Engineering Bias in AI

Limiting representation in machine learning has the potential for harm on multiple levels

Cynthia Weber

After working at Apple designing circuits and signal processing algorithms for products including the first iPad, Timnit Gebru (Figure 1) received her Ph.D. from the Stanford Artificial Intelligence Laboratory in the area of computer vision. She recently completed a postdoc with Microsoft Research in the FATE (Fairness, Transparency, Accountability, and Ethics in Artificial Intelligence (AI)) group, was a cofounder of Black in AI, and is currently working as a research scientist in the Ethical AI team at Google. Her research in algorithmic bias and the ethical implications of data mining have appeared in multiple publications, including The New York Times *and* The Economist. IEEE Pulse *recently spoke with Gebru about the role societal bias plays in engineering AI, the deficits and dangers in the field caused by limited diversity, and the challenges inherent in addressing these complex issues.*

IEEE Pulse: What first drew you to working with Artificial Intelligence (AI)?

Timnit Gebru: It was never a conscious decision to focus on AI. I was a hardware engineer and worked with analog circuit design and also was interested in medical imaging and device physics. At some point, my primary interest shifted to image processing and signal processing, and I began taking image processing classes, which introduced me to computer vision. Back then, I wasn't thinking about it as AI, I was thinking about it as computer vision, which was typical of the field at the time. Eventually, my work led me to think about issues of bias and how algorithms and automated decision-making tools are being used to impact people's lives, and that's when I started working on the issues of bias and fairness, which are issues that affect all aspects of AI.

Figure 1. Timnit Gebru.

IEEE Pulse: In what ways did you see bias appearing in the computer vision work that you were doing?

TG: One way of understanding this bias is that it's not only human bias injected into AI, it is a direct result of limited representation in the field. There are very few people of color and very few women in the room, and this has always been a problem for engineering. Then, when you consider the development and implementation of automation in civil society, these technologies are more likely to adversely affect people who are already marginalized and who are not necessarily given the opportunity to work in this field. Although I was already involved in initiatives that were trying to address diversity in the field, I never really saw the connection between bias and AI until I was doing data mining with images.

Digital Object Identifier 10.1109/MPULS.2018.2885857
Date of current version: 12 March 2019.

2154-2287/19©2019 IEEE

At the time, we were working on a project that automatically detected and classified cars in 50 million images in 200 cities of the United States, and we trained a model to determine the potential characteristics of people who lived in certain zip codes based on the cars we saw in these images. For example, in this zip code, we think that there are this percentage of people who are college-educated or this percentage of people who voted Democrat versus Republican, and the accuracy of this model widely varied depending on how many cars we saw and whether these zip codes were outliers. This led me to think about accuracy in general and how we could make this type of model even more accurate; one thing I didn't consider was the inherent bias of the data we were using. But then we talked about using a similar process to determine crime rates. To do this, we would need to get crime rates from a particular website. The problem is that these data show only who was arrested or what crime was reported, not the data for who actually committed a crime. This made me realize that by using these kinds of limited data sets in your models, you can unintentionally perpetuate the existing bias that already exists in society.

Toward the end of my Ph.D. program, I read a *ProPublica* article that talked about machine bias and crime recidivism, and this made me very concerned. I wasn't aware at the time about the widespread use of machine learning-based algorithms in real life. And then, I read the book *Weapons of Math Destruction* (2016, Cathy O'Neill) and, at the same time, was also getting to know Joy Buolamwini at the MIT Media Lab. In her work, Joy showed that open-sourced face detection software wouldn't detect her face unless she put on a white mask.

These experiences helped me to recognize that these systems we were creating were only as good as the data they were being trained on, and so, for example, they weren't working well with darker skinned individuals. This is how I started seeing bias in my own work that at the time wasn't really being explored in computer vision, although now a lot more people are starting to acknowledge and work on mitigating bias.

IEEE Pulse: This idea that the results are only as good as the data you put into it seems to parallel some of the biases we're detecting in other areas—clinical trials, for example.

TG: Exactly. In a recent paper called "Data Sheets for Data Sets," we explored bias across different fields and provided case studies of other industries, including clinical trials. The fact that we are creating drugs that don't work well and that may have adverse effects on many people because the clinical trials consist of very homogenous groups is a problem. This potentially means that many of the drugs that will be available in the future will not work for a large part of the world's population.

