BLOCKNUBIE

Human And AI Bias

Human biases are well-documented, from implicit association tests that demonstrate biases that we may not even be aware to all the forms of explicit discrimination based on sex, race, religion, country of origin, social class, etc. Since AI is a human product, is not immune to being biased just as the human intelligence is..

Bias is our responsibility. It clearly hurts those discriminated against, but it also negativity affects everyone else by reducing people's ability to take part and add value in the economy and in the society. In the case of AI bias, it lessens the potential of AI by encouraging mistrust and producing distorted results. To maximize the utility that AI could provide to society, the bias in AI has to be addressed and minimized. But it has to be done in a smart way, otherwise the intended effect will be hampered.

Most of the people in AI heard of Tay, Microsoft racist AI. It was an artificial intelligence chat bot released in 2016; it caused controversy when it began to post inflammatory and offensive tweets through its Twitter account. It was so offensive that Microsoft had to shut it down just 16 hours after its launch since it was unable to correct its bias. Was it intended to be racist and obnoxious? No, certainly not, but the ML algorithm had little to no protection against being fed biased data. When actual people interacted with Tay in a twitter feed with racist "opinions", Tay became racist. In a few hours started tweeting neo-Nazi, racist and derogatory comments at a surprising rate.
Microsoft sort of learned from the mistake and decided that Zo, the successor of Tay, will not be as vulnerable as its predecessor. Every attempt to mention anything that was flagged controversial, and almost anything that could be considered offensive by anyone was flagged, prompted Zo to drop the conversation. Microsoft prevented Zo from becoming racist by curtailing the effectiveness of the AI. In Microsoft defense, it is true that many people, just as Zo was programed, will also avoid the same conversation topics. The difference is that people understand context, and will not drop off a conversation just because you mention that you come from a "controversial" country or that you saw a girl with a hijab..

Microsoft was not the only one who took the route of limiting functionality instead of fixing the underlying problem.. In 2015, Google was rightfully criticized when their image-recognition algorithm began labeling black people as gorillas. Google explained that their engine was trained to recognize and tag content using a vast number of pre-existing photos. The data bank contained all sorts of pictures: humans, animals, foods, inanimate objects, etc. The problem was that the dataset of human faces were mostly white and hence not a diverse enough representation to accurately train the algorithm. The algorithm then internalized this proportional bias and did not recognize some black people as being human. Google explained the mistake and apologized for the error. The solution that google implemented was not to retrain the algorithm by feeding it with enough faces of people of color to improve it and let it recognize people of color as humans.

Google's solution was to eliminate the possibility of labeling the data (pictures) as monkeys, gorillas or chimp… So the algorithm defaulted to human black, Latino, etc.…

It is clear that when artificially intelligent / ML machines are fed our systemic biases on the scales needed to train the algorithms that run them; the results are AI's that perpetuate our bias.

Artificial intelligence (AI) has the potential to revolutionize healthcare delivery. AI powered applications in decision support, patient care, fraud detection and disease management are fast becoming an industry standard. AI helps clinicians work smarter while improving patient outcomes based on machine learning algorithms that feeds on the enormous amount of data that is being created.
The questions arise, what if the data that the AI is fed incorporate part of the human bias? What if the algorithm, knowingly or unknowingly, is biases as its human programmers? What happens if an AI detected in the data that it been fed that, for example, those who have private insurance get better treatment, or those with darker skin tone get less pain treatment, or that men should not be nurses -since it is a woman's profession- ? What will happen is that, if unchecked, the Ai will learn that and perpetuate the tendency.

Human bias have been investigated for a long time, these are some who have potential a stronger effect on the quality of the data. In turn, this will have a strong effect on the fairenss of the AI's.

1. Confirmation bias
Occurs when the person performing the data analysis, previous to be fed to the AI, wants (consciously or unconsciously) to prove a predetermined assumption. They keep looking in the data, cleaning it and applying labels until it matches what he/she expects. It could happen by excluding particular variables or data sources from the analysis or contextual information.

2. Selection bias
This occurs when data is selected subjectively. As a result, the sample used is not a good reflection of the population. As example, in many social research studies, researchers have a tendency to use students as participants to test their hypotheses. If the universe of the research is students in that university it will be fine, it they want to apply conclusion to the rest of the population it is not.
As an example, if the data happens to be originated in the US and UK then, even if it is a fine representation of the population, it will majority white. The conclusion will be unfair or less applicable to other demographics.
You should always ask what sort of sample has been used for research, an in the AI case, what sample was fed to the AI.

3. Average vs Median or the Values of the gauss bell
An outlier is an extreme data value. It could be a customer with an age of 110 years, a consumer with €10 million in their savings account or a patient the cures from cancer without explanation. You can spot outliers by inspecting the data closely, and particularly at the distribution of values. Values that are much higher, or much lower, than the region of almost all the other values. Outliers can change the representation of an AI of the average and act accordingly

4. **Framing Effect** — Survey questions that are constructed with a particular slant.

People are more likely to save a guaranteed 200 lives compared to a 33% chance of saving everyone, if the question was framed **positively**. Or they will be more willing to do something to prevent losing X amount of money than to win the same amount of money.

There are many more bias both in the data creation, Data Collection, Data Preprocessing, Data Analysis and Modeling. Each step has to be controlled to minimize the impact of bias. How to correct the algorithm or the data to avoid bias? Human judgment is still needed to ensure AI supported decision making is fair. That human intervention is and will be always biased based on the different interpretation of what is fairness.