An investigation of social Biases in Supervised Machine Learning

A thesis presented in partial fulfilment of the requirements for the degree of Bachelor of Science

Ashwin van Gool

S.B. Degree Candidate in Artificial Intelligence

Supervised by Niels van Miltenburg and Dong Nguyen

Faculty of Humanities, Utrecht University

Utrecht, The Netherlands

April 13, 2020

7.5 ECTS

Contents

1 Understanding Machine Learning **1**

1.1 Applications in society 1

1.2 Origins 3

1.3 A New Approach 4

1.4 Learning 5

1.5 Verification 8

1.6 Example: The loan problem 9

2 On Bias **11**

2.1 Towards a definition 12

2.2 Exploring Bias 13

2.3 When is Bias Problematic? 15

2.4 Causes of Bias 19

2.5 Conclusion 26

Abstract

The accelerated development and rapid societal integration of machine learning over recent years has left insufficient time for considerations of widespread reliance on these techniques with regards to social and societal consequences. This thesis provides an investigation into bias in supervised machine learning, assessing its facilitating role in societal injustice and identifying root causes. The first chapter serves as an inspection into the current state of supervised machine learning with regards to societal impact. Furthermore, basic technicalities and concepts within supervised machine learning are discussed. The second chapter presents an examination of the concept of bias itself, along with an assessment of the problemacy of bias and identification of key causes.

Three main issues concerning societal usage of machine learning arise: Preservation of both individual recognition and equal treatment are important, as is the right allotment of credibility. Four root causes in bias in supervised learning have been identified: Usage of unrepresentative data, propagation of bias, adopted bias and bi-enhancing bias. It seems that conscious collection of data, complete, inclusive curating of datasets, and thoughtful development of algorithms are the best methods to minimizing bias. Awareness will better allow for the curating of balanced, complete representative datasets, as well as the conscious development and deployment such algorithms.

Acknowledgements

I would like to take the time to thank my first supervisor Dr. Niels van Miltenburg. He's been of great help, both in person as well as online during the corona crisis, in structuring and shaping this thesis. Second, I'd like to thank second supervisor Dong Nguyen for providing feedback and suggestions. General thesis coordinator Dr. Janneke van Lith has also provided me with useful suggestions and tips along the way. Finally, special thanks goes to Kathleen de Boer for keeping me motivated and focused, as well as proofreading most of the work.

Sincerely, Thanks.

Introduction

Machine learning has opened up possibilities that seemed impossible only decades ago, and its potential applications are deemed to be of unrivaled magnitude. But with great power comes great responsibility. For all the incredible utility machine learning might bring, greater societal reliance on it could be problematic, or even dangerous. Elon Musk, one of the founders of the renowned OpenAI, remarked that building safe and fair artificial intelligence will be this generations' greatest challenge (Friedman, 2016).

The widespread incorporation of machine learning into almost all facets of society has seen a steady increase over the past decade, and shows no signs of slowing down. Furthermore, the rapid technical developments within the field, fueled by a race between a variety of companies and countries to stay ahead of the curve, has left other important topics such as ethics by the wayside (Lyam, 2020; Sharma, 2018). Yet, an increasing societal reliance on these techniques means that any risks with regards safety and fairness should be kept to a minimum, especially if machines are to be placed in a position of being stronger, faster more trusted or smarter than humans (Bostrom & Yudkowsky, 2011).

Machine learning has already made an profound impact on elementary procedures within commerce, justice, business, science and law enforcement, among others. Through these mediums, machine learning directly and indirectly influences matters of societal significance. This thesis is concerned with *bias*, a relevant issue that encompasses systemic, discrimina-

tory deviations within systems utilizing machine learning. The central question throughout this thesis is *What is bias in supervised machine learning, and how does it occur?*.

A firm grasp of core concepts within machine learning is necessary in order to fully understand and evaluate the problem of bias. The first chapter was put in place to provide this, in addition to illustrating societal relevance. The second chapter examines bias itself, providing an investigation into the nature of bias, as well as touching on its problemacies and causes.

Hopefully, this thesis will serve as an instrument to raise both understanding and awareness of bias.

Chapter 1

Understanding Machine Learning

The aim of this chapter is to provide an insight into societally impactful applications of machine learning, as well as a basic overview of basic supervised learning concepts, so as to act as a solid basis of understanding required for the later parts.

1.1 Applications in society

Some general purpose applications of machine learning include but are not limited to image, video and face recognition, text processing and understanding, trait and phenomenon prediction as well as outlier and anomaly detection. Various forms of such applications have been integrated into many facets of society, as well as our daily lives. This section provides a selection of some examples of socially and societally influential applications.

Let us start at Google's search engine. Machine learning stands at the core of multiple systems involved here: It is used in understanding the content and context of query text, rank-order the pages shown to users based on prior interests as well as in RankBrain, special software for rare or hard queries (Schachinger, K. 2017). Sites that provide visual content, such as Youtube or Facebook, widely use image recognition for a multiplicity of reasons,

such as detecting unwanted content or even assisting blind users (Wu, S. & Wieland, J. 2016). Services that offer a wide range of creative content, such as Netflix and Spotify, almost universally utilize machine learning-based recommendation systems to suggest new content (Chong, D. 2020).

An increasingly greater amount of sectors and enterprises have taken substantial interest in the possibilities offered by machine learning. The commercial sector is now dominated by machine-learning governed personal advertising, adaptive pricing, chatbots and content personalization (Chiu, J. 2019). Some large companies such as Amazon have even started experimenting with using machine learning to hire new employees (Dastin, J. 2018). Healthcare has adapted the techniques to improve risk on diagnoses, patient risk identification and even to detect tumors in scans (Kourou, K. 2015). The applications in the financial sector range from stock prediction to assessing loan and mortgage worthiness to fraud detection (Cheung, K. 2020). Scientific research too, has benefitted from machine learning in solving complex problems such as protein folding or discovering new medicine (Hassabis et. al. 2020). Justice systems have found new tools in assisted verdicts and recidivism risk assessment (Miller, 2020). Some divisions of law enforcement use *predictive policing*, which is the premature deployment of police forces in areas that are predicted to have higher risk of crimes (Khan et. al. 2019). Such social control is taken to an extreme by certain governments, monitoring citizens in real time through an extensive net of cameras equipped with facial recognition (Campbell, C. 2019). Some schools have even started using such tools to monitor student's attention and emotional state (Chan, T.F. 2018).

