Fixing Algorithmic Bias Requires Fixing Societal Injustice

Gabbrielle Johnson (Claremont McKenna College)

Artificial intelligence is often believed to rely purely on objective mathematical processes operating upon raw data, leaving no role for social biases. But algorithms exhibit the same problematic social biases that humans harbor: Algorithms used by doctors throughout the United States are significantly more likely to determine that a white patient needs extra care than a black patient, even when the two patients have almost identical health issues. Facial-recognition software performs poorly on images of women and non-white people because male, white faces are overrepresented in the training data. COMPAS, a recidivism risk-assessment algorithm, used throughout the country, erroneously flags black defendants as high-risk at nearly twice the rate of white defendants.

To date, philosophical work on bias has focused exclusively on top-down biases. While well-documented, data-driven bottom-up biases are poorly understood. As reliance on AI deepens, understanding the sources of and solutions to such biases is ever more pressing. It requires both technical knowledge and a deep understanding of the underlying ethical and sociopolitical issues. Schellenberg and I have been working jointly on a project straddling these domains for several years. We are applying for a Lebowitz Prize to facilitate collaborating on this project in a focused manner.

By contrast to Schellenberg, I argue that no biases are purely bottom-up; all are intertwined with human bias. Adjudicating our dispute will allow us to gain a clearer understanding both of the nature of algorithmic bias and of what interventions are effective. While there are deep differences between our two ways of understanding bottom-up biases, there is significant common ground—especially when it comes to strategies for making algorithms less biased.

I acknowledge that interventions that affect only the algorithm, absent interventions on the wider patterns of societal injustice, are our best practical approach to curbing algorithmic bias: such interventions pinpoint a precise location in a complex causal network at which we can intervene. In her forthcoming book, Schellenberg argues that pure bottom-up biases interact with societal patterns and that changing such societal patterns will affect and perhaps even eliminate the algorithmic bias. We both believe that given the ever-growing feedback loops that exist between humans and machines, interventions on algorithms are necessary to dismantle biases harbored by humans and machines alike.

Creating Less Biased and More Intelligent Algorithms

Susanna Schellenberg (Rutgers)

We are living in the age of Al. Algorithms are used to make decisions about criminal sentences, loans, policing, credit card applications, job applications, and medical care. We do not just use the fruits of AI systems. Our actions provide the data on which they operate, and they are being used to make decisions about us. Since the pandemic, our lives have moved online to an unprecedented degree. As a consequence, we are ever more subject to the positive and negative consequences of AI.

All AI systems have at least some biases. The reason is that there is a mismatch between processing power and the quantity of data on which they operate. To deal with this mismatch, AI systems use statistical generalization to operate efficiently.

Contrary to what is frequently assumed, both among the public and researchers, most algorithmic biases stem not from the effects of a programmer's beliefs, goals, and background views, but rather from how AI systems function at the lowest level: the patterns that the algorithm detects within the data it receives, the connections it forges between data points, and the generalizations that emerge from these connections. In short, they are bottom-up biases rather than top-down biases. While top-down biases are due to the goals, expectations, or beliefs of humans, bottom-up biases are due to incoming data and its processing at the lowest level.

Johnson and I agree that most biases in AI are bottom-up biases and that such biases affect not only AI systems but also human cognition and perception. We disagree, however, as to what relation, if any, there is between bottom-up and top-down biases. According to Johnson, any bottom-up bias is intrinsically intertwined with top-down biases. Drawing on recent research at the intersection of neuroscience and AI, I argue that at least some are pure bottom-up biases: they are entirely independent of any top-down biases.

This difference has ramifications on how best to mitigate AI's harmful effects. If there are pure bottom-up biases, we can make an algorithm less biased simply by improving on the algorithm itself: correcting how the algorithm links data and ensuring that the training data is unbiased. According to Johnson, this approach will never be successful: if all bottom-up biases are intertwined with top-down biases, any ameliorative interventions will require intervening on wider societal patterns that give rise to the bottom-up biases.

Johnson and my skillsets complement each other optimally to address this topic. If we are fortunate enough to receive a Lebowitz Prize, we will be able to take full advantage of the recent academic and media interest in biased AI by having the time to focus on writing our joint papers.