Ultimately, believing that technology is blind or that technology is not biased or that engineers are objective is dangerous because nobody is truly objective. Often, it's the most highly educated people in the world, including scientists, who have historically caused the most harm in this area because our education and our elitism doesn't allow us to self-reflect and accept that we don't know certain things or that we are biased. Consider the harm caused by eugenics as one example.

IEEE Pulse: What do you hope having these data sheets for data sets will achieve?

TG: This is where my hardware background comes in. When you design a component or a chip for resale, there are associated data sheets with it that describe stress tests that have been done on the component as well as various characteristics (tolerance, etc.) so that the engineer can make a decision about how appropriate this component is for a specific use case. We were proposing a similar thing for data sets.

Currently, there are many publicly available data sets being used to train models that are then employed in high-stake scenarios. However, we have no idea what the characteristics of the data set that the model was trained on are and we have no idea what the characteristics of the model are. In the paper, we identify cases where this issue of bias in "training data" has resulted in very serious consequences, which perhaps could have been avoided if the bias had been recognized. We hope that implementing this concept of data sheets will help identify bias before problems arise.

IEEE Pulse: What are some of the challenges in changing current models to be more inclusive and diverse?

TG: I don't think that one could just focus only on the technical aspects of machine learning case models. For me, there are a few issues that need to be considered at the same time. One is that in the

same way we have regulatory bodies that test foods and drugs before they are unleashed, we need to have a similar type of regulatory agency that tests algorithms and trains on them to determine whether or not they should be used in specific scenarios.

For example, consider the use of facial recognition systems by law enforcement. Currently, we don't know when these systems are being used, by whom and for what purpose, what kinds of face recognition characteristics these systems have, how they were trained, and so on. Given that the use of facial recognition by law enforcement can have serious consequences on someone's life, and given that we have shown that these systems work worse for people with certain facial characteristics, unregulated use will likely lead to discrimination lawsuits—it is basically racial profiling transferred into this technology from inappropriate data.

Right now, we are not well equipped to be using these algorithms and we are not effectively examining how to regulate them and in what scenarios they should be used. In fact, the reason we wrote this article on data sheets was to talk about what industry should do now that AI is being used in everyday scenarios and not just for research. People need to take process and documentation very seriously and standardization very seriously. Those two things for me come before any algorithm model changes. Transparency is very important.

After that, then of course there is the issue of how to uncover potential bias that might exist. If you have a black box system, how do you test it, what are the kinds of tests we should have to potentially uncover disparate impacts among certain groups of people, and what is the best way to do this testing? And once you've uncovered bias, what are the characteristics that you should have for a particular algorithm that is used for a particular scenario? The community is not even in agreement as to what it means to be biased or fair. Does being fair mean that if your algorithm misclassifies a certain group of people, then it should have equal rates of misclassification among different groups of people? All of this is very context-dependent, and it's a very complex issue and you have to take a holistic approach to this process of change.

The last thing is that, again, I can never have this conversation without the understanding of why we're here in the first place. We're here because of the societal biases that exist and also because of the homogeneity of the field. There needs to be an understanding that it's not only about wanting to have diversity out of the goodness of your heart, actually you want to have diversity to advance the field as a whole and make better products and better algorithms.

We also need to remember that the final application informs the types of research that would advance the field. For example, the fact that now we are in a place where unless you see someone with a particular characteristic in your training data you cannot really classify them as a person, that means the type of technique we are using is too simplistic—it has to see some characteristics in the training data for it not to misclassify it in the test data. Perhaps, if we can reason more about what it means to be human, then we could have better classifications of whether or not a picture consists of a human or not. This also affects the types of models and the types of algorithms that you think work well. The technical questions I think would change if we had a different notion of what it meant for AI to work well as would the kinds of applications we were focused on.

IEEE Pulse: Where do you see this work advancing in the next few years and what are you excited about regarding the potential for change?

TG: I'm hoping that there would be some standardization as well as regulation and guidelines about the technologies that can be used and where and how they can be used. In terms of models, there's a lot of work to be done in trying to make "fair" machine learning models or even more interpretable machine learning models. There is a lot more interest now in causality in terms of technical work and I'm excited about that, and people are starting to examine the effects of the data they train on in more detail. I'm hoping that from a technical standpoint, from a standardization standpoint, or even a use case standpoint, things will change, and that we will be able to recognize that it's not about importing a technology assuming it already works, but that we all need to participate in the creation of this technology as well for AI to be successful. ■

■ ***Cynthia Weber*** *(clweberb@mtu.edu) holds a doctorate in rhetoric and technical communication and is Associate Editor of* IEEE Pulse.