Tackling bias in AI (and in humans)

The goal of the study was to show how algorithms can be used to abate the biases caused by human, underlining the issues resulting from algorithmic bias due to the data they are being trained on, showing that human biases can creep into the AI system, and suggesting ways that can be used to reduce biases from the system. The authors pose a question of whether the AI can be a better unbiased decision-maker as compared to humans or will the human biases exacerbate the AI's biased decisions.

The factual study indicates that even though a program was able to match the human decisions with extremely high accuracy, it still produced a biased result against the applicants on the basis of their gender and names. There has been a major divide among people who think that algorithms can prove to be a countermeasure against human biases and people who think that algorithms can pose a serious threat to the decision making process by involving human biases. However, the authors noted that these two issues can present opportunities to enhance our AI systems as well as human decision-making abilities.

Certain examples about how judges and employers can unconsciously favor a certain group based on their own experiences rather than looking at the factual data present has given support to their claim about how humans can be prone to including factors that might not be relevant in the decision-making process. In this case, artificial intelligence can prove to be highly useful by considering only the variables that have a strong correlation with the decision output.

Although these algorithms can be useful, the people who make these algorithms have certain biases attached to them which can easily sneak into their algorithms, and when these algorithms are used to tackle real-world problems these biases are amplified. They illustrated their claim by giving an example of COMPAS, which classified American African defendants as “high-risk” twice as much as it did to white defendants.

The authors state that it's not always the algorithm's fault for introducing the biases but the data it is trained on can also be a major contributor to the issue. The data collection and sampling can result in distorting the actual results. Another factor to consider is the independent variables of our model. Furthermore, a feedback loop created by the users themselves can also result in making a model more biased as it is continuously learning from the data provided. Societally unacceptable or illegal statistical correlations can be picked up by the model and cause further complications.

A concrete definition of fairness is still not defined although Arvind Narayan has tried to list down 21 definitions of fairness. Every model can not adhere to the fairness principles and a tradeoff between other objectives and fairness definitions exists. To make a model very fair it must lose some utility. Fairness constraints can make the model less efficient in choosing a particular output and thus reducing its usefulness.

Multiple approaches are made to ensure that the model sticks to the fairness principles. These include preprocessing, post-processing the data, and imposing a fairness constraint on the optimization process. Researchers have also tried to ensure that they try to explain how the AI system reaches its conclusion to identify and minimize bias.

As AI does have the capacity of having hidden biases, human intervention is essential. An AI can not be left to make every decision by itself. However, another question posed is who are the people to keep a check on these systems. To address this issue, the authors have mentioned that efforts have been made to integrate ethics modules into the computer science curriculum and regular audits are

conducted. Another solution could be to have a diversity of people involved in the decision-making process.

To tackle the issue of biases, the person should be aware of the domains and contexts of the surroundings before deploying the system. Tests would have to be conducted on the AI systems by external or internal parties to audit the data the model predicts. People should engage more in a conversation about the long-held human biases and how models can rectify our behavior. Humans and machines should work in harmony and no fully automated system should exist. More investment is also needed to research the biases and the study of AI itself and it should have people from all disciplines involved to better evaluate the AI's fairness.