MILLIMAN WHITE PAPER

Developing fair models in an imperfect world

How to deal with bias in AI

Daniël van Dam
Raymond van Es
Jan Thiemen Postema

Artificial intelligence (AI) is increasingly used in data-based decision making as we move away from general rule-based models to machine learning (ML) models. Decisions made by ML models are thought to be better, faster, and more consistent than those arrived at by humans. However, as AI becomes an integral part of our lives, concerns over potentially biased and unfair models are growing. Those concerns can be managed by employing proper development methodologies and continuous human oversight.

Introduction

Can we recognize a criminal by their facial features? That is the question Wu and Zhang attempted to answer in their 2016 research paper titled “Automated Inference on Criminality using Face Images.” Their answer was a wholehearted “yes,” after determining that a ML algorithm could recognize criminals based on their facial features with some degree of accuracy. However, as Wired's Katherine Bailey so aptly pointed out,[1] Wu and Zhang’s results could also be used to prove the criminal justice system is biased against people with certain facial characteristics.

This perspective is something neither the authors nor the peer reviewers considered, which goes to show the level of blind trust we often put in the data we use—all of it collected by humans within the inherently messy and often biased world we live. Moreover, the endless examples of bias in AI applications aren't limited to crime-related problems. Insurance, being the data-hungry industry it is, faces the same challenge.

If this problem can occur in such a seemingly inconspicuous data set, it’s likely to also be commonplace in larger data sets. One study from the George Washington University found that the dynamic pricing algorithms used in Chicago by ride hailing companies Uber and Lyft charged more for trips taken to or from primarily non-white neighborhoods[2]. Both companies denied that their algorithms were biased and promised to investigate these results. Uber also added that there might be a host of reasons why such effects could occur. Nevertheless, this study forced both companies to make a public comment and put them in a bad light[3]. This shows us that the societal pressure to do something about the issue of disparate impact is increasing. It also highlights the importance of addressing such issues, not just internally, but also in communication with the outside world. In the meanwhile, legislation is being drafted[4] and companies such as Amazon and Google have been called out on their use of biased algorithms[5]. We refer to these issues as *algorithmic bias*, which describes an algorithm that treats one group unfairly, such as people of a certain race or gender, compared to other groups. In other words, we must find a way to detect and avoid this algorithmic bias.

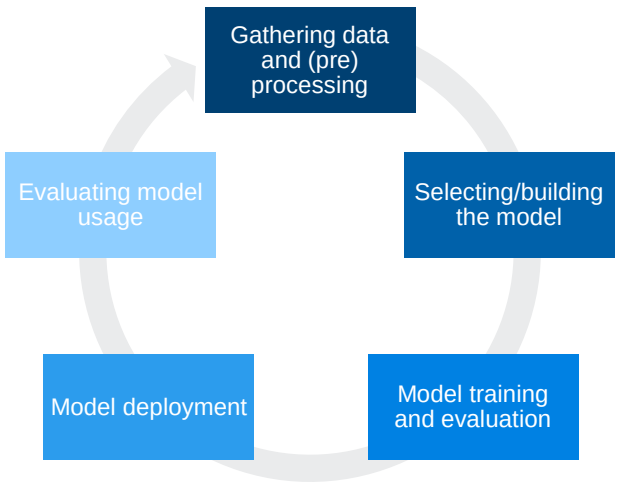
How to detect bias

As mentioned in the introduction, the primary source of bias in models stems from using biased data. In the world of IT there's a saying: "Garbage in, garbage out," meaning that you should always check your input data. The same goes for preventing bias—if the underlying data is biased, the model is likely to be as well. This bias does not have to be explicit, more often than not it’s hidden in other variables such as income or even preferences for music. There are many reasons why data could be biased, which generally fall into two categories: 1) stereotyping, favoritism, and prejudice; and 2) errors in the sampling or reporting procedures. Because data is often viewed as a given in our line of work, we will focus in this paper on the techniques to detect and mitigate biased data.[6]

Detecting bias in data isn't straightforward and can't be condensed into just one step. Instead, we must treat identifying bias as a process that is intertwined with the regular ML lifecycle depicted in Figure 1 below. Out of the five steps in the lifecycle, two are relevant when detecting bias: 1) Gathering data and (pre) processing, and 2) Model training and evaluation.