The widespread social impact of machine learning is thus enormous. Some of the most important aspects of our lives, such as civil liberties, financial well-being and employment are increasingly governed by algorithms that are inherently opaque due to their sheer complexity. With such usage of and reliance on these techniques, one could now start to see how

any related disadvantages should be of serious societal concern. Yet, the high rate at which the field is developing and the techniques are adapted leaves insufficient time for paramount considerations regarding topics such as privacy, fairness and equality of opportunity. This lag is also reflected in most bodies of law, which currently fail to properly account for the so-called 'datafication' of society (Tricoles, R. 2019). The relevancy of such concerns becomes apparent in cases such Amazon's hiring algorithm being biased against women, or Google Photo's tendency to label people of color as 'Gorillas' (Dastin, J. 2018; Grush, L. 2015). These topics will further be covered in chapter 2. First, the remaining sections will cover the basic approach and principles of supervised machine learning.

1.2 Origins

The term *machine learning* was coined by Arthur Samuel in a paper that investigated ways in which computers could be taught to play checkers. The paper described an implementation of an algorithm that learned to improve by observing its own games, which in time allowed it to beat national level players. At this point in time where most programs were written explicitly to perform a certain task, Samuel observed that '*Programming computers to learn from experience should eventually eliminate the need for this much detailed programming effort*' (Samuel, 1959).

It was not until years later that another computer science pioneer by the name of Tom Mitchell provided a formal definition: '*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E*' (Mitchell, 1997). Note how both scientists emphasized the experience-based nature of machine learning. It was this fundamentally different approach to problem solving that would go on to characterize the

field within computing sciences.

1.3 A New Approach

Where conventional programs and algorithms explicitly prescribe a computer what to do, machine learning algorithms specify how the computer can learn using sample data instead. This learning could be thought of akin to how a human would learn: By appropriately adjusting its behaviors or beliefs based on experiences and observations. In this parallel, data is to computers what experience is to humans. The behaviour-adjusting incorporation of this data would then be the equivalent of learning.

Conventional programs and algorithms are thus explicitly programmed to perform a certain task. A fundamental limitation to this approach is captured in Polanyi's paradox, which states that *human knowledge and understanding exceeds their ability to explicate it.* This has major consequences for conventional software development, as it implies an upper limit to the problem solving potential of such software problems. Furthermore, within the context of artificial intelligence, it means that computational devices programmed this way could never truly rival human understanding (Vardi, 2016).

A second limitation stems from the inherent complexity of some problems. This complexity makes it infeasible to develop conventional programs to perform these tasks. An illustrative example is found in *Computer vision.* Computer vision is the field that aims to enable computers to interpret and understand the meaning behind visual imagery. As it turns out, this problem is incredibly complex. So complex in fact, that explicitly programmed classification algorithms fail miserably on tasks that humans would consider trivial, such as recognizing symbols or discriminating between images of trains and planes.

Machine Learning largely circumvents the problems mentioned above. Because these algorithms are instructed how to learn on their own, explication of human knowledge is not required. This ability to learn, combined with their ability to represent complex non-linear decision spaces through advanced features such as sigmoid functions, hidden layers or the kernel trick enables unrivaled complex problem solving potential. And it shows: A specialized deep neural network known as ResNet was able to obtain an error rate of 3.51 percent as early as 2015 in the annual ImageNet Challenge for image recognition (He et. al. 2016). By comparison, human error rate was estimated to be 5.1 percent (Russakovsky et. al. 2015).

1.4 Learning

By understanding by which principles machines learn, it might be understood how biases can emerge or be propagated. Machine learning is commonly divided into three categories: reinforcement learning, unsupervised learning, and supervised learning. Reinforcement learning is a form of action-outcome learning, in which the result of every chosen action is evaluated and future behaviour is adjusted accordingly. Both supervised and unsupervised learning are concerned with learning underlying patterns in data, in which the former uses data with some identified classification while the latter does not. This thesis is concerned with supervised learning only. The reasons for this are twofold: The first being that this method is commonly used in applications with direct social impact. The second reason is that it is prone to incorporating preventable biases (Losifidis et. al. 2018).

Supervised learning is used for classification problems and regression problems. Problems whose solution can be represented by a discrete value are known as classification problems. For banks, the decision whether or not to lend a loan is a classification problem in which the solution takes the value of either 0 (NO) or 1 (YES). Another classification problem is the

recognition of handwritten letters, in which the solution space would consist of all integers inclusively between 0 (A) and 25 (Z). All decision and categorisation problems can thus be converted to an instance of a classification problem, in which each possible solution is in some way represented by an integer.

Likewise, problems with a continuous value solution as their solution are known as regression problems. Intuitively, these can be thought of as the problem of predicting some quantity. A prime example of a regression problem is the prediction of house prices. The notorious COMPAS system also uses regression-based analysis. It accounts for a variety of factors to assess the defendant's risk of recidivism. This type of regression could be thought of as a complex variant on linear regression and is used for problems that require the estimation or prediction of some value.

Supervised learning by definition exclusively utilizes labeled data. *Labeled* denotes a pairing between the problem's input and its respective output. That is, each data point incorporates one or more input values as well as their corresponding output value. Intuitively, one could think of this as presenting the algorithm with some specific instance of the problem as well the solution to that instance. The complete collection of data points is referred to as the *data set*.

X_1	X_2		Y
1	*3*		*5*
4	*1*		*33*
3	*5*		*23*
2	*3*		*11*

Table 1.1: Example of a small labeled data set

Table 1.1 presents a simple example of a labeled data set.The first $n = 2$ columns, commonly numbered as X_n for $n \in N$ denote the input variables, also known as *features*. These features are collectively referred to as the input. The third column, denoted by Y, gives the

corresponding output. Each row, which also contains output in the labeled case, represents a *data point*. Note that data sets used in practice might consist of thousands of data points, each supplying hundreds of input variables.

