ARTIFICIAL INTELLIGENCE AND ETHICS

Ashar Farooq
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
afarooq@mit.edu

August 12, 2020

EXECUTIVE SUMMARY

The modern world has seen an explosion in the usage of Artificial Intelligence(AI) in many different domains and fields. These innovative technologies offer the best of what the human civilization has to offer. However, the rapid developments of these tools is causing various issues that have only begun to be looked at seriously. There are serious flaws in many AI tools that perpetuate unfairness and discrimination.

Artificial Intelligence is the broadest scope of a field that relies on predictions and training. The typical AI pipeline begins well before the actual algorithm: the data. The AI models are trained on training data and tested using the testing data. The original dataset from the real world is often incompatible and unusable for the AI algorithm. More work in pre-processing the data and making it ready for the training of a model is usually required. After getting the data and formatting it appropriately, the AI algorithm trains on the training data, learning some correlations between the inputs and outputs, also known as features and labels. After many iterations, the algorithm produces a model that can be used to make predictions. This AI pipeline is often relatively accurate. However, the nature of Artificial Intelligence is statistics and predictions, thus getting complete accuracy is not feasible. Nevertheless, there are significant issues in addition to this that certainly needs to be looked at and fixed.

There are many entry points of bias and discrimination in the AI pipeline. The data collection process is often the most important manner of attaining algorithmic discrimination as without an inclusive dataset, the AI model will train on biased data and produce biased predictions. This is often challenging as much data is accurate, but reflects the systemic prejudice and discrimination of various stakeholder groups. Another method of introducing bias and discrimination into AI models is to utilize features that should not be utilized in various cases. Other methods include sample bias, inaccurate data, and flawed AI algorithm with an exclusive design thinking process.

These entry points and other factors of algorithmic discrimination have already created problematic cases of AI usage. The COMPAS algorithm, which stands for Correctional Offender Management Profiling for Alternative Sanctions, is utilized in the criminal justice system in order to determine a defendant's likelihood of reoffending. While this seems like an efficient solution to a problem, the algorithm was inherently biased as this tool predicted higher risk scores for African American defendants twice as likely as compared to others. The Impact Pro Healthcare Algorithm by Optum is a tool designed to predict high-risk patients in order to enroll them in a program with more healthcare resources. This tool was problematic because African American patients identified at the same level of risk as other patients were actually sicker to a significant degree. Amazon's internal AI tool designed to screen candidates was problematic because the model effectively learned to prefer male candidates and not prefer resumes with language relating to women. This reflects the dominance of men in the technology industry. Various cases of algorithmic bias in various search engines are existent as many

search results of certain demographics return sexist and problematic results. Sexism is also rooted in some translation programs with some correlating certain roles, such as doctor, with males and other roles, such as nurse, with females. Gender classification via facial recognition technology is also inaccurate to a significant degree for many members of the African American and Asian communities.

These cases of algorithmic discrimination makes it clear to invest in research and studies of other AI tools. With more knowledge and understanding of the AI Ethics issue, more progress can be made to ensure fairness and equity for all in terms of these AI tools. The best future pathways for combating algorithmic bias include an inclusive design thinking process, fairer datasets, inclusive algorithms that take into account various edge cases, and more human intervention in a holistic manner where multiple factors produce a decision. AI is valuable for narrowing down the larger problem into a smaller problem, thus making it efficient. However, this human intervention at the last stages can make the overall decision making much more effective and also equitable as a machine currently does not have the human intuition and understanding to be able to make highly consequential decisions that cannot be only based on mathematics. In addition, more collaboration between computer scientists, ethicists, educators, lawyers, and other community members is needed for fairer and effective future AI tools. AI is both very good and very bad at times, thus the human mission should be to make it very very good with equity and inclusion in mind.