Bias in AI Systems

Submitted by -Saachi

With the surge of the fourth wave of the Industrial Revolution-Machine intelligence, blockchain-based decentralized governance, genome editing and AI in healthcare are among the top trending arenas and has tremendously grabbed the interest of the researchers and enthusiastic techies. But as much as these solutions have greatly impacted, automated the tasks and improved the learning in the respective areas along with a lot more interesting insights and findings, it also poses new ethical challenges and blockers. Want to know what those might be?

Biases in the psychological world are quite common. Humans are more or less inclined towards a certain object or an opinion. But do you agree with the thought of an algorithmic model favouring something over the other?

Being able to generalize the problem over a set of inputs represents the key characteristic of a machine learning or an artificial intelligence algorithm, meaning it must be able to accurately predict the output of new data based on learning it has gained from the training data. But, if the incoming new data contains features not previously seen in the training dataset, the AI will have trouble classifying what this new data is. If this all sounds a bit abstract, here is a quick example. Imagine you want to develop an Image classification model for identifying a cat or a dog. In the training dataset, you feed it hundreds of thousands of labelled images of dogs of different breeds, but just very few of cat breeds; it is very likely for your model to misclassify an image of a persian cat as that of a dog. The model is not generalizable enough to all cat breeds because it has not been trained with the sufficient images of other cat breeds. Now, this represents a bias in an AI Image Classification Model.

Root cause for introducing bias in AI systems

Data imbalance is one of the major problems in introducing bias. For instance, in 2016, Microsoft released an AI-based conversational chatbot on Twitter designed to interact with people through tweets and direct messages. However, within a few hours of its release, the replies became quite offensive and loaded with racist messages. The chatbot was trained on anonymous public data and had a built-in internal learning feature, which led to a coordinated attack by a group of people to introduce racist bias in the system. This incident was an eye-opener to a broader audience of the potential negative implications of unfair algorithmic bias in the AI systems.

Class imbalance is the leading issue in many of the classification problems. Apart from algorithms and data, researchers and engineers developing these systems are also responsible for AI bias. This is generally termed as human bias.

Bias Prevention measures-

Bias is all of our responsibility. Awareness and good governance can help prevent machine learning bias by cultivating best practices to mitigate it. Selecting a representative sample will counteract common types of machine learning and artificial intelligence biases. Also, monitor the learning systems as they perform their tasks to ensure biases don't creep in over time as the systems continue to learn as they work. Use of additional resources, such as Google's What-if Tool or IBM's AI Fairness 360 Open Source Toolkit, to examine and inspect models will be a step towards trusted AI.