require good understanding of the problem, and an ability to convert this understanding into appropriate entities [9]. Unintentionally or intentionally biased choices may negatively affect performance, and also systematically disadvantage *protected classes* in systems building on these choices [3].

Related to the selection of features, the notion of *proxies* deserves some comments. When designing a decision support system, one approach to prevent bias with respect to a protected attribute, such as race, is to simply remove race from the features used for training. One problem with this approach is that the result may still be biased with respect to race, if other features are strongly correlated with race and therefor act as *proxies* for race in the learning [14, 45]. Proxies for race could, for example, be area code, length, and hairstyle.

3.3.2 Measurement bias

In epidemiology, *Measurement bias*, *Observational bias*, and *Information bias* refer to bias arising from measurement errors [42], i.e. errors occurring in the process of making observations of the world. In the reviewed material on bias and machine learning, such bias was rarely mentioned, although this process can be biased in very many ways. In epidemiology and medicine, the data gathering process is central, and the Dictionary of Epidemiology [39] lists 37 different types of biases that may influence data. While most of the listed biases are specific for medicine and epidemiology, we identified the following fundamental types of measurement related bias that are highly relevant also for machine learning. *Bias due to instrument error* is a 'Systematic error due to faulty calibration, inaccurate measuring instruments, contaminated reagents, incorrect dilution or mixing of reagents, etc.'. *Observer bias* is defined as 'Systematic difference between a true value and the value actually observed due to observer variation'. The related *investigator bias* is defined as 'Bias on the part of the investigators of a study toward a particular research result, exposure or outcome, or the consequences of such bias'. Hence, a measurement bias can occur either due to the used equipment, or due to human error or conscious bias.

3.3.3 Sampling bias

Sampling bias occurs when there is an underrepresentation or overrepresentation of observations from a segment of the population [15]. Such bias, which is sometimes called *selection bias* [8], or *population bias* [35], may result in a classifier that performs bad in general, or bad for certain demographic groups. One example of underrepresentation is a reported case where a New Zealand passport robot rejected an Asian man's eyes because 'subject eyes are closed'[7]. A possible reason could have been that the robot was trained with too few pictures of Asian men, and therefor made bad predictions on this demographic group.

There are many reasons for sampling bias in a dataset. One kind is denoted *self-selection bias* [15] and can be exemplified with an online survey about computer use. Such a survey is likely to attract people more interested in technology than is typical for the entire population and therefor creates a bias in data. Another example is a system that predicts crime rates in different parts of a city. Since areas with more crimes typically have more police present, the number of reported arrests would become unfairly high in these areas. If such a system would be used to determine the distribution of police presence, a viscous circle may even be created [11, 41].

An opposite example demonstrates how the big data era with its automatic data gathering can create 'dark zones or shadows where some citizens and communities are overlooked' [12]. The author Kate Crawford points to *Street Bump*, a phone app that uses the phone's built in accelerometer to detect and report information about road problems to the city. Due to the uneven distribution of smartphones across different parts of the city, data from Street Bump will have a sampling bias.

It is important to note that sampling bias does not only refer to unbalanced categories of humans, and furthermore not even to unbalanced categories. Unbalances may also concern features that have to appear in a balanced fashion. One example is given in [46], and is there denoted *dataset bias*. Focusing on image data, the authors argue that '... computer vision datasets are supposed to be a representation of the world', but in reality, many commonly used datasets

[6] Wikipedia:Neutral point of view. Accessed July 29, 2020.

[7] CNN World, Dec. 9, 2016. Accessed July 29, 2020.

represent the world in a very biased way. Objects may, for example, always appear in the center of the image. This bias makes it hard for a classifier to recognize objects that are not centered in the image [46]. The authors compared six common datasets of images used for object detection, and found that performance on another dataset than the one used during training in average was cut to less than half. A similar effect is reported in [2]. If all images in a dataset containing a snowmobile also contain snow, a machine learning algorithm may find snow cues useful to detect snowmobiles. While this may work fine for images in the dataset used for training, it becomes problematic to analyze images with snowmobiles placed indoors.

