Dehumanizing AI: Removing Bias in Machine Learning Using Counterfactual Simulation

Julian TszKin Chan and Weifeng Zhong

May 31, 2019

Long Abstract

In recent years, machine learning algorithms are sometimes found to be biased—often dubbed as “artificial intelligence (AI) bias.” When applied to criminal justice, hiring, and lending, for example, those algorithms can be discriminating against protected groups (by gender, race, etc.).

Driving such bias is the fact that, when the algorithm is trained with historical data, which may contain bias in human behavior from the past, the bias is “successfully” recovered from the history of our human decisions.

This kind of bias presents a unique challenge to policymakers. Forbidding the use of protected variables (i.e., banning disparate intent) is apparently under-regulating, because the model could still infer the bias from unprotected variables through their correlations with protected ones. However, if policymakers moved to ban disparate impact, they would be over-regulating; firms would have to remove so much information from their data that the machine learning applications may be appealing. Over-regulating the use of these algorithms would further hinder innovation and, by extension, productivity growth.

In this paper, we first discuss various methods in the literature that attempt to remove unwanted bias in machine learning and demonstrate why those methods may either under-serving or over-killing. We then propose a three-step procedure to remove unwanted bias using counterfactual simulation.

The procedure the following steps: (1) Use all variables, protected and unprotected, to estimate a machine learning model—the “true” model—that truthfully reflect the patterns in historical data including the unwanted bias. (2) Use the “truth” model to simulate a counterfactual world where the unwanted bias is removed, which we call the “ideal” data. (3) Estimate another machine learning model using the “ideal” data, which does not contain the unwanted bias anymore, for the future predictive tasks.

Finally, we demonstrate the three-step procedure using an example data set, and we show that not only the resulting algorithm does not discriminate against protected groups but its performance is not significantly poorer than the under-regulated models in the literature that do discriminate protected groups.