Diego Perez

CS 59866: Ethical, Legal, and Policy Issues in Technology and Cybersecurity

9/26/2022

Position Paper Week 5: Bias in AI

As technology keeps advancing and expanding in the field of artificial intelligence, allowing us to move towards an even more complex and interconnected world every day. These new inventions and applications have proved to be of almost indispensable use not only on a personal level. Especially in the professional levels including sensitive areas such as hiring, criminal justice, and the healthcare system. However, despite all the improvement and utilities AI can provide in the professional setting, AI also comes with its limitations. The main topic of discussion and controversies originated from the use of these technologies is the capability of AI to be “bias” and “unfair”.

From the article titled “Notes from the AI frontier: Tackling bias in AI (and in humans)”, by Jake Silberg and James Manyika, my purpose in this position paper will be to present and further discuss and explore the topics touched in this article. As of today, the use of AI is expected to keep growing across a wide range of sectors, affecting individuals, businesses, organizations, and economies everywhere. While many believe that the use of these technologies will provide a solution to the human biases, others think that in the contrary, it will increase and propel human and societal biases.

Corroborated by the article, both sides of the argument are right. In its majority AI can help reduce humans’ subjective interpretation of data., and they improve decision making. In addition, they help reduce racial disparities in the criminal justice system. Unlike human decisions, the decisions made by AI can be examined. On the other hand, there is plenty of evidence where AI has been subject for being biased. For example, hiring algorithms making exclusions between genders, and race, and recidivism prediction algorithms incorrectly labeling groups of people because of race.

If this is the case where algorithms are making bias decisions, then it is clear for me to say that this is a simple problem with a simple solution. These algorithms are either not working properly and were not implemented correctly, hence they should be discarded, or improvements and further implementation need to be made. The article suggests that one possible reason for the presence of bias in these technologies’ decisions, is the presence of bias in the underlying data used to train these algorithms. There are different ways the data can come to contain biases.

Bias is most commonly introduced into data through how it was collected and selected for use. Also, data that is generated by users can create feedback loops that lead to bias. Other important cause is the presence and choice of variables. These algorithms are sometimes capable of picking up on statistical correlations between variables, therefore resulting in incorrect assumptions/decisions. Sometimes these correlations may be societally unacceptable or illegal.

Regarding this issue the article discusses a number of approaches that can be applied to minimize the existence of bias in the algorithm’s decisions. One approach would be to pre-processing the data to maintain as much accuracy as possible while reducing any relationship between outcomes and protected characteristics. Also producing representations of data that do not contain information about sensitive attributes. Other approach is to transform some of the algorithm’s decisions after they are made in order to satisfy the given definition of fairness constrains.

Other approach is to add fairness constraints on the optimization process itself or make use of an adversary to minimize the system’s ability to predict the sensitive attributes. Others have focused on encouraging impact assessment and audits to check for fairness before systems are deployed and to review them on an ongoing basis, as well as on fostering a better understanding of legal framework and tools that may improve fairness. In addition, other possibility that is explored and sometimes used is human-in-the loop AI. This latest is used to assist decisions rather than leaving the algorithms to do the decisions, that is humans and machines working together.

In response to the multiple solutions discussed, multiple questions need to be answered to make better sense of how they can be put to use. The main concern is regarding the fairness constraints since many and different definitions of what is constituted as fair can be found. This is a problem because as the article corroborates models cannot conform to more than a few group fairness metrics at the same time. Deciding on the best metric or combination of metrics to determine if a system can demonstrate group, individual or other notion of fairness is complex.

In addition to this, ensuring that an AI system satisfies measures of groups fairness could create trade-offs with measures of individual fairness and could reduce the utility of the algorithm. Other questions originated from the possible solutions is, who decides when an AI system has sufficiently minimized unfair bias so that it can be safely released for use. Other question is, in which situation should fully automated decision making be allowed at all. In this case it is important to recognize that there are circumstances in which Artificial Intelligences cannot make decisions on their own, requiring human intervention.

In conclusion, biased decision making either by human or AI can have devastating consequences. Minimizing bias is a crucial prerequisite to allow AI to reach its maximum potential. Being aware of where to deploy these AIs, establishing processes and practices to test for bias, further exploring how to implement a human-machine environment. Also, investing in bias research, making more data available, and investing in diversifying the AI field are just some of the suggestions can be looked at to maximize fairness in AI.