[1] Put Away Your Machine Learning Hammer | WIRED
[2] Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms | arXiv
[3] Researchers find racial discrimination in 'dynamic pricing' algorithms used by Uber, Lyft, and others | VentureBeat
[4] EUR-Lex - 52021PC0206 - EN - EUR-Lex (europa.eu)
[5] E.g. Amazon scraps secret AI recruiting tool | Reuters, Google ‘fixed’ its racist algorithm - The Verge
[6] Machine Learning Glossary | Google Developers

FIGURE 1: ML LIFECYCLE



In the first step (Gathering data), we must determine what kind of data we are dealing with, where it comes from, and how and for what purpose it was collected. This will help us understand the kind of data we are processing, why we're processing it, and what kind of biases the authors of the dataset could have that may have seeped into the data. Lastly, it's important to identify the sensitive groups in the dataset.

When we're fully aware of the data we're using, we can then take a deep dive into it. The first thing on our list should be to check the dataset for skewness and outliers, which could indicate a reporting bias. Next, we must examine the correlations between variables which we know correspond to a sensitive group (e.g., race or gender) and other relevant variables, as these could create an implicit bias in the resulting model. If, for example, we want to avoid unfairly discriminating on gender, it will most likely not be enough to remove this specific feature from the model as it may have possible correlations with other features.

How to build a fair ML model

The next step in our fight against bias happens during the evaluation phase (Evaluating model usage) of the ML lifecycle, what is called the *Disparate Impact Analysis* (DIA). This umbrella term refers to the primary tool we use to encapsulate many metrics that try to measure whether the model (adversely) affects a sensitive group compared to other groups. Three of the most widely used metrics are *Disaggregated Evaluation, Equality of Opportunity,* and *Demographic Parity*. The first of these metrics separates the dataset into components based on the sensitive groups and evaluates them individually. In an unbiased model, the results for each component should be similar. In the case of equality of opportunity, all else equal, we measure whether the chance of a sample ending up in the positive group would be the same, regardless of the sensitive attribute. In a two-category classification problem, this means that the true positive rate would be equal between groups. Conversely, demographic parity doesn't look at the other attributes, it just measures the chance of a sample ending up in the positive group. In a two-class problem, this is the positive rate.

These metrics are just a few techniques that can be used to evaluate bias. Unfortunately, such methods depend on having access to sensitive variables. Many companies do not currently collect the sensitive data required to calculate the metrics, as it's often undesirable or even unlawful in some cases to store those sensitive characteristics. This lack of a ground truth makes it harder to perform a DIA.

When bias has been detected during the model evaluation, there are several mitigation techniques that can be used to avoid it. The first and most obvious option is to go back to the data and fix the underlying problems. This can be accomplished by collecting more or better data, oversampling underrepresented groups, or by employing techniques such as the *Disparate Impact Remover*, which attempts to remove hidden bias. It does this by editing feature values such that they can't be used to identify sensitive subgroups. Alternatively, the impact of the bias can be minimized by setting different thresholds for different subgroups. For example, if our goal is to predict whether someone will default on their loan we often say: if the chance of defaulting is higher than 50% (i.e., the threshold), we classify that loan as a default. By using different thresholds for different subgroups, we can remove or reduce disparate impact.

Luckily for we practitioners, there are a wealth of tools available that can guide us through the process. Major players such as IBM, Facebook, Microsoft, and Google provide toolsets that implement the techniques we just discussed. Additionally, there are several open-source projects that can help—FairML, debia-ml, and ML-fairness-gym, to name a few.

Lastly, pursuing a robust understanding of the model is important. When we have a clear, overall picture of the model and why certain predictions are made, it is easier to detect and mitigate bias. Explainable AI (XAI) plays an important role here. Certain models, so-called *black boxes*, are not easy to understand, as the importance of each feature is difficult to trace using traditional methods. Thankfully, due to the rising interest and importance of XAI, several methods and tools are now available to help us navigate that process.

Conclusion

Society is becoming increasingly more skeptical about the societal impact of Big Data and ML algorithms. Those developments force us to reconsider how we train ML models. No company intends to purposefully develop a biased model. However, since models are based on datasets collected in the real, imperfect world, bias can seep in unnoticed. Luckily there are multiple techniques available to help us detect potential bias. Even if unintentional bias sneaks into a model, there are several evaluation methods available to mitigate its presence. Of course, this effort is not a one-time occurrence. Model bias should be managed over time and incorporated into model governance procedures. Moreover, these methods merely serve as technical tools. Human oversight is and will continue to be the most reliable safeguard.

Such practices represent a change from the status quo. Instead of developing models that best reflect the world we live in, we should start training models that reflect the world we want to live in. Accomplishing this goal has some obvious drawbacks, especially considering model performance, but it is an essential step in the move to greater equity and fairness in an imperfect world.

CONTACT

Daniël van Dam
daniel.vandam@milliman.com

Raymond van Es
raymond.vanes@milliman.com