Alexis D. Torres
Prof. Christopher Mykolyk
Ethics and Computer Science
Hunter College, CUNY

3 September 2020

De Cremer, D. (2020, September 3). *Harvard Business Review*. What Does Building a Fair AI Really Entail? Retrieved 03 September 2020, from https://hbr.org/2020/09/what-does-building-a-fair-ai-really-entail?ab=hero-subleft-2

This article deals with the increasingly cumbersome and expanding scope of algorithms used to make decisions and predictions in the real world—outcomes that have significant implications and ramifications for the people at the center of these programs who, ironically, are the ones targeted to benefit most from them. According to the author, these algorithms embedded in machine learning artificial intelligence have been created to remove the harmful aspects of human intervention in activities such as hiring or decisions around criminal justice issues. In other words, these applications are supposed to remove bias, discriminatory leanings, and other injurious human propensities when taking action in a human-involved arena. However, he concedes, these systems often perpetuate the same negative outcomes and "mindsets" they were designed to prevent or ameliorate. Appropriately so, the author focuses on the most significant aspect of the problem in the way these algorithms function: the data sets.

In the recent past, it has become ever more clear to computer scientists, citizens, business, and governmental institutions alike that there is an inherent danger in relegating life-altering decisions to the musings of a supposedly dispassionate algorithm. Research has demonstrated that due to the biases embedded in the original data sets, artificial intelligence continues to learn to be essentially more biased—or at least value biased outputs that best reflect the biased inputs it received. This renders these machine learning algorithms ineffective at addressing societal issues such as racism and sexism. More importantly, the impact on the individuals whose lives are shaped in large part by the outputs generated by these programs are made to suffer life-long consequences that in some instances may be irreversible.

This issue has serious implications for my students and the communities from which they originate. Every year, the number of companies and government systems utilizing artificial intelligence algorithms with biased data is increasing. These applications are used to determine things like where my students can attend college, the kind of jobs they will qualify for, the places and dynamics of communities in which they will be permitted to live, and so forth. It is clear that entrusting these decisions to an algorithm will automatically place my students in difficult and precarious situations in different areas of their life—throughout their entire lives. This may mean that we will be using technology to codify inequities and systems of oppression with more exacting and devastating precision. The implications, however, go beyond the individual student for the broad application of this technology within already marginalized communities will serve to isolate and oppress them further. It would not be hyperbolic speech, therefore, to speak of a future society that is more segregated; experiences more economic strife, and relegates its neediest to a more permanent diminished status in our nation. Furthermore, similar repercussions can be witnessed in other societies and communities around the world since these technologies are regularly exported to other nations.

Unlike many articles on this topic, the author does propose two important starting points for addressing the issue. The first is developing "interpretable AI", which will allow people to see the mechanism by which the algorithms make their determinations. Essentially, programmers could institute numerous corrective actions to continuously refine an algorithm towards ever "fairer" states. While this seems arduous and unworkable over the long term, it still represents a good start. The second offering is developing protocols and principles that focus on developing algorithms that prioritize fairness in their development. Similar to the first proposal, this too is an arduous and error-prone solution. Nonetheless, what the existence of this article (and especially its author’s suggestions) signal is that the awareness of this grave problem is increasing. This is an important point because it helps my students and their community members become better prepared to understand how these technologies can impact their lives—and what they can do to protect themselves. However, these are not solutions. My students, their communities, our nation, and the world must demand better from programmers and the end users of their products because these technologies currently cause more damage than fix problems.