Formally, the machine learning algorithm is now tasked with finding some function that, given an input, correctly classifies or estimates the corresponding output value. This could be thought of as estimating the relationship between problem instances and their solutions. In feed-forward neural networks, the most commonly used supervised learning technique, this estimation starts out as some arbitrarily chosen function at initialization[1]. This function is then iteratively adjusted in order to better account for the data points utilizing a process called back-propagation[2]. This is the phase where the actual learning takes place, and is referred to as *training* the network. The intricacies of this process are mathematically complex and not directly relevant to understanding later topics, and have therefore been omitted. Reconsider table 1.1. The outputs in the table were generated by the *true* function F presented below.

$$F : Y = 2X_1^2 + X_2$$

In practical applications of machine learning when F is unknown, the algorithm attempts to find a function F' so that $F' \approx F$. In other words, F' approximates the underlying pattern through which the data was generated. The full range of F' is referred to as the *model*. The application of supervised learning lies in then using F' to solve new instances of the problem. That is, if such a function F' is obtained, it could then be used to approximate solutions to instances of the problem for which only input is known. Had the algorithm come up with the true function during training on the data in table 1.1, so that F = F', and we were given $X_1 = 1$ and $X_2 = 2$, we might predict that the corresponding output should be 2. This

[1]Taken strictly, this is incorrect. Rather, the initial function is a result of an arbitrary allocation of weight values to nodes in the network. However, this explanation suffices for the purpose of explanation and avoids unnecessary complexity.

[2]Again, this is slightly incorrect. Rather, after each pass through the network that leads to an incorrect classification, weights of the nodes involved are adjusted, which directly affects the function.

process of extrapolating solutions to unknown instances based on known instances is known as *generalization*. Generalization could be thought of as akin to inductive reasoning in the sense that a limited amount of observations are used in order to infer general truths, and is a key process in machine learning. Generalization however, also proves to fulfill a key role in the origination or propagation of bias, as will be shown in chapter 2.

1.5 Verification

Now that an approximation is made, a second problem emerges: how does one verify that $F' \approx F$? That is, how does one know that the obtained approximation is representative of the real underlying function? The second major phase of developing supervised learning algorithms is verifying that the obtained function F' generalizes well. That is to say, F' also accurately predicts outputs for data not used in training. This phase is known as *testing*. It is common practice to, prior to training, randomly split the data set into two parts: the *training set* and the *holdout set*[3]. The holdout set usually accounts for about 20 percent of the data, which leaves the remaining 80 percent for the training set. During testing, the accuracy of F' is verified by presenting it with the unlabeled versions of holdout set data. Since the labeling on this data is known, the accuracy of the model can then be calculated as the percentage of instances in which the prediction of F' and the original labeling agree[4]. When the model's accuracy on the holdout set approaches the model's accuracy on the training set, the model is said to generalize well[5]. Good generalization is very important, as it verifies that the model didn't do well on the training data by accident, or because it

[3]Sometimes, a third part known as the *validation set* is added. The validation set may be used when the algorithm sports additional adjustable parameters, such as learning rate. These are then calibrated using the validation set prior to testing. This is not directly relevant to the purposes of this thesis and has therefore been omitted.

[4]This is the accuracy measure for classification problems. Accuracy is calculated slightly different for regression problems, but the idea is the same.

[5]Unless F' = F and there is no noise, which never happens in practice, the model will also make classification errors on the training data as well as the holdout data.

picked up on patterns in the training data that are not true in the general case.

It is important to note that the example presented in table 1.1 is not fully representative of data used in practice, as it was generated by a *noiseless* mathematical function. *Noise* refers to accidental inaccuracies in the the data, and is almost omnipresent. Data used in practice is seldom perfect due to measurement errors, human judgement errors and other inaccuracies. The true function, stripped from noise, is referred to as the signal. A high signal to noise ratio is one of the characteristics of high-quality data, as this will allow the algorithm to learn true patterns in the data rather than also having to account for deviations caused by noise. Accounting for these deviations caused by noise is referred to as *overfitting*. Overfitting is problematic as it causes bad generalization, since the algorithm learns patterns that are only true for training data that will not hold in the general case. This may cause the algorithm's outputs to be unrepresentative of reality. Thus, proper verification is another key step in understanding and preventing bias.

1.6 Example: The loan problem

To bring it all together, consider a classical application of supervised learning: the loan problem. In this decision problem, a bank is given some multitude of values such as *age*, *gender*, *current income* and *loan history* about their client, and should decide based on these factors whether or not to loan them money. Banks used to employ dedicated human experts for this. Nowadays, most banks use machine learning to make these decisions. The banks' record of supplied loans can be used as a data set in which the client's characteristics form the input, and whether their loan was successfully paid back determines the output. A well trained supervised learning algorithm could then be employed to predict whether new clients will be able to successfully pay off their loans.

Although such algorithms, among many other applications of machine learning, are commonplace in the banking world nowadays, they are not perfect. Every client is different, and no amount of information is sufficient for perfect predictions. One might even envision a situation in which two clients input values are very comparable, yet only one of them was able to pay off the loan. Such occurrences make it impossible to correctly classify all data. Machine learning is thus not without its limitations and restrictions.

Chapter 2

On Bias

The presumption that computers or algorithms are inherently neutral is a dangerous one, especially now that their footprint on society is greater than ever. If left unchecked, *bias* might prove to be a root cause in social injustice and societal destabilization (Binns et. al. 2017). This chapter provides an investigation into the emergence and nature of bias in supervised learning.

