strategy+business

APRIL 12, 2019

TECH & INNOVATION

What is fair when it comes to AI bias?

It's not the algorithm behaving badly, but how we define fairness that determines an artificial intelligence system's impact.

BY ANAND RAO AND ILANA GOLBIN

pwc | strategy&

What is fair when it comes to AI bias?

It's not the algorithm behaving badly, but how we define fairness that determines an artificial intelligence system's impact.

by Anand Rao and Ilana Golbin

Bias is often identified as one of the major risks associated with artificial intelligence (AI) systems. Recently reported cases of known bias in AI — racism in the criminal justice system, gender discrimination in hiring — are undeniably worrisome. The public discussion about bias in such scenarios often assigns blame to the algorithm itself. The algorithm, it is said, has made the wrong decision, to the detriment of a particular group. But this claim fails to take into account the human component: People perceive bias through the subjective lens of fairness.

Here it can be helpful to consider how we are defining both *bias* and *fairness*. Bias occurs when we discriminate against (or promote) a defined group consciously or unconsciously, and it can creep into an AI system as a result of skewed data or an algorithm that does not account for skewed data. For example, an AI system that reviews job applicants by learning from a company's historical data could end up discriminating against a particular gender or race if that group were underrepresented in the company's hiring in the past.

Fairness, meanwhile, is a social construct. And in fact, when people judge an algorithm to be "biased," they are often conflating bias and fairness: They are using a specific definition of fairness to pass judgment on the algorithm. There are at least 20 mathematical definitions of fairness, and when we choose one, we violate some aspect of the others. In other words, it is impossible for every decision to be fair to all parties.

No AI system can be universally fair or unbiased. But we can design these systems to meet specific fairness goals, thus mitigating some of the perceived unfairness and creating a more responsible system overall. Of course, responsible AI is a complex topic. It encompasses a number of factors beyond establishing processes and training teams to look for bias in the data and machine learning models. These factors include integrating risk mitigation and ethical concerns into AI algorithms and data sets from the start, along with considering other workforce issues and the greater good.

But there may be another factor holding companies back from establishing responsible AI: At most organizations, there exists a gap between what the data scientists are building and what the company leaders want to achieve with an AI implementation. It is thus important for business leaders and data scientists to work together to select the right notion of fairness for the particular decision that is to be made and design the algorithm that best meets this notion.

This approach is critical, because if the data scientists know what leaders' fairness goals are, they can more effectively assess what models to build and what data to use. In every situation, this requires bringing together various business specialists with AI system ar-

Anand Rao
anand.s.rao@pwc.com
is a principal with PwC US based in Boston. He is PwC's global leader for artificial intelligence and innovation lead for the U.S. analytics practice. He holds a Ph.D. in artificial intelligence from the University of Sydney and was formerly chief research scientist at the Australian Artificial Intelligence Institute.

Ilana Golbin
ilana.a.golbin@pwc.com
is a manager with PwC US based in Los Angeles. As part of PwC's Artificial Intelligence Accelerator, she specializes in developing solutions, frameworks, and offerings to support responsible AI.

Also contributing to this article was PwC US associate Amitoj Singh.

chitects and data scientists to help them develop or select the appropriate machine learning model. Thinking through this issue before you roll out new AI systems — both for optimal performance and to minimize bias — could produce business benefits and help you establish and maintain ethical standards for AI.

You'll need to determine which definition of fairness you, as the business leaders, will focus on — and which attributes should be protected (for example, gender and race). For instance, take the issue of group fairness versus fairness to the individual. Let's consider a credit card company that is implementing an algorithm using historical data to predict whether individuals applying for a certain credit offer are "good" or "bad" risks. In any scenario, leaders need to determine which protected variables to consider in achieving a fair outcome — and whose decision it is to determine who that group should be. In this example, the goal is to treat male and female credit card applicants fairly.

The AI developer must be aware of two potential outcomes: true positives and false positives. True positives are instances in which the model correctly predicted an applicant to be a good risk. False positives occur when bad-risk customers are assigned a good-risk score. People in the company's risk management group will be concerned with minimizing false positives to limit their risk exposure, because wrong risk assignments directly translate to potential losses to the company. When they try to minimize these losses, however, they do not want to discriminate based on gender.

Another definition of fairness involves a balance between treatment