Another kind of sampling bias is *survivorship bias* [15]. It occurs when the sampled data does not represent the population of interest, since some data items 'died'. One example is when a bank's stock fund management is assessed by sampling the performance of the bank's current funds. This leads to a biased assessment since poorly-performing funds are often removed or merged into other funds [32].

3.3.4 Annotator bias

Annotator bias refers to the manual process of labelling data, e.g. when human annotators assign 'approve' or 'do not approve' to each y_i to be used to build a classifier for approval of loan applications [43]. During this process, the annotators may transfer their prejudices to the data, and further to models trained with the data. Sometimes, labelling is not manual, and the annotations are read from the real world, such as manual decisions for real historical loan applications. In such cases bias rather falls into the category historical bias (see Section 3.2).

3.3.5 Inherited bias

It is quite common that tools built with machine learning are used to generate inputs for other machine learning algorithms. If the output of the tool is biased in any way, this bias may be inherited by systems using the output as input to learn other models. One example is if the output of a smile detector based on images is used as input to a machine learning algorithm. If the smile detection is biased with respect to age, this bias will propagate into the machine learning algorithm. We suggest the term *inherited bias* to refer to this type of bias. The authors of [44] identify a number of Natural Language Processing tasks that may cause such inherited bias: *machine translation, caption generation, speech recognition, sentiment analysis, language modelling,* and *word embeddings*. For example, tools for sentiment analysis have been shown to generate different sentiment for utterances depending on the gender of the subject in the utterance [27]. Another example is word embeddings, which are numerical vector representations of words, learned from data [33, 38]. Word embeddings are often used as input to other machine learning algorithms, and usually provide a powerful way to generalize since word embeddings for semantically close words are close also in the word vector space. However, it has been shown that several common word embeddings are gender biased. The authors in [5] show that word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. For example, in the embedding space, the word 'nurse' is closer to 'female' than to 'male'. The authors in [7] propose a method called WEAT (Word Embedding Association Test) to measure such bias. Two sets of so-called *target words* (e.g. *programmer, engineer, scientist & nurse, teacher, librarian*) and two set of so-called *attribute words* (e.g. *man, male & woman, female*) are considered. The null hypothesis is that there is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words.

Methods that reduce this kind of bias in word embeddings have been suggested, and either modify already trained word embeddings [5] or remove parts of the data used to train the embeddings [6]. However, bias may still remain [18] after applying these methods, and may propagate to models generated by other machine learning algorithms that rely on word embeddings as input.

4 ASSESSING MODEL BIAS

The result from an inductive learning process, i.e. the function f^* in Eq. 1), is often referred to as a 'model'. As described in the previous sections, bias may propagate from the biased world, through a biased data generation, to the learning step with its inevitable inductive bias and other biases. The observed bias of a resulting model is often simply denoted 'bias' [8, 10, 11, 21]. To distinguish this from other types of bias discussed in this paper, we propose using the term *model bias* to refer to bias as it appears and is analyzed in the final model. An alternative would be the existing term *algorithmic bias* [13]. However, typical usage of that term usually refers to the societal effects of biased systems [36], while our notion of bias is broader. Nevertheless, most suggestions on how to define model bias statistically consider such societal effects: how classification rates differ for groups of people with different values on a *protected attribute* such as race, color, religion, gender, disability, or family status [21]. As we will see in the following, classification rates may differ in very many respects, and a large number of bias types have been defined based on the condition that should hold for a model not to have that particular type of bias. For a binary classifier we can for example require that the *overall misclassification rate* (OMR) is independent of a certain protected attribute A (that takes the values 0 or 1). The corresponding condition for a classifier not being biased in this respect is [49]:

$$P(\widehat{Y} \neq y|A=0) = P(\widehat{Y} \neq y|A=1), \tag{3}$$

where $\widehat{Y}$ is the classifier output $f(x)$ (see Eq. 2), and y is the correct classification for input x. Both $\widehat{Y}$ and y take the values 0 or 1. For example, the fact that a person is female ($A = 0$) should not increase or decrease the risk of incorrectly being refused, or allowed, to borrow money at the bank. Several similar conditions can be defined to describe other types of unwanted bias in a classifier model [49]:

false positive rate (FPR):

$$P(\widehat{Y} \neq y|A=0, y=0) = P(\widehat{Y} \neq y|A=1, y=0), \tag{4}$$

false negative rate (FNR):

$$P(\widehat{Y} \neq y|A=0, y=1) = P(\widehat{Y} \neq y|A=1, y=1), \tag{5}$$

false omission rate (FOR):

$$P(\widehat{Y} \neq y|A=0, \widehat{Y}=0) = P(\widehat{Y} \neq y|A=1, \widehat{Y}=0), \tag{6}$$

false discovery rate (FDR):

$$P(\widehat{Y} \neq y|A=0, \widehat{Y}=1) = P(\widehat{Y} \neq y|A=1, \widehat{Y}=1). \tag{7}$$

Each one of these equations focuses on that an incorrect ($\widehat{Y} \neq y$) classification should be independent of A and a specific value of $\widehat{Y}$ or y. The advantageous classifier output (for example being accepted a loan) is here coded as 1. A classifier that does not satisfy one of these equations is said to be biased in the corresponding sense[8]. For example, a classifier is biased with respect to FDR if the value of A affects the probability of incorrectly being allowed to borrow money.

A related condition is the *equalized odds*, which appears in the literature with slightly different definitions (see [21] and [31]). In [21], equalized odds is defined by the following two conditions (slightly modified notation):

$$P(\widehat{Y} = 1|A = 0, y = 0) = P(\widehat{Y} = 1|A = 1, y = 0), \quad (8)$$

and

$$P(\widehat{Y} = 1|A = 0, y = 1) = P(\widehat{Y} = 1|A = 1, y = 1). \quad (9)$$

Note that Eq. 8 is equivalent to FPR in Eq. 4, and Eq. 9 is equivalent to TPR in Eq. 5.

Several other indicators of model bias have been proposed. Loftus et al. [31] define *Calibration, Demographic Parity/Disparate Impact*, and *Individual Fairness*. For a binary classification $\widehat{Y}$, and a binary protected group A, demographic parity is defined as follows:

$$P(\widehat{Y} = 1|A = 0) = P(\widehat{Y} = 1|A = 1). \quad (10)$$

That is, $\widehat{Y}$ should be independent of A, such that the classifier in average gives the same predictions to different groups. If the equality does not hold, this is referred to as *disparate impact*. An example is a software company that wants to reach a better gender balance among their, mainly male, programmers. By following the principle of demographic parity, when recruiting, the same proportion of female applicants as male applicants are hired.

Taken all together we conclude that there is a large number of different types of model biases, each one with its own focus on unwanted behavior of a classifier. Furthermore, many of these biases are related, and it can also be shown that several of them are conflicting in the sense that they cannot be avoided simultaneously [10, 28, 49]. Hence, it is problematic to talk about 'fair' or 'unbiased' classifiers, at least without clearly defining the meaning of the terms. It can also be argued that a proper notion of fairness must be task-specific [16].

5 TOWARDS A TAXONOMY OF BIAS

In this section we summarize and discuss the various notions of bias found in the survey, and propose a taxonomy, illustrated in Figure 1.

5.1 Terminology

While it used to be the case that 'Bias in machine learning' usually referred to the inductive bias we describe in Section 3.1, this is no longer the case. As the survey shows, there is a multitude of usages with different meanings of bias in the context of machine learning. We summarize our proposed taxonomy in Figure 1, with different types of biases organized in the three categories *A biased world, Data generation*, and *Learning*. In several cases the meaning of terms differed between surveyed papers, and in some cases specific and important types of biases were only referred to as 'bias'. In these cases, we propose descriptive names.