Prior to further reading, i would like to make a pair of remarks. First off, the term *bias* may be used to refer to different kinds of bias: A first, fundamental kind that arises purely through inconsistencies and stochastic noise in the training data and a second, more conceptual kind that arises through the learning or propagation of concepts within incomplete or unrepresentative data. The first kind is inherent to generalization and thus machine learning itself, and computer scientists have come a long way in minimizing such bias (Hüllermeier, 2014). It is of little further interest for the purposes of this paper. Instead, this thesis is concerned with the second, conceptual kind that seems to be causally related to systemic injustices and harmful stereotyping. This will further be discussed in the rest of chapter 2.

Second, it should be noted that the term *to discriminate* in essence means *to make a distinction*. Thus, any preference for one value over another is discriminatory. This means

in turn that there is no such thing as non-discriminatory machine learning: There should always be factors on which a classification or prediction is made. In this sense, discrimination is therefore an essential part of machine learning. That is not to say there are no grounds on which should not be discriminated. The right question is: What are the right and wrong grounds on which this distinction should be based?

2.1 Towards a definition

Bias was defined by Tom Mitchell to be *any basis for choosing one generalization over another, other than strict consistency with the observed training instances* (Mitchell, T. 1980). However, it seems this definition falls short for the purposes of this thesis. It seems reasonable to assume that whenever machine learning is used in a social context, the predictions or decisions made by the algorithm should reflect the society's core social values, such as those captured in legislation. This means that any algorithm which directly or indirectly influences social matters should act in accordance to these captured values. So, at the very least, these values should include protected classes such as age, sex, race, sexual orientation, religion or national origin, as citizens are protected by law against discrimination based on such factors. Yet, we might imagine a situation in which an algorithm's outcomes are in strict consistency with the training data, but such standards do not hold, leaving Mitchell's definition insufficient to properly account for such socially discriminatory bias. Suppose some company employs a machine learning algorithm to decide who to invite for a job interview, and further suppose this algorithm was trained on previous human decisions. Would these decisions have been biased, in regards to gender for example, the algorithm might learn this bias. This is where Mitchell's definition falls short for the purposes of this thesis: In theory, the generalization is indeed in strict consistency with the observed training instances. It might be clear however, that the kind of discrimination resulting from such a scenario would

not be acceptable in any egalitarian society.

This disparity between Mitchell's definition and such discriminatory bias is thus captured in the fact that what we do and do not perceive as bias is partly dependent on what a given society perceives to be just and unjust grounds for discrimination. What would then exactly count as just or unjust within this context is a complicated philosophical and social manner that is beyond the scope of this thesis. For the sake of simplicity, the assumption is made here that the set of unjust discriminatory grounds should at least include protected factors. This is why I decided to expand on Mitchell's definition. We could say that bias refers to discrimination based on *unjust* grounds, where unjust refers to *inconsistency with regards to some combination of the observed data and the ethical framework within which it operates*. Such data or ethics based grounds are usually interwoven in a variety of ways when it comes to social or societal applications of supervised learning, depending on the way in which the bias arose in the first place (Corbett-Davies & Goel, 2018). The definition presented above will be used for the remainder of this thesis.

2.2 Exploring Bias

Before further examination of problemacy and causality, there are some more considerations concerning bias to be made.

Implicit and Explicit Bias

One might wonder why it would be hard to prevent bias in the first place. After all, it is known what the values that serve as the inputs to a supervised learning algorithm represent. For instance, if a recidivism risk scoring system such as COMPAS was found to exhibit racist behaviour, could we not just prevent it from considering *race* as an factor? Unfortunately, it is not that simple. In fact, race was never an input to begin with, nor were

any directly related factors (Brennen et. al. 2009). This is where an important distinction comes into play. If an algorithm were to directly consider some unwanted factor and use this in prediction, we refer to this as *explicit* bias. This sort of bias would indeed be easily preventable: Just stop using it as an input factor, as this will avert the algorithm from explicitly considering the factor. The challenging form of bias is *implicit* bias. An algorithm is implicitly biased if it does not directly account for some factor through direct input, yet it exhibits discriminatory behaviour on that same factor in regards to its output. This was the case in the COMPAS system for recidivism risk, which was trained on seven factors, including criminal record and age of first offense. Gender, race, age or other protected factors were never taken into consideration, as would be forbidden by law. Yet, independent research found the algorithm labeled black defendants to be a high risk score twice as often as would be expected when looking at the actual re-offending rates (Angwin et. al. 2016). Similarly, white defendants were found to be labeled low risk scores way less often compared to what the eventual re-offending rates would end up being. The possibility of these skewed results seems staggering, especially considering that race or directly related variables were not taken into consideration. It may thus be clear that the problem of bias is not as simple as it might seem on first glance. This problem is further examined in section 2.3.

Fairness and Base Rates

What exactly constitutes unbiased or fair within the realm of socially related machine learning is a complex problem. A number of fairness metrics have been proposed (Binns, 2017). Such metrics all embody the idea that similar people should be treated similarly. A second plausible restriction to the fairness of a system is that it should be reflective of reality. That is, they are in accordance to the true base rates (Zemel et. al. 2013). There is a delicate balance in obtaining fairness that makes for a hard but interesting problem. If one is to be concerned about fairness within a classifier system like machine learning, there is an a priori incentive to minimize bias in such a system. While enforcing fairness through

some chosen metric might seem like a good idea, it turns out to not be. In fact, I will argue below that enforcing equality through base rates or other means in itself introduces bias.

Base rates denote statistical means over groups that represent some aspect of that group. For example, when considering the aspect of biological sex over the entire population, it is found that the base rate of women in the world is 49.75 percent. When considering bias, it is important to keep base rates in mind. Consider a scenario in which some engineering company's employees are for ninety percent male. An obvious case of bias, it might be presumed. That is until it is considered that engineers are overwhelmingly male to start out with. One might then presume that the disparity between groups should itself be objectionable, and that the bias is in the base rates to begin with. It might be argued however, that inequality should not be objectionable as long as it stems from equality of opportunity. That is, as long as the disparity results from individual, unforced and fair choices (Segall, 2012).

