TELECOM PARIS

MOD207 Project

Artificial Intelligence's Bias

Realized By :
Chiheb AOULED AHMED BEN ALI

professor :
Winston Maxwell

Supervisor :
Astrid Bertrand

Academic year: 2020/2021

Contents

1 Introduction

Artificial intelligence (AI) and machine learning are advancing at a breakneck pace, with enormous potential benefits. However, if we are to avoid unforeseen, negative repercussions and hazards coming from the adoption of AI in the workplace, we must investigate all ethical, social, and legal elements of AI systems in the society. We are living in an era of artificial intelligence (AI) democratization, in which enterprises and individuals have unprecedented access to AI technologies ranging from deep neural nets (Brynjolfsson et al. 2018) to virtual assistants (Wilson and Daugherty 2018). Increased access has resulted in a dramatic growth in the usage of AI to inform business choices in enterprises, resulting in numerous commercial wins (Davenport and Katyal 2018). However, many businesses have forgotten, in the midst of its fast acceptance, that artificial intelligence, like any other technology, has ethical consequences (Khalil 1993; Martin 2018).

Artificial intelligence's propensity to spread bias and prejudice (Campolo et al. 2017; Koene 2017) is one of the most important ethical challenges (Campolo et al. 2017; Koene 2017), notably against individuals from protected classes such as those of minority race, gender, and national origin (Andreeva et al. 2004). Artificial intelligence has the potential to spread bias and discrimination due to a gap between existing anti-discrimination legislation and the technology's increased capabilities (Barocas and Selbst 2016; Zarksy 2014).Material culture, such as physical equipment or production techniques, in this case AI, advances more quickly than non-material culture, such as ethics, philosophy, belief systems, values, and laws, which has been seen before (Marshall 1999). Cultural lag arises as a result of the difficult process of enacting laws and policies, which lags behind the quick rate of technical advancements (Wellman and Rajan 2017). Unfortunately for businesses, the cultural divide means that following the law isn't enough to prevent discrimination when using AI to make business choices.

This study provides an overview of artificial intelligence's bias, including contemporary applications and definitions.

2 Motivation

The private and public sectors are increasingly turning to artificial intelligence (AI) systems and machine learning algorithms to automate simple and complex decision-making processes.The mass-scale digitization of data and the emerging technologies that use them are disrupting most economic sectors, including transportation, retail, advertising, and energy, and other areas. AI is also having an impact on democracy and governance as computerized systems are being deployed to improve accuracy and drive objectivity in government functions.

In recent years, we've all heard about examples of bias in machine learning (ML) and artificial intelligence (AI), such as Amazon's hiring algorithm that favored men, Facebook's charge of housing discrimination in targeted ads, and this well-known healthcare algorithm that exhibited significant racial bias. It is then essential for the humans to keep the upper hand and not get discriminated because of this bias.

In this study, we will explain at first the different reasons that created this bias.Then we will explain the different methods of reducing this bias or even eliminating it. At last, we will try to use these techniques on a real dataset and study the effects of these techniques on the bias.

3 State of the art: previous literature

The artificial intelligence's bias has been a very common topic and has been referenced by many studies that try to eliminate it.

3.1 Examples of AI bias

We will begin by stating some examples where bias has been demonstrated:

Amazon's biased recruiting tool :
Amazon began an AI project in 2014 with the goal of automating the hiring process. Their initiative was entirely focused on assessing resumes and grading prospects using AI-powered algorithms, allowing recruiters to spend less time on human resume screening. By 2015, however, Amazon had learned that their new AI recruiting algorithm was not fairly grading candidates and was biased against women.
To train their AI model, Amazon used historical data from the previous ten years. Because there was a male dominance in the IT industry and men made up 60% of Amazon's workforce, historical data had biases against women. As a result, Amazon's hiring algorithms wrongly assumed that male applications were preferred. It penalized resumes with the phrase "women's" in them, such as "women's chess club captain." As a result, Amazon no longer uses the algorithm for recruiting.

Racial bias in healthcare risk algorithm :
Because it relied on a poor criteria for establishing the need, a health-care risk-prediction algorithm that affects more than 200 million Americans showed racial bias.

The algorithm was created to predict which patients would require more medical attention, however it was later discovered that the program was giving incorrect findings that favored white patients over black patients.

Previous patients' healthcare spending was utilized as a proxy for medical demands by the algorithm's creators. This was a poor interpretation of historical data because wealth and race are highly associated metrics, and basing assumptions on only one variable of correlated metrics led to erroneous conclusions from the algorithm.

Bias in Facebook ads :
Human bias can be seen in a variety of situations, including tech platforms. These biases result in biased machine learning models since data from tech platforms is later used to train machine learning models.

In 2019, Facebook began enabling advertisers to target advertisements based on gender, race, and religion. For example, women were favored in employment advertisements for nursing and secretarial positions, whereas janitors and taxi drivers were largely advertised to men, particularly those from minority backgrounds.
As a result, Facebook will no longer allow employers to specify age, gender or race targeting in its ads.