In the Biased world category, the main term is *historical bias*. We identify five named types of historical bias. If we define bias as things that 'produce outcomes that are not wanted' [45], this list could of course be made considerably longer. We suggest the term *co-occurrence bias* for cases when a word occurs disproportionately often together with certain other words in texts (see Section 3.2).

In the Data generation category, we found five types of sources of bias. This list should also not be taken as complete, but rather as containing some of the most common and representative examples used in the literature. Several sub-types were also identified (see Section 3.3). We propose the term *specification bias* to denote bias in the specifications of what constitutes the input and output in a learning task (see Section 3.3.1), and we suggest the term *inherited bias* to refer to existing bias in previously computed inputs to a machine learning algorithm (see Section 3.3.5).

In the Learning category, we have the classical inductive bias, but also what we name *hyper-parameter bias*, the bias caused by, often manually set, hyper-parameters in the learning step (see Section 3.1).

We propose using the term *model bias* to distinguish the bias detected in the outcome of a machine learning system, from the possible reasons for this bias. Specific remarks concerning model bias are presented below.

5.2 On model bias

A wast majority of published research refer to social discrimination when talking about bias in machine learning. A typical, and frequently discussed, example of such model bias is COMPAS, a computer program used for bail and sentencing decisions. It has been labeled biased against black defendants [26] [9].

Model bias is caused by bias propagating through the machine learning pipeline. Bias in the data generation step may, for example, influence the learned model, as in the previously described example of sampling bias, with snow appearing in most images of snowmobiles. This may cause an object classification algorithm to use irrelevant features as shortcuts when learning to recognize snowmobiles (in this case snow cues) [2]. This in turn leads to a classifier that is biased against snowmobiles placed indoors, and biased for snowmobiles placed outdoors. While this, at first, may not be seen as a case of social discrimination, an owner of a snowmobile shop may feel discriminated against if Google does not even find the shop's products when searching for 'snowmobiles'.

In our survey we identified nine aspects of model bias, defined by statistical conditions that should hold for a model not being biased in a specific way. In addition, several causal versions exist. Some of the identified conditions are contradictory such that any attempt to decrease one bias will increase another. This is not totally surprising since the conditions are related to common performance measures for classifiers, such as precision and recall, which are known to have the same contradictory relation [22, pp. 405]. The contradictory conditions is not a statistical peculiarity, but a very real phenomenon. The COMPAS system mentioned above is indeed biased by certain conditions, but fair by others[10,11] [1].

In some cases, certain types of bias violates intuitive notions of fairness, and may even be prohibited by law. One example is demographic parity (Eq. 10), which aims at classifiers with the same predictions to different groups. As noted in [31], this may require

[8] In practise, the requirement is usually that the left and right hand side of the equation should be approximate equal.

[9] *Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica May 23, 2016. Accessed Feb. 10, 2020.

[10] *False Positives, False Negatives, and False Analyses. Community Resources for Justice*. Accessed Feb. 15, 2020.

[11] *A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear*. Washington Post Oct. 16, 2017. Accessed Feb. 15, 2020.

positive discrimination, where individuals having different protected attributes are treated very differently. In some cases, this may be a consciously chosen strategy to change societal imbalances, for example gender balance in certain occupations. However, it would probably not be seen as a good idea to apply the same reasoning to correct arrest rates for violent crimes, where men are significantly overrepresented as a group.

Given this complex situation, one should view the different aspects of model bias as dimensions of a multi dimensional concept. They should, together with traditional performance measures, be selected, prioritized and used to guide the design of an optimal, albeit not necessarily statistically 'unbiased' machine learning system. As noted in [10], '... it is important to bear in mind that fairness itself ... is a social and ethical concept, not a statistical one'.