The greatest employment disparities between groups are actually to be found in the freest and most egalitarian nations such as the Scandinavian countries (Sanandaji, 2018). This makes sense: if general preferences exist between groups and choices are made freely, group disparities will inevitably arise. For example, if boys and girls were given a free choice between a blue or a pink toy pet, a preference disparity between these groups would very likely occur. As such disparities thus arise despite, or sometimes even because of unbounded general preferences between groups, i would argue that merely a difference between groups should not be objectionable. Furthermore, group differences and even bias seem to be a necessity in certain learning situations (Mitchell, 1980). But when should objections arise?

2.3 When is Bias Problematic?

In this section, the problemacy inherent to bias is examined and discussed.

Failure of Equal Treatment

It might be argued that the greatest societal danger of biases is in their potentially unjust discriminatory nature. What does it mean to be unjustly discriminated against? Unequal treatment might be evaluated in a variety of ways. A compelling definition is given by Kusner's counterfactual scenario (Kusner et. al. (2017), which is akin to how we might intuitively classify equal treatment. Here, a system is considered discriminatory with regards to some variable in proportion to how altering that variable might influence the final decision or prediction. For example, a system would be considered fair in regards to the variable *gender* if a man and woman with otherwise exactly equal characteristics would obtain the exact same treatment by that system.

It is of course exactly this specification that is not met within a biased system, as this is in effect the very definition of bias: *Discrimination based on unjust grounds.* Thus, biased systems per definition fail this specification of equal treatment. The objections to usage of biased systems in social matters might thus seem trivial, as doubtlessly, such systems will serve in the promotion or maintenance of inequality of opportunity. Segall argued that discrimination is bad for this and only exactly this reason, as the undermining of equality of opportunity is both unfair and sub-optimal for parties involved (Segall, 2012). This seems like a right assessment to me. For example, would some employer unjustly discriminate on some non-relevant factor like gender, this deprives both the employer as well as the employee of an optimal situation in which the best suited candidate, regardless of gender, is selected for the job. In a similar vein, biases in justice assistance systems might lead to an unfair and sub-optimal scenario in which the very goal of justice systems, bringing righteousness, is jeopardized. As true fairness and optimality were to be goals of introducing automatized assistance, bias is then a root of counterproductivity to such ideals.

Failure of Individual Recognition

Generalization is an integral part of supervised learning, and indeed more generally, learning. Generalization is the very strength of machine learning, as it allows for the of prediction of characteristics that would likely be present in new instances. The very nature of generalization is in opposition to handling individuality, as it inherently concerned with patterns that hold in general cases. That is, it is concerned with identifying the characteristics that are found generally, and are thus not unique. Note that this is in exact opposition to individuality, which is defined by the unique characteristics of an individual. Sometimes, the removal of the individualistic element is not problematic when using machine learning. For example, in recommendation systems such as Netflix's, where new series are recommended based on individual watch history and the trends in watch history of other users, there is nothing wrong with accounting solely for general trends, as there is nor the impact nor need for individual treatment. In fact, it seems likely that most users use such recommendation systems not despite but because of this service, as it allows them to be find series that are new yet approximately suit their liking. However, there are definitely instances in which the gravity of the decision's impact or the nature of the decision might mandate that a individualistic stance towards the problem is taken. This might be the case for hiring, where a fair review based on individualistic traits should arguably be a higher priority than any other. Regarding algorithmic assistance in judicial systems, in which decisions are of great impact and should differ case by case, i would argue that there should not be a great reliance on generalized trends and information. Of course, jurisprudence should still be taken into account. It is just that the accuracy of jurisdiction for any case should not be blurred due to generalization.

Opacity and Overestimation of Credibility

It might be argued that the problems considered above are not necessarily inherent to machine learning itself. As noted, failure of equal treatment is a complex problem with multiple

roots, some of which are beyond direct control, and failure of individual recognition is inherent to all generalizations. Rather, these problems' social impacts are uncovered and exacerbated by the rise of machine learning. Where this might prove more problematic is through the misplaced confidence that is sometimes put into technology that is deemed advanced. The sometimes-made assumption that computerized learning and prediction is more reliable than human judgment is very dangerous. The sheer complexity of machine learning does not necessarily translate to success, especially in careless hands. Reconsider the COMPAS software that was used state-wide in the USA in courts to determine recidivism risk. This algorithm came under scrutiny after claims that it was implicitly severely biased against people of color. This was not the only problem, however. Further research showed that the algorithm was just not very good at its only task. (Dressel & Farid, 2018) In this research, the algorithm's predictions were compared to prediction of humans without any background in justice whatsoever. These human participants were given the exact same data and were asked the exact same question: based on this data, do you predict this defendant's risk of recidivism to be low or high? The human novices did not do particularly well, achieving an accuracy rate of about 69 percent. However, the algorithm did even worse, with an accuracy rate of 67 percent. As it turns out, judges would be better off incorporating random stranger's judgments into their verdicts, rather than a high tech state of the art algorithm. Yet, these and similiar algorithms are still in use in justice systems in multiple states in the USA (Jackson & Mendoza, 2020). Furthermore, as of yet, there is another important aspect to decision making that is lacking on the end of learning algorithms, which is the ability to point out their exact, articulated reasons for making decisions or predictions. Usually, the sheer complexity and lack of clear meaning for any particular inner component within a supervised machine learning network means that not even human experts are consistently able to do this. However, some strides have recently been made on this front (Caelli & Bischof, 2013).

It may be clear that misplacement of credibility might lead to unwanted scenarios in which systems are relied on for the sole reason of being technologically advanced, rather than actual results to back up this confidence. Furthermore, the opaque, 'black box' nature of supervised learning means that it is ill-suited for any purposes that require clearly justified reasoning. Machine learning is a very useful and powerful technique, but should be handled with care. Awareness of this fact seems crucial in the justifiable, correct and fair application of supervised machine learning.

2.4 Causes of Bias

The problem of bias is very complicated, not less so due to the variety of ways in which it might arise. This section analyses and lists major underlying causes of biases.