3.2 Causes of bias

According to the previous studies there are 3 main issue types that can lead to the bias of the artificial intelligence algorithms:

- **Presentation Problem :**
  A company's real goal can be to target its advertising to the people who are most likely to buy its goods. Because there is no simple method to convert this into an AI implementation , businesses must choose which hypothesis, input attributes, and training labels or reinforcement criteria would best achieve this goal. For example, a video game firm might believe that their product will sell best to young males, so they look for clients who are male and between the ages of 15 and 25.For this problem there exist 3 main causes for the bias:

  - **Proxy Goals**: Basing the goal choice on historical information without factoring in suitable context will necessarily expose the system to historical bias
  - **Feature Selection**: The choice of which attributes to include can lead to the AI bias.
  - **Surrogate Data**: converting a non numerical feature into a numerical one

- **Dataset Problem :**
  There may be issues with training or production datasets, in addition to dataset issues that can cause problems during the mapping step. Creating training sets is a difficult task. It usually includes cleaning up a huge data set and, in the case of supervised learning, obtaining labels. It may include ensuring that unusual cases are proportionally over-represented in deep learning systems to offer the model the opportunity to learn such cases during training. Creating a training data collection can be the bulk of the labor necessary in AI systems and is frequently the source of difficulties due to its volume, complexity, and occasionally timeliness.The main causes of this problem are :
  - Unseen Cases
  - Mismatched Data Sets
  - Manipulated Data
  - Unlearned Cases
  - Non-Generalizable Features

- **Individual Samples problem :**
  Individual sample issues are ones that can be seen by looking at the data from a single sample. The issue could be isolated to that sample or endemic to all of them. When data sets contain personal information, this classification is critical since the complete data collection cannot be made publicly available, but people may be able to access their own personal data.
  There are 2 main causes of this problem :
  - **Inaccurate Data**: To guarantee that the model learns properly, the data utilized for training is frequently heavily chosen. Unfortunately, real-world data is rarely so spotless. It could be missing or corrupted. Data that is manually entered has the potential to be input improperly. Automatically acquired data can have erroneous origins.
  - **Stale Data**: It's possible that the data utilized for both training and production input is outdated. This is especially true if huge "dictionaries" are stored for quick access. Credit scores, for example,

might be downloaded from a third-party site and saved locally for quick access. Unfortunately, developers may be hesitant to update the dataset because it will reset the baseline for a current training trial.

4 Exploration of AI bias(Example on the US Homicide Reports dataset)

In this part of the report,we will describe our work an the notebook attached to it. In our work we used a python library named "aif360" (artificial intelligence fairness 360) in which we used some the different methods to facilitate the manipulation of the AI bias.

4.1 Visualisation of the Artificial intelligence's bias

In our dataset we can see clearly in the notebook that we have a massive dataset issue where the race,sex and country are not well distributed.It is then very obvious that the algorithm will have a huge bias for the most frequent category. In the first part of the preparation of the data, we did some visualisations and frequency counting to examine our dataset and to make the bias very clear.

4.2 Creating a new dataset with AIF360

With the library AIF360 we can create a new dataset that can correct the problem that we had in the first place to control the feature that we want to mitigate the bias from.In our example we chose the "Perpetrator Sex".We then recreated the model with this new dataset.

4.3 Checking the bias fairness with the different metrics

In this part we checked if the model is biased with utilization of 5 different metrics to check if any of these metrics is detected for our model which checks if our model is biased or not. So with aif360 we have some metrics that indicate if our data or model are bias. These are the different metrics:

- Statistical Parity Difference
- Equal Opportunity Difference
- Average Absolute Odds Difference
- Disparate Impact
- Theil Index

4.3.1 Statistical Parity Difference

This measure is based on the following formula :

$$Pr(Y = 1|D = unprivileged) - Pr(Y = 1|D = privileged)$$

Here the bias or "statistical imparity" is the difference between the probability that a random individual drawn from unprivileged is labeled 1 (so here that he has more than 50K for income) and the probability that a random individual from privileged is labeled 1. So it has to be close to 0 so it will be fair.

4.3.2 Equal Opportunity Difference

This metric is just a difference between the true positive rate of unprivileged group and the true positive rate of privileged group so it follows this formula :

$$TPR_{D=unprivileged} - TPR_{D=privileged}$$

Same as the previous metric we need it to be close to 0.

4.3.3 Average Absolute Odds Difference

This measure is using both false positive rate and true positive rate to calculate the bias. It's calculating the equality of odds with the next formula :

$$\frac{1}{2}[|FPR_{D=unprivileged} - FPR_{D=privileged}| + |TPR_{D=unprivileged} - TPR_{D=privileged}|]$$

It needs to be equal to 0 to be fair.

4.3.4 Disparate Impact

For this metric we use the following formula :

$$\frac{Pr(Y = 1|D = unprivileged)}{Pr(Y = 1|D = privileged)}$$

Like the first metric we use both probabities of a random individual drawn from unprivileged or privileged with a label of 1 but here it's a ratio.

It changes the objective, for the disparate impact it's 1 that we need.

4.3.5 Theil Index

This measure is also known as the generalized entropy index but with α equals to
1 (more informations on [the Wikipedia page](https://en.wikipedia.org/wiki/Generalized_entropy_index)).$Sowec$

$$\frac{1}{n}\sum_{i=0}^{n}\frac{b_i}{\mu}ln\frac{b_i}{\mu}$$

Where $b_i = \hat{y}_i - y_i + 1$

So it needs to be close to 0 to be fair.

4.4 Fixing the bias problems

In this section we used the different techniques of fixing the fairness of a model:

- Pre-processing algorithms : they are used before training the model
- In-processing algorithms : they are fair classifiers so it's during the training
- they are used after training the model

4.5 Results and conclusion

In our notebook we observed that the fairness is hugely corrected with the visualisations that we made but in the other hand the performance did slightly decrease.

5 conclusion

In this project we observed a very clear bias in the artificial intelligence algorithms with different types of discrimination (race,sex,skin color).This bias is mainly because of the historical observations and the unbalance of the different types of categories in the dateset.However, there exist many solutions to tackle this problem and to mitigate this bias but in the cost of the decrease of the performance of the model.It is then the data scientist's job to choose wisely between the performance and the fairness of the model.