AINA

AI and Analytics

Volume 1 • Edition 2019-20

Annual analytics magazine from the students of PGDBA

Post Graduate Diploma in Business Analytics: *Jointly offered by IIM Calcutta, ISI Kolkata, IIT Kharagpur*

The world is not fAIr

BY SOWRYA REGANA

The fast paced and multi directional development of AI applications is supporting humans in areas ranging from buying a simple toothbrush to sending reusable spacecrafts into the space. As AI is being increasingly adopted in crucial and sensitive applications, it raises the need to study about the biases that these systems bring to the table. Recent times have seen significant increase in the examples of AI systems reflecting or exacerbating machine biases, from racist facial recognition to sexist natural language processing. Recently, American Civil Liberty Union filed a case against Detroit Police for falsely arresting an African American due to mis-identification attributed to its facial recognition software.

According to IBM Research, there are more than 180 human biases in AI systems. Biases, of any kind threaten to overshadow AI's technological gains and potential benefits and hence have become a primary matter of concern for Policy Makers (governments), Auditors, Businesses and their end-users. Businesses often tend to shorthand their explanation of AI bias by blaming it on biased training data. But the reality is more nuanced. Biases can creep in long before the data is collected, sometimes at various stages of the data collection process. In the current article, we collate the various bias and fairness notions defined and proposed by researchers actively working in the field of ethical AI. In the interest of readers, these definitions have been adapted verbatim from the paper. (see note at the end of the article)

Historical Bias is the already existing bias and socio-technical issues in the world that can seep into from the data generation process even given a perfect sampling and feature selection. An example of this type of bias can be found in a 2018 image search result where searching for women CEOs ultimately resulted in fewer female CEO images due to the fact that only 5% of Fortune 500 CEOs were woman—which would cause the search results to be biased towards male CEOs.

Representation Bias happens from the way we define and sample from a population. Lacking geographical diversity in datasets like ImageNet is an example for this type of bias. This demonstrates a bias towards Western countries

These are few of the many existing biases that affect AI systems. However, there are also some fairness notions which when followed ensure that AI systems are free from few of these biases.

Algorithms do what they're taught, unfortunately some are taught prejudices and unethical biases by societal patterns hidden in the training data. To build algorithms responsibly, we need to pay close attention to various sources of potential discrimination or unintended harmful consequences. Due to increasing usage of AI in judicial systems, health care and other crucial domains, it is very important to ensure that it gives an unbiased output. Businesses and organizations should also ensure that their AI systems are be free from bias both from data and also the algorithmic perspective.

*The groups that are often victims of discrimination are termed as protected groups. It varies based on context. (Ex. African-Americans, females etc.)

Interested readers can kindly go through the paper titled "A Survey on Bias and Fairness in Machine Learning" available on ArXiv : https://arxiv.org/pdf/1908.09635.pdf

Fairness Notions	
Equalized Odds	Protected and unprotected groups should have equal rates for true positives and false positives.
Equal Opportunity	Protected & unprotected groups posses equal true positive rates
Demographic Parity	The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group
Fairness through Awareness	Any two individuals similar w.r.t. a similarity (inverse distance) metric defined for a task must have similar outcome
Fairness through Unawareness	No explicit usage of protected attributes