On Causes

Biases in supervised learning generally stem from inaccuracies in or unrepresentativeness of training data. This is not unexpected, as data is the most essential and influential component of learning and therefore of model construction. However, bias might stem from a range of sources for a variety of applications. I have made a distinction into four kinds of causes, which will be presented below. The issue of *unrepresentative data* is the most general case in which the dataset as a whole is unrepresentative of reality in some way. In the more specific case of using human-based data, the propagation of *human bias* leads to the issue of individual data points might be skewed. *Adopted bias* is a process-generated form of bias in which the continual modification of the system itself through new data leads to biases. Lastly, the problem of *positive feedback loops* arises as another process-generated form of bias in which stored information and algorithmic predictions enhance one another, spiraling to unbalanced proportions. Bias is complex problem and there might be additional roots that I missed or that will arise in future applications of supervised learning. Nevertheless,

these seem to form an exhaustive list for now of general and more specific applications of supervised learning biases.

Unrepresentative Data

The most general way in which bias might arise is through the usage of unrepresentative data to train the algorithm. This could be thought of as a imbalance in the data leading to an imbalance in the model. To better understand this concept, let's start with an example. In the 1936 presidential election of the United States of America, Alfred Landon was pitted against Franklin D. Roosevelt. In order to predict who would win, a magazine known as *Literary Digest* conducted one of the largest polls in human history, sporting more than 2.4 million respondents. The magazine's results estimated that Landon would receive 57 percent of the votes. It was calculated that based on these results and sample size, Landon's victory was all but guaranteed. When the actual results came in they showed a landslide victory, but not for Landon: A substantial majority of votes had gone to Roosevelt. So what went wrong? The problem had been the way the data was collected. The conductors of the polling had used the telephone directory to create a mailing list that was used for gathering responses. In 1936, telephones were still a luxury item and thus used almost exclusively by the middle and upper class. Thus, despite the huge sample, the poll had been very unrepresentative of the general population. Their prediction actually was very representative for middle and upper classes, but had completely failed to generalize by not accounting for the lower class voters, which had overwhelmingly voted for Roosevelt (Ozbey, 2018).

Although this example does not involve direct application of machine learning, the parallels might be clear. This example was chosen because it is very insightful into how bias might occur. In the same way that usage of unrepresentative data led *Literary Digest* to make some seriously wrong predictions, usage of unrepresentative data will lead supervised

learning algorithms to be biased towards the instance that is over-represented. Data seems to be unrepresentative in this way when there is a significant imbalance between data and reality. In this case, the dataset was severely under-representative of lower class households or, in converse, over-representative of upper class households. This unrepresentativeness of training data will inevitably lead to bad generalization, as learned patterns will fail to hold for the entire population.

The 1936 presidential election far from the only example of bias based on unrepresentative data: Many voice recognition systems were found to react better or more consistently to male voices than female ones (Bajorek, 2019). Examples of 'smart' soap dispenser systems that failed to recognize hands with darker skin tones also surfaced on internet (Hale, 2017). Such biases in all likelihood stem from the under-representation of females or people of color in training and testing environments. Not even IBM's jeopardy-winning supercomputer known as Watson escaped this bias. IBM's researchers had attempted to make it speak in a more human-like manner by incorporating the Urban Dictionary, a dictionary for internet and street language, into its vocabulary. However, the severe over-representation of offensive and swear words made Watson speak in a rather vulgar way, which had the researchers initiate a complete reset (Falk, 2013).

Indeed, such biases stem from imbalances that are unrepresentative of reality in the dataset. This kind of bias seems to be the unfortunate and hopefully unintentional byproduct of naivety or carelessness. In order to maximize chances of obtaining data that is generally representative, statistical mathematics prescribes the usage of a large data set that is sampled truly random over the population. This might not always be possible in practice, however. Recall the loan problem: a practical machine learning problem in which banks assess loan worthiness. Since data is only gained here in situations in which the loan was approved, data collection automatically fails for anyone that did not receive initial approval. Thus, it is

impossible to consider the full range, as it will never be found out whether those clients would indeed default or not. As such, the algorithm's predictions might skew towards approval as the majority of data it is trained on will inherently be biased through this selection.

Learning Human Bias

For some applications, algorithms will be tasked with the emulation of cognitive tasks classically performed by humans. Consequently, these algorithms will be trained on previous human decisions and predictions. This is problematic, as these human decisions are almost invariably biased in themselves. Furthermore, good intent might not suffice to prevent these biases, due to the phenomenon of *unconscious bias*. Unconscious bias is formed by the *immediate, automatic associations that tumble out before we've even had time to think* (Gladwell, 2005). Like all human biases, they usually arise as a heuristic *to quickly discern who is enemy and who is friend, for in the past — and certainly in many places in the world today — the ability to quickly identify friend or foe may be a matter of life or death* (Begley, 2004). The existence of human biases is thus definitely understandable and at times excusable. Other times however, this might not be the case. It would for example be highly desirable for some employer to hire not based on their preferences with regard to race or gender, but rather competence and other relevant traits. In fact, this is enforced by anti-discrimination laws in most countries. Enforcing protection of discrimination on these protected grounds is a completely different manner without straightforward solutions. Any decisions made due to conscious biases held by the employer might be explained away. For example, one could always find excuses or make up fake reasons to not hire a female applicant rather than admitting their sexist motives. Even with the right intentions, there is the issue of unconscious bias. It was found that around 90 percent of Americans exhibit racial bias to some extent, although many consider themselves to be unbiased (Begley, 2004). This non-introspective nature is characteristic of unconscious bias and further complicates the problem, as it implies that not even conscious and good-willed attempts to prevent biases might suffice to do so.

Machine learning is inherently concerned with learning through the discovery of underlying patterns. This is a definite upside, as the combination of data processing and sheer computational power might allow for observations, predictions and pattern finding that would be impossible for humans or conventional programs to make or do. This does mean however, that any biases that were underlying in human decision making data, whether consciously or unconsciously, will in all likelihood be propagated through training on such data. This means that preventing bias in supervised learning algorithms based on human decisions is as at least as hard as preventing bias in human data.

