Dealing With Bias in Artificial Intelligence (Published 2019)

Three women with extensive experience in A.I. spoke on the topic and how to confront it.

- Give this article

Credit... Harriet Lee-Merrion

Published Nov. 19, 2019
Updated Jan. 2, 2020
This article is part of our Women and Leadership special section, which focuses on approaches taken by women, minorities or other disadvantaged groups challenging traditional ways of thinking.

Bias is an unavoidable feature of life, the result of the necessarily limited view of the world that any single person or group can achieve. But social bias can be reflected and amplified by artificial intelligence in dangerous ways, whether it be in deciding who gets a bank loan or who gets surveilled.

The New York Times spoke with three prominent women in A.I. to hear how they approach bias in this powerful technology. Daphne Koller is a co-founder of the online education company Coursera, and the founder and chief executive of Insitro, a company using machine learning to develop new drugs. Dr. Koller, an adjunct professor in the

computer science department at Stanford University, spoke to bias through the lens of machine-learning models.

Olga Russakovsky is an assistant professor in the Department of Computer Science at Princeton University who specializes in computer vision and a co-founder of the AI4ALL foundation that works to increase diversity and inclusion within A.I. Dr. Russakovsky is working to reduce bias in ImageNet, the data set that started the current machine-learning boom.

Timnit Gebru is a research scientist at Google on the ethical A.I. team and a co-founder of Black in AI, which promotes people of color in the field. Dr. Gebru has been instrumental in moving a major international A.I. conference, the International Conference on Learning Representations, to Ethiopia next year after more than half of the Black in AI speakers could not get visas to Canada for a conference in 2018. She talked about the foundational origins of bias and the larger challenge of changing the scientific culture.

Their comments have been edited and condensed.

Daphne Koller

You could mean bias in the sense of racial bias, gender bias. For example, you do a search for C.E.O. on Google Images, and up come 50 images of white males and one image of C.E.O. Barbie. That's one aspect of bias.

Another notion of bias, one that is highly relevant to my work, are cases in which an algorithm is latching onto something that is meaningless and could potentially give you very poor results. For example, imagine that you're trying to predict fractures from X-ray images in data from multiple hospitals. If you're not careful, the algorithm will learn to recognize which hospital generated the image. Some X-ray machines have different characteristics in the image they produce than other machines, and some hospitals have a much larger percentage of fractures than others. And so, you could actually learn to predict fractures pretty well on the data set that you were given simply by recognizing which hospital did the scan, without actually ever looking at the bone. The algorithm is doing something that appears to be good but is actually doing it for the wrong reasons. The causes are the same in the sense that these are all about how the algorithm latches onto things that it shouldn't latch onto in making its prediction.

To recognize and address these situations, you have to make sure that you test the algorithm in a regime that is similar to how it will be used in the real world. So, if your machine-learning algorithm is one that is trained on the data from a given set of hospitals, and you will only use it in those same set of hospitals, then latching onto which hospital did the scan could well be a reasonable approach. It's effectively letting the algorithm incorporate prior knowledge about the patient population in different hospitals. The problem really arises if you're going to use that algorithm in the context of another hospital that wasn't in your data set to begin with. Then, you're asking the algorithm to use these biases that it learned on the hospitals that it trained on, on a hospital where the biases might be completely wrong.

Over all, there's not nearly as much sophistication as there needs to be out there for the level of rigor that we need in terms of the application of data science to real-world data, and especially biomedical data.

Credit... David Crow

Olga Russakovsky

I believe there are three root causes of bias in artificial intelligence systems. The first one is bias in the data. People are starting to research methods to spot and mitigate bias in data. For categories like race and gender, the solution is to sample better such that you get a better representation in the data sets. But, you can have a balanced representation and still send very different messages. For example, women programmers are frequently depicted sitting next to a man in front of the computer, or with a man watching over her shoulder.

I think of bias very broadly. Certainly gender and race and age are the easiest to study, but there are all sorts of angles. Our world is not fair. There's no balanced representation of the world and so data will always have a lot of some categories and relatively little of others.

Going further, the second root cause of bias is in the algorithms themselves. Algorithms can amplify the bias in the data, so you have to be thoughtful about how you actually build these systems.

This brings me to the third cause: human bias. A.I. researchers are primarily people who are male, who come from certain racial demographics, who grew up in high socioeconomic areas, primarily people without disabilities. We're a fairly homogeneous population, so it's a challenge to think broadly about world issues. There are a lot of opportunities to diversify this pool, and as diversity grows, the A.I. systems themselves will become less biased.

Let me give one example illustrating all three sources. The ImageNet data set was curated in 2009 for object recognition, containing more than 14 million images. There are several things we are doing with an eye toward rebalancing this data set to better reflect the world at large. So far, we went through 2,200 categories to remove those that may be considered offensive. We're working on designing an interface to let the community flag additional categories or images as offensive, allowing everyone to have a voice in this system. We are also working to understand the impact that such changes would have on the downstream computer vision models and algorithms.

I don't think it's possible to have an unbiased human, so I don't see how we can build an unbiased A.I. system. But we can certainly do a lot better than we're doing.

Credit... Cody O'Loughlin for The New York Times

Timnit Gebru

A lot of times, people are talking about bias in the sense of equalizing performance across groups. They're not thinking about the underlying foundation, whether a task should exist in the first place, who creates it, who will deploy it on which population, who owns the data, and how is it used?

The root of these problems is not only technological. It's social. Using technology with this underlying social foundation often advances the worst possible things that are happening. In order for technology not to do that, you have to work on the underlying foundation as well. You can't just close your eyes and say: "Oh, whatever, the foundation, I'm a scientist. All I'm going to do is math."

For me, the hardest thing to change is the cultural attitude of scientists. Scientists are some of the most dangerous people in the world because we have this illusion of objectivity; there is this illusion of meritocracy and there is this illusion of searching for objective truth. Science has to be situated in trying to understand the social

dynamics of the world because most of the radical change happens at the social level.

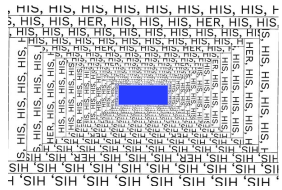
We need to change the way we educate people about science and technology. Science currently is taught as some objective view from nowhere (a term I learned about from reading feminist studies works), from no one's point of view. But there needs to be a lot more interdisciplinary work and there needs to be a rethinking of how people are taught things.

People from marginalized groups have been working really hard to bring this to the forefront and then once it's brought to the forefront other people from nonmarginalized groups start taking all the credit and pouring money into "initiatives." They're not going to take the kinds of risks that people in marginalized communities take, because it's not their community that's being harmed.

All these institutions are bringing the wrong people to talk about the social impacts of A.I., or be the faces of these things just because they're famous and privileged and can bring in more money to benefit the already privileged.

There are some things that should be discussed on a global stage and there should be agreements across countries. And there are other things that should just be discussed locally. We need to have principles and standards, and governing bodies, and people voting on things and algorithms being checked, something similar to the F.D.A. So, for me it's not as simple as creating a more diverse data set and things are fixed. That's just one component of the equation.

Craig S. Smith is a former correspondent for The Times and now hosts the podcast Eye on A.I.


, HIS, HIS, HIS, HIS, HIS, HIS, HIS, HIS, H
S, HIS, HIS, HER, HIS, HIS, HIS, HER, HIS, HIS,
, HIS, HIS, HIS, HIS, HIS, HIS, HIS, HIS, HIS, HIS