Most used notions of model bias share a fundamental shortcoming: they do not take the underlying causal mechanism that generated data into account. This is serious not least since the legal system defines discrimination as an identified causal process which is deemed unfair by society [21]). Furthermore, the importance of causality in this context is widely recognized among ethicists and social choice theorists [31]. Unfortunately, correlations between observed entities can alone not be used to identify causal processes without further assumptions or additional information. Several researchers have recently developed causal approaches to bias detection. A causal version of equalized odds, denoted *Counterfactual Direct Error Rate*, is proposed in [50], together with causal versions of other types of model biases. Causal versions of additional types are suggested in [21,31]. Due to space constraints we will not discuss these further, although causal reasoning is seen as critical both for identification and reduction of model bias.

5.3 The world as it should be vs. the world as it is

It is important, but not always recognized, that most statistical measures and definitions of model bias, such as Equations 3-9, use the correct classifications y as baseline when determining whether a model is biased or not. If y are observations of humans' biased decisions in the real world (such as historical loan approvals), or humans' biased manual labels created in the data generation process, Eq. 3 could be perfectly satisfied, which may be interpreted as the model being free of bias (with respect to overall misclassification rate). However, a more correct interpretation would be that the model is no more, or less, biased than the real world. Assessing the 'true' degree of biasedness of a model, requires a notion of an ideal 'world as it should be', as opposed to the observed 'world as it is'. Demographic parity (Eq. 10) has such a notion built in, namely that the classifier output should be independent of the protected attribute.

There are at least two fundamentally different approaches to address the problem with a biased model. We may debias the computed model, based on an understanding of what 'the world as it should be' looks like. For example, word embeddings may be transformed such that the distance between words describing occupations are equidistant between gender pairs such as 'he' and 'she' [5]. Another approach to address biased models is to debias the data used to train the model, for example by removing biased parts, such as suggested for word embeddings [6], by oversampling [17], or by resampling [30]. Debiasing input data can be seen as a technical introduction of (good) bias in the data generation process, but it can also be seen as an attempt to model an ideal 'world as it should be' rather than the biased 'world as it is'.

The distinction between these two worlds is related to when a specific model is useful or not. If the model is going to be used to predict 'the world as it is', model bias may not be problem. Such a model may, for example, be used to predict whether a given loan application will be accepted or not by the bank. Good predictions should model also biased decisions made by the bank. On the other hand, if the model is going to be used in a decision support system, we may want it to mimic 'the world as it should be', and bias is then highly relevant to detect and avoid in the design of the system. One example of how 'the world as it should be' is chosen as norm, is Google's image search algorithm. Since only 5% of Fortune 500 CEOs were women (2018), a search for 'CEO' resulted in images of mostly men. Since then, Google has reportedly changed the algorithm to display a higher proportion of women [45].

5.4 What's wrong with discrimination?

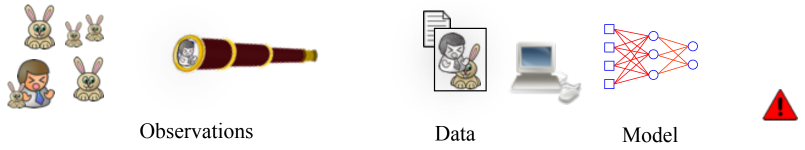
The necessity of inductive bias in machine learning was mentioned in Section 3.1. The same holds at the level of human learning, as discussed in the area of philosophical hermeneutics [23]. In [20] the author argues that we always need some form of prejudice (or bias) to understand and learn about the world. Returning to the example in Section 3.1, a decision support system for approval of bank loans is sometimes described as biased and discriminating if it treats certain groups of people differently. It is important to realize that this difference in treatment, in a general sense, is inevitable and rather the main purpose of such a decision support system: to approves some people's applications, and reject others[12]. For example, it may be the bank's policy to not approve applications by people with very low income. While this technically is the same as rejecting people based on ethnicity, the former may be accepted or even required, while the latter is often referred to as 'unwanted' [21], 'racial' [43], or 'discriminatory' [10, 37] (the terms *classifier fairness* [10, 16, 49] and *demographic parity* [21] are sometimes used in this context). The difference between features such as 'income' and 'ethnicity' has to do with the, already cited, normative meaning of the word bias expressed as 'an identified causal process which is deemed unfair by society' [8]. This is further reflected in the notions of *protected groups* and *protected attributes* [21], which simply define away features such as 'income', while including features that are viewed as important for equal and fair treatment in our society.