The issue of human bias is relevant but complicated. The intricacies of the issue will not be further addressed here, but it is important to note that it seems very hard to fully prevent human bias. One might suppose that different solutions might exist. However, a retroactive approach has major issues, among which the enforcement problem. There is thus no straightforward solution to this problem. One could supposedly average over decisions made by a diverse group of humans in some way. But the objections are of both practical and utilitarian nature. First, it might simply not be possible to do this due to scarcity and costs of collecting a large, diverse but well-suited group. Second, even succeeding to do so is far from a guarantee in preventing bias. In fact, it is not obvious at all whether we could even suppose that biases could 'cancel out' between them. Furthermore, there are a number of psychological fallacies leading to biases that any human is prone to exhibiting. And the problems do not end there, but I will not go in to further depth here due to scope constraints.It may be clear however, that the problem of transitioning between human and computational decision without adapting bias is complicated.

Adopted Bias

In some applications of machine learning, such as natural language chatbots or loan accreditors, it is useful to incorporate sequential learning. This process is known as *incremental learning* and involves constantly updating the system as new data comes in. This is useful for problems such as loan accreditation, in which the incorporation of more or more recent data should in theory lead to better predictions. Similarly, chatbots might learn from their own conversations and consequently adjust their style or formulation. This approach requires additional care, upkeep and constraints to be in place in order to properly work. In March 2016, Microsoft launched an experimental chatbot named Tay. Tay was an artificial intelligence that interacted with other users via general messages known as *tweets* on the social media platform Twitter. The intention was that Tay would learn in real time, and adapt human tweeting behaviour over time. However, Microsoft had failed to put in place adequate constraints to block malicious intent. Immediately after launch, people would send vulgar, racist and toxic tweets at Tay. Consequently, Tay had to be taken offline a mere 16 hours after launch, having devolved to exhibiting offensive behaviour after being exposed to it.

The possibility of acquiring bias in this way is specific to a handful of applications: those that adapt their own behaviour to better match that of their environment. Currently, there is only a limited selection of such applications, mainly consistent of mostly experimentally oriented chatbots and other forms of guided digitized interactivity. Nevertheless, with the current rate of automation and digitization, there might soon be a plethora digitized agents such as robots for all sorts of tasks and interactive purposes. As such, adaptation of bias through user-input sequential learning, which would in all likelihood take on a central role within such applications, should be of concern. Luckily, I presume that little else than awareness and conscious development should be required in combating such bias. The existence and learning of malign contents, which might be intentionally malevolent, give rise to

adopted bias. Although control and restriction should be specific to the platform and purposes, these should largely account for preventing the roots of adopted bias. Had Microsoft put in place appropriate constraints, it is very unlikely that Tay would have ever had to be taken down due to toxic behaviour. Once more, awareness and conscious development thus seem to be key in preventing bias.

Positive Feedback Loops

A second application of machine learning that involves incremental expansion and updating of a system in which databases and algorithms work in tandem. Examples of such systems would be predictive policing, which involves a database of past crime per area, or job hiring, which involves a database of past applicants. This is the only cause of bias discussed here that does not directly arise through training on data. Instead of iteratively changing the model with new data, these applications work off of an already trained algorithm that makes predictions based on constantly updating data in the database. For most other applications of supervised learning, there is a one-sided relation between data and prediction. This means data only influences prediction, and not the other way around. In these database-based systems, this might not always be the case. Instead, both data and predictions influence each other. This may lead to a positive feedback loop that magnifies some aspects while disparaging others, invoking bias into the system.

To see how this works, consider *predictive policing*, which is the act of predicting crime based on trends in the past and using this to influence allocation of law enforcers. The act of prediction is done by algorithms trained through supervised learning. These algorithms factor in past data collected by law enforcement based on aspects of the crimes like place, time, date and type to assign risk predictions of future crimes to areas (Huet, 2015). This is useful as it allows for more effective and efficient law enforcement, as police will be at the right place at the right time more often. As some areas will be more heavily patrolled than

others, naturally, the number of arrests in these areas will increase. Conversely, the number of arrests in lesser patrolled areas will decrease. It is at this point that the positive feedback loop establishes as incorporation of this new data, influenced by these changes, will lead the algorithm to double down on its prediction, which then further polarizes the deployment of ground forces.

It is in this bi-enhancing nature that feedback loops might lead to invoking of bias into the system. Strictly speaking, one could say that this is just a process that leads to unrepresentativeness bias. This might be seen in the hypothetical scenario in which two areas have an equal street crime rate, yet the elevated number of patrols in one of them will lead to more arrests. This skews the obtained data in comparison to the true rates, with unrepresentativeness bias as a result.In practice, there will be a number of inhibiting factors that prevent this bias from straying off too far. Nevertheless, there is good reason to be aware of the possibility of the emergence of positive feedback loops as a root cause to bias.

2.5 Conclusion

This thesis was written with a central question in mind: *What is bias in supervised machine learning, and how does it occur?*. There are some inherent characteristics to both the ways of machine learning itself as well as the nature of bias that make for a complex, multi-faceted problem. Understanding bias in supervised machine learning starts with understanding the processes through which it operates, as well as accounting for the social context. As bias might aid in the maintenance and advancement of social injustice and inequality of opportunity, it should therefore be of great societal concern to minimize biases in current and future applications. Keeping this social aspect in mind, i defined bias to be discrimination based on *unjust* grounds, where unjust refers to *inconsistency with regards to some combination of the observed data and the ethical framework within which it operates.* Four ways in which

bias might arise were identified: Bias may emerge utilizing unrepresentative data, or be propagated through usage of biased human data. Additionally, processes of sequential or bi-enhancing data processing might in themselves lead to biases. While machine learning is an incredibly potent tool, it ought to be wielded with great care. Bias is problematic if it leads to failure of equal treatment or failure of individual recognition. Furthermore, anyone should restrain from using such techniques for their appeal as 'high-tech' or 'advanced' instruments, if there are little or no proven benefits from doing so.

