Algorithms and Bias

Julie M. Smith
University of North Texas, USA

INTRODUCTION

On March 23, 2016, Microsoft released a bot called Tay on Twitter that was capable of interacting with other Twitter users. Just two days later, a Microsoft Vice President, Peter Lee, had to publicly apologize because, as Lee (2016) explained in a blog post, "Tay tweeted wildly inappropriate and reprehensible words and images," including support for genocide. Lee explained that, while this was not the first chatting bot Microsoft had released, they were unprepared for users who exploited Tay's tendency to parrot extremely offensive messages. Lee noted that Tay, as an artificial intelligence, learned from "both positive and negative interactions with people. In that sense, the challenges are just as much social as they are technical."

Tay is but one example of the fact that by the second decade of the twenty-first century, algorithms were responsible for making a nearly infinite number of decisions, from deciding who would be given a bank loan to determining what advertisements joggers hear on their music streaming service. It is fair to say that most people do not appreciate the extent to which the choices they make are influenced by algorithms. Microsoft's experience with Tay serves as a cautionary tale for the ability that algorithms, especially those that deploy machine learning and/or artificial intelligence, have to engage in patently discriminatory behavior--even despite the intentions of their creators. Megan Garcia, a senior fellow emphasizing cybersecurity at New America CA, wrote, "Computer-generated bias is almost everywhere we look" (Garcia, 2016, p. 112). This article will examine algorithmic bias, which is the potential for algorithms to engage in discrimination, and it will also suggest ways to avoid this problem.

BACKGROUND

While there is widespread agreement that algorithmic bias exists, defining it precisely is tricky, as there are different ways to measure bias. Using the example of an algorithm that assesses risk for criminal behavior, one might measure bias in several different ways (Huq, 2019). One could look at aggregate risk scores for various groups and see if they differ. Or, one could determine if the same initial risk score resulted in the same final risk score for people with different demographic characteristics. One could determine whether the rate of false positives and/or false negatives varied for each demographic group. Efforts to improve the fairness of an algorithm on one measure may actually lead to worse performance on another measure. For example, Speicher, et al (2018), borrowed the concept of inequality indices, which are used by economists, and applied them to biased algorithms. This provided a way to quantify bias. But they also found that efforts to minimize between-group bias may actually increase within-group bias.

For purposes of this article, a biased algorithm will be defined as one that that unfairly and/or inaccurately discriminates against a certain person or group of people, especially on the basis of protected

DOI: 10.4018/978-1-7998-3473-1.ch065

categories such as race and/or gender. In some cases (as will be discussed below), the bias is not in the algorithm per se but rather in the data which it uses.

Unfortunately, examples of algorithmic bias are not difficult to find. A study found that names associated with people of color were far more likely to return search results that were negative (in this case, related to arrest records) than neutral or positive. A black-identified name was 25% more likely to return an ad for an arrest record (Sweeney, 2013). In an experience that went viral in 2015, Jacky Alcine and his friend, both African American, were tagged as "gorillas" by Google Photos (Garcia, 2016). Google responded promptly by removing the auto-tags for terms that might be offensive, but they didn't fix the underlying problem, which was the algorithm's inability to properly identify people with darker skin (Monea, 2019).

In addition to racial disparities, some algorithms display gender bias. In a study of smartphone-based personal assistants (including Siri, S Voice, and Google Now), the assistant was able to recognize statements such as "my foot hurts" but was not able to respond appropriately to statements such as "I was raped" or "I was beaten up by my husband" (Garcia, 2016).

And sometimes the bias is intersectional: in a study of gender classification systems, there was an error rate for darker-skinned women of over one-third, while the error rate for lighter-skinned men was less than one percent (Buolamwini & Gebru, 2018). Similarly, a Google search for "unprofessional hairstyles for work" featured almost all women of color, while a search for "professional hairstyles for work" pictured white women (Noble, 2018).

ALGORITHMIC BIAS

The examples above illustrate that sometimes algorithmic bias is easy to see; in other cases, however, it persists unrecognized. The latter situation is due to the complex constellation of possible sources of algorithmic bias, combined with several complicating factors. These will be explored below.

Sources of Algorithmic Bias

The source(s) of algorithmic bias in any particular situation is not always discoverable, especially in the presence of the complicating factors catalogued below. But, in general, there are five major sources of algorithmic bias.

Deliberate Choices

It is, of course, possible that programmers—or those who supervise their work—inject deliberate bias into their work. Because algorithms are usually hidden from scrutiny, companies and individuals can use them "to hide anticompetitive, discriminatory, or simply careless conduct behind a veil of technical inscrutability" (Pasquale, 2015, p. 163).

For example, Hicks (2019) shows how the British government was actually more accommodating of transgender people before record systems were digitized, but when computers were deployed to manage the social welfare system, they "helped reinstantiate binary gender and all of its attendant inequalities" (p. 22). While computers are theoretically capable of providing greater flexibility in record-keeping than paper-based systems, choices made by program designers can result in less flexibility instead.