Maneesh Wijewardhana 2023-01-21

Week 3

This week's readings allowed me to learn more about the bias and transparent nature of AI systems and surprised me with how the two are so closely related to one another. What caught my attention the most was when the author stated “It often means that the algorithms we produce are fueled by aggregating the same intuitive and sometimes prejudiced human decision-making we’re trying to improve on” (Zerelli 48). This quote opened my eyes to the fact that most AI systems we are developing today are in fact, to put aside human bias and get non-altered decisions but still fed off of none other than, the human bias. In my opinion, this is a problem because it can disrupt accurate information in important contexts such as policing, job applications, credit loans, and much more.

After conducting some more research, a clear example of this bias was brought up to me where “Businesses use such tools to inform stock trades and other pointed decisions. But after training his tool, Dr. Bohannon noticed a consistent bias. If a tweet or headline contained the word ‘Trump,’ the tool almost always judged it to be negative, no matter how positive the sentiment” (https://www.nytimes.com/2019/11/11/technology/artificial-intelligence-bias.html). This was referring to a startup named Primer utilizing Google’s BERT AI to develop an automated sentiment judge for businesses. So how can we fix this issue?

You might think a clear-cut solution to this would be to remove certain variables from the training model, specifically those that might be more controversial than others. The problem with this, however, is now the model cannot output accurate results without certain key variables. Another solution might be to process the data beforehand. This would verify, in another domain, where sensitive attributes were changed if any bias took place. It is hard to say whether one solution will work over another since “As of 2019, major tech companies including Google, Facebook, and Microsoft, have all announced their intention to develop tools for bias detection, although it’s notable that these are all ’in-house’” (Zerelli 51).

The bottom line here is transparency. The mere fact that these bias detection tools are in-house, is detrimental to the public perception of how AI is used for decision-making, especially coming from big corporations. In my opinion, we should demand more transparency when it comes to the inner workings of these technologies as the result of it will be end users’ trust, loyalty, and support for the use of AI in the contexts mentioned above.