It seems that conscious collection of data, complete, inclusive curating of datasets, and thoughtful development of algorithms are the best methods to minimizing bias. Furthermore, increased awareness of bias on the societal end could greatly aid in better and more considerate deployment of machine learning. Both awareness and understanding will be major factors bringing about a society in which machine learning and fairness go hand in hand. I hope this thesis might contribute in raising awareness and deepening the understanding of this interesting problem called bias.

References

Academic Articles

1. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

2. Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. 3 (3): 210–229.

3. Khan, J. R., Saeed, M., Siddiqui, F. A., Mahmood, N., & Arifeen, Q. U. (2019). PREDICTIVE POLICING: A Machine Learning Approach to Predict and Control Crimes in Metropolitan Cities. University of Sindh Journal of Information and Communication Technology, 3(1), 17-26.

4. Mitchell, T. (1997). Machine Learning. McGraw Hill. p. 2. ISBN 978-0-07-042807-2.

5. Vardi, M. Y. (2016). The moral imperative of artificial intelligence. Communications of the ACM, 59(5), 5-5.

6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

7. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A.

C. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

8. Iosifidis, V., & Ntoutsi, E. (2018). Dealing with bias via data augmentation in supervised learning scenarios. Jo Bates Paul D. Clough Robert Jäschke, 24.

9. Mitchell, T. (1980) .The need for biases in learning generalizations. Tech. rep. CBM-TR-117, Rutgers University. Reprinted in Readings in Machine Learning, J. W. Shavlik and T. G. Dietterich (eds.), Morgan Kaufman, San Mateo, CA, 1990.

10. Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. arXiv preprint arXiv:1712.03586.

11. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, February). Learning fair representations. In International Conference on Machine Learning (pp. 325-333).

12. Segall, S. (2012). What's so Bad about Discrimination?. Utilitas, 24(1), 82-100.

13. Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017, September). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In International Conference on Social Informatics (pp. 405-415). Springer, Cham.

14. Hüllermeier, E. (2014). Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. International Journal of Approximate Reasoning, 55(7), 1519-1534.

15. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

16. Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. Criminal Justice and Behavior, 36(1), 21-40.

17. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). Machine bias. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

18. Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1), eaao5580.

19. Caelli, T., & Bischof, W. F. (2013). Machine learning and image interpretation. Springer Science & Business Media.

20. Huet, E. (2015). Server And Protect: Predictive Policing Firm PredPol Promises To Map Crime Before It Happens. Forbes Magazine.

Resources

1. Friedman, J. (2016). *Interview with Elon Musk from Y Combinator's How To Build The Future series.* Retrieved from https://www.ycombinator.com/future/elon/

2. Lyam, Michael. (2020). *The AI Supremacy.* Retrieved from https://www.technative.io/the-ai-supremacy-who-will-take-the-lead-in-this-global-race/

3. Sharma, Munish. (2018). The Global Race for Artificial Intelligence: Weighing Benefits and Risks.

4. Bostrom, Nick & Yudkowsky, Eliezer (2011). The ethics of artificial intelligence. Cambridge University Press, 17

5. Schachinger, Kristine. (2017). *A Complete Guide to the Google Rankbrain Algorithm.* Retrieved from https://www.searchenginejournal.com/google-algorithm-history/rankbrain/

6. Wu, Shaomei & Wieland, Jeffrey. (2016). *Using Artificial Intelligence to help blind people see Facebook.* Retrieved from https://about.fb.com/news/2016/04/using-artificial-intelligence-to-help-blind-people-see-facebook/

7. Chong, David. (2020). Deep Dive into Netflix's Recommender System. Retrieved from https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48

8. Chiu, Joyce. (2019). *The many business applications of Machine Learning.* Retrieved from https://www.datacamp.com/community/blog/machine-learning-tracks

9. Dastin, Jeffrey. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women.* Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

10. Cheung, K. (2020). *10 applications of Machine Learning in Finance.* Retrieved from https://algorithmxlab.com/blog/applications-machine-learning-finance/

11. Hassabis, D., Jumper, J., Kohli, P., Senior, A. (2020) *Alphafold: using AI for scientific discovery.* Retrieved from https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

12. Miller, Susan. (2020). *AI pulls ahead in recedivism prediction.* Retrieved from https://gcn.com/articles/2020/02/21/ai-vs-humans-recidivism-prediction.aspx

13. Campbell, Charlie. (2019). *How China Is Using "Social Credit Scores" to Reward and Punish Its Citizens.* Retrieved from https://time.com/collection/davos-2019/5502592/china-social-credit-score/

14. Chan, Tara Francis. (2018). *A school in China is monitoring students with facial-recognition technology.* Retrieved from https://www.businessinsider.nl/china-school-facial-recognition-technology-2018-5/

15. Tricoles, Robin. (2019). *Smart tech sprints forward, but the law lags behind.* Retrieved from https://research.asu.edu/smart-tech-sprints-forward-law-lags-behind

16. Grush, Loren. (2015). *Google Engineer apologizes after Photos app tags two black people as gorillas.* Retrieved from https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas

17. Ozbey, Ozan. (2018). *Two lessons of sampling bias* Retrieved from https://medium.com/@ozanozbey not-to-sample-11579793dac

18. Bajorek, Joan Palmiter. (2019). *Voice recognition still has significant race and gender bias.* Retrieved from https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

19. Hale, Tom. (2017). *This Viral Video Of A Racist Soap Dispenser Reveals A Much, Much Bigger Problem.* Retrieved from https://www.iflscience.com/technology/this-racist-soap-dispenser-reveals-why-diversity-in-tech-is-muchneeded/

20. Falk, Tyler. (2013). *Why IBM's Watson learned curse words.* Retrieved from https://www.zdnet.com ibms-watson-learned-curse-words/

21. Jackson, Eugenie & Mendoza, Christina. (2020). *Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not.* Retrieved from https://hdsr.mitpress.mit.ed

22. Sanandaji, Nima. (2018). *The Nordic Glass Ceiling* Retrieved from https://www.cato.org/publicatio analysis/nordic-glass-ceiling