Sumona Halder | MBB 2020 | Computer Science

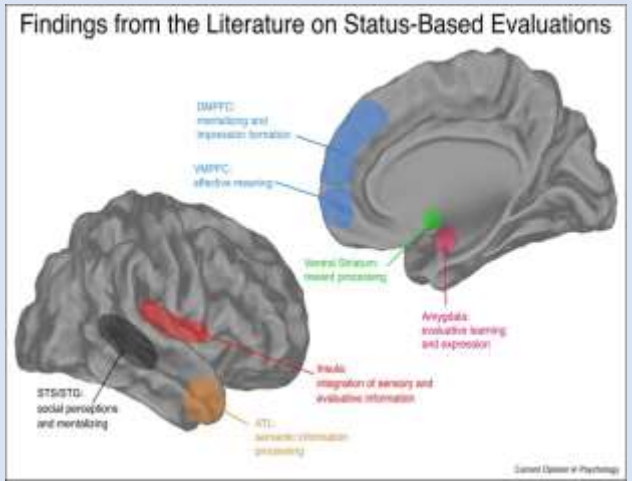
The Dangers of Algorithmic Autonomy

Abstract :

In current research surrounding machine learning algorithms, we have faced a large ethical and moral issue. Our algorithms, which are intended to provide us with the most optimal and unbiased decision has started to emulate the implicit biases of humans. Inductive bias on the other hand, offers some insight as how we can generalize from specific examples and maximize future predictions. While inductive reasoning is not immune to being affected by implicit bases, it can be used to properly train algorithms to produce better outcomes. In this paper we will explore a theoretical solution through a mechanism-first (building upon a foundation of cognitive processes) strategy.

What is Implicit Bias ?

Implicit biases can form through early memories or subliminal messaging, they are truly a part of our unconscious state as they are triggered as automated responses to specific stimuli. Since the amygdala is known not only to detect fear and danger but also is a primary role in visual processing (as most studies as noted in this paper have used visual stimuli) it can be concluded that it plays a central role in the automatic and non-conscious processing of emotional and social stimuli. Therefore, the amygdala can be attributed to demonstrating the almost automatic and inherent reactions that implicit biases illicit.



Implicit Bias in AI?

Though we have tried to create machine learning algorithms (modeled after inductive bias) to create better predictions and hopefully come to unbiased conclusions. Unfortunately, however, because it is humans who create such algorithms and feed it the training sets or data, the AI learns from these inputs and makes predictions based on this data. Due to the inherent, automatic nature of implicit bias and the difficulty of correcting implicit bias with external stimuli and extensive training such bias seeps into our algorithms rendering them just as biased as we are.

Importance of Inductive Bias and Reinforced learning in Machine Learning Algorithms

Inductive bias is essential to both human and machine learning. In this case, bias woks in a positive way and is necessary for recognizing patterns and learning to label future possibilities or options not presented or known . It tends to cut down time and energy for us as we don't spend time speculating and further investigating or stressing on these new choices. Based on our previous observations we can often count on this predictions to be correct and beneficial for us thus leading to efficient decisions. This type of search mechanism has been developed to be the basis for neural networks and machine learning algorithms as it can now be applied to much more complex and extensive datasets that accelerate the search of valuable options.

Conclusions

The apparent solution would be to remedy our own implicit biases and produce unbiased (in the implicit sense) data and training sets. This would allow for machine learning algorithms to have "clean" data sets to learn and make more accurate generalizations from and ultimately accurately predict outcomes from few examples. However, based on our research we have seen how implicit bias an unconscious, automatic cognitive process and so trying to undo years of childhood memories and subliminal messaging does not seem to be a realistic solution. In addition to this, as I have discussed in this paper, while implicit bias has ways to be retrained and reversed in humans, there have been mixed effects on trying to change implicit bias. Thus, the first solution I propose is that with proper data and training sets even bad algorithms can outperform good ones.

My Proposed Theoretical Model

A model that I have been exploring is the union of ML with PBE. PBE is programming by examples (PBE), a subset of program synthesis. Program Synthesis is the task of synthesizing a program that satisfies a given specification. The main strength of program synthesis is that not only can it function well with only a small number of specific functions and aims to output a program that satisfies those specifications, but the synthesized program is interpretable and hence editable. ML on the other hand, requires a large amount of training data and tries to minimize loss. But ML is robust to noise and handles generalization which is something that program synthesis does not have the capabilities to do. Implementing ML in a PBE framework thus could work to combat implicit biases.