5.5 Fighting bad bias with good

With the possible exception of inductive bias, the various types of biases described in this paper are usually used with negative connotations - to describe unwanted behavior of a machine learning system. However, several of the types of biases described are not necessarily bad. For example, some kind of specification bias is necessary to setup a machine learning task. The alternative would be to observe everything observable in the real world, which would make learning extremely hard, if not impossible. The choice of features to include in the learning constitute a (biased) decision, that may be either good or bad from the point of view of the bias of the final model.

Likewise, annotator bias is usually regarded as a bad thing, where human annotators inject their prejudices into the data, for example by rejecting loan applications in a way that discriminates members of a certain demographic group. However, there is of course also a possibility for the human annotators, to consciously or unconsciously,

[12] In machine learning this general ability to distinguish between varying input data is even called 'discrimination', but without any negative connotations (see for example [29]).



A BIASED WORLD → DATA GENERATION → LEARNING → EVALUATION

Historical bias:

- *Co-occurrence bias*
- Framing bias
- Epistemological bias
- Language bias
- Coverage bias

…

- *Specification bias*
- Annotation bias
- Measurement bias
- Sampling bias
- *Inherited bias*

…

- Inductive bias
- *Hyper-parameter bias*
- Uncertainty bias

…

Model bias:

- Overall misclassification rate
- False positive rate
- False negative rate
- False omission rate
- False discovery rate
- Equalized odds
- Calibration
- Demographic Parity
- Individual Fairness

+ Causal versions

…

Figure 1. Proposed taxonomy of the different types of bias that appear in the machine learning pipeline. Starting with existing bias in the world, biased observations are made, and data is generated in a, possibly biased, process. The data is then used by a biased learning mechanism to produce a model. The bias of the model is finally evaluated. When necessary for clearness or completeness, we propose new terms (here written in italics).

inject 'kindness' by approving loan applications by the same members 'too often'. Depending on the context, this could be described as a good annotator bias.

Increasing the inductive bias in the learning step can even be shown to be a general way to reduce an unwanted model bias. Imposing requirements on f, such as Eq. 3, can be expressed as constrained minimization [49] in the inductive learning. Eq. 1 may be rewritten as

$$f^* = \underset{\substack{\text{s.t. } f \in \Omega, \\ P(f(x) \neq y | A=0) = \\ P(f(x) \neq y | A=1)}}{\arg\min} \sum_{i=1}^{N} (f(x_i) - y_i)^2. \tag{11}$$

While the minimization problems 1 and 11 seem to be identical, the latter is unfortunately much harder to solve. The constraints are non convex, as opposed to the normal concave case which can be solved by several efficient algorithms. The authors in [49] approximate the additional constraints such that they can be solved efficiently by convex-concave programming [48].

However, the imposed requirements on f can also be seen as unconstrained minimization over a restricted function space Ω'

$$f^* = \underset{f \in \Omega',}{\arg\min} \sum_{i=1}^{N} (f(x_i) - y_i)^2, \tag{12}$$

where Ω' is the original Ω, with all functions not satisfying the imposed requirements removed. Hence, in order to decrease unwanted (bad) model bias, we increase the inductive (good) bias by restricting the function space Ω appropriately.

6 FINAL REMARKS

Our survey and resulting taxonomy show that 'bias' used in conjunction with machine learning can mean very many different things, even if the most common usage of the word refers to social discrimination in the behavior of a learned model. Even this specific meaning of the word deserves careful usage, since it comes in a variety of types that sometimes even contradict each other. Regarding bias in the steps leading to a model in the machine learning pipeline, it may or may not influence the model bias, in a sometimes bad, and sometimes good way.

A final remark is that humans are deeply involved in all parts of the machine learning process illustrated in Figure 1: the biased world, the data generation process, the learning, and the evaluation of bias in the final model. *Cognitive biases* are systematic, usually undesirable, patterns in human judgment and are studied in psychology and behavioral economics. They come in a large variety of shades, and the Wikipedia page[13] lists more than 190 different types. Only a small number of them are directly applicable to machine learning, but the size of the list suggests caution when claiming that a machine learning system is 'non-biased'.

REFERENCES

[1] Kristin Bechtel Anthony W. Flores and Christopher T. Lowenkamp, 'False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks"', *Federal Probation Journal*, **80**(2), (September 2016).

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh, 'Learning de-biased representations with biased representations', *ArXiv*, **abs/1910.02806**, (2019).

[3] Solon Barocas and Andrew D. Selbst, 'Big data's disparate impact', *California law Review*, **671**, (2014).

[4] Hadrien Bertrand, *Hyper-parameter optimization in deep learning and transfer learning : applications to medical imaging*, Theses, Université Paris-Saclay, January 2019.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai, 'Man is to computer programmer as woman is to homemaker? debiasing word embeddings', in *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*, (2016).

[6] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel, 'Understanding the origins of bias inword embeddings', in *Proc. of the 36th Int. Conf. on Machine Learning*, (2019).

[13] Wikipedia *List of cognitive biases*. Accessed July 29, 2020.

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

[8] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford, 'Ai now 2017 report', in *AI Now 2017 Symposium and Workshop*, (January 2018).

[9] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas Reinartz, C. Russell H. Shearer, and Robert Wirth, *CRISP-DM 1.0: Step-by-step data mining guide*, SPSS, 2000.

[10] Alexandra Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', *Big data*, **5 2**, 153–163, (2016).

[11] Ignacio N. Cofone, 'Algorithmic discrimination is an information problem', *Hastings Law Journal*, **70**, 1389–1444, (2019).

[12] Kate Crawford, 'Think again: Big data', *Foreign Policy*, (May 2013).

[13] David Danks and Alex John London, 'Algorithmic bias in autonomous systems', in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, p. 4691–4697. AAAI Press, (2017).

[14] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen, 'Proxy non-discrimination in data-driven systems', *CoRR*, **abs/1707.08120**, (2017).

[15] Rice University David M. Lane, 'Chapter 6 research design - sampling bias', in *Online Statistics Education: A Multimedia Course of Study*, 235, Rice University.

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, 'Fairness through awareness', in *Proc. of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, p. 214–226, New York, NY, USA, (2012). Ass. for Comp. Machinery.

[17] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel, 'Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.', in *Int. Conf. on Learning Representations*, (2019).

[18] Hila Gonen and Yoav Goldberg, 'Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them', in *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (June 2019).

[19] Bryce Goodman and Seth Flaxman, 'European union regulations on algorithmic decision-making and a "right to explanation"', *AI Magazine*, **38**, 50–57, (2017).

[20] Gadamer Hans-Georg, *Truth and Method*, Continuum, 1975.

[21] Moritz Hardt, Eric Price, and Nati Srebro, 'Equality of opportunity in supervised learning', in *Advances in Neural Information Processing Systems 29*, eds., D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 3315–3323, Curran Associates, Inc., (2016).

[22] Aboul Ella Hassanien, Aboul Ella Hassanien, and Tarek Gaber, *Handbook of Research on Machine Learning Innovations and Trends*, IGI Global, USA, 1st edn., 2017.

[23] Mireille Hildebrandt, 'Privacy as protection of the incomputable self: From agnostic to agonistic machine learning', *Theoretical Inquiries in Law*, **20**(1), 83–121, (2019).

[24] Christoph Hube and Besnik Fetahu, 'Detecting biased statements in wikipedia', in *WWW'18: Companion Proceedings of the The Web Conference 2018*, pp. 1779–1786, (2018).

[25] Christoph Hube, Besnik Fetahu, and Robert Jäschke, 'Towards bias detection in online text corpora', in *International Workshop on Bias in Information, Algorithms, and Systems (BIAS)*, number 2103 in CEUR Workshop Proceedings, pp. 19–23, Aachen, (2018).

[26] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner, 'Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks', *ProPublica*, (May 2016).

[27] Svetlana Kiritchenko and Saif M. Mohammad, 'Examining gender and race bias in two hundred sentiment analysis systems', *ArXiv*, **abs/1805.04508**, (2018).

[28] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan, 'Inherent trade-offs in the fair determination of risk scores', *CoRR*, **abs/1609.05807**, (2016).

[29] Tim Kovacs and Andy J. Wills, 'Generalization versus discrimination in machine learning', in *Encyclopedia of the Sciences of Learning*, ed., Norbert M. Seel, 1352–1355, Springer US, Boston, MA, (2012).

[30] Yi Li and Nuno Vasconcelos, 'Repair: Removing representation bias by dataset resampling', *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9564–9573, (2019).

[31] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva, 'Causal reasoning for algorithmic fairness', *CoRR*, **abs/1805.05859**, (2018).

[32] Burton G. Malkiel, 'Returns from investing in equity mutual funds 1971 to 1991', *The Journal of Finance*, **50**(2), 549–572, (1995).

[33] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', *CoRR*, **abs/1301.3781**, (2013).

[34] Tom M. Mitchell, 'The need for biases in learning generalizations', Technical report, Rutgers University, New Brunswick, NJ, (1980).

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

[36] Trishan Panch, Heather Mattie, and Rifat Atun, 'Artificial intelligence and algorithmic bias: implications for health systems', *Journal of global health*, **9**(2), 010318–010318, (12 2019).

[37] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini, 'Discrimination-aware data mining', in *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD '08, p. 560–568, New York, NY, USA, (2008). ACM.

[38] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, 'Glove: Global vectors for word representation', in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, (2014).

[39] *A Dictionary of Epidemiology*, ed., Miquel Porta, Oxford University Press, USA, 2008.

[40] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky, 'Linguistic models for analyzing and detecting biased language', in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 650–1659. ACL, (2013).

[41] Rashida Richardson, Jason Schultz, and Kate Crawford, 'Dirty data, bad predictions: How civil rights violations impact police data', *Predictive Policing Systems, and Justice*, (February 2019).

[42] K.J. Rothman, S. Greenland, and T.L. Lash, *Modern Epidemiology*, Wolters Kluwer Health/Lippincott Williams & Wilkins, 2015.

[43] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith, 'The risk of racial bias in hate speech detection', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678, Florence, Italy, (July 2019). Association for Computational Linguistics.

[44] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang, 'Mitigating gender bias in natural language processing: Literature review', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1630–1640. Association for Computational Linguistics, (2019).

[45] Harini Suresh and John V. Guttag, 'A framework for understanding unintended consequences of machine learning', *ArXiv*, **abs/1901.10002**, (2019).

[46] A. Torralba and A. A. Efros, 'Unbiased look at dataset bias', in *Proc. of the 2011 IEEE Conf. on Computer Vision and Pattern Recognition*, CVPR '11, p. 1521–1528, USA, (2011). IEEE Comp. Soc.

[47] Claudia Wagner, David García, Mohsen Jadidi, and Markus Strohmaier, 'It's a man's wikipedia? assessing gender inequality in an online encyclopedia', in *ICWSM*, (2015).

[48] Shen Xinyue, Diamond Steven, Gu Yuantao, and Boyd Stephen, 'Disciplined convex-concave programming', in *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 1009–1014, (2016).

[49] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi, 'Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment', in *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pp. 1171–1180, (2017).

[50] Junzhe Zhang and Elias Bareinboim, 'Equality of opportunity in classification: A causal approach', in *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, (2018).

[51] Xu Zhang, Xiaocong Chen, Lina Yao, Chang Ge, and Manqing Dong, 'Deep neural network hyperparameter optimization with orthogonal array tuning', in *ICONIP*, (2019).

[52] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang, '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*, pp. 2979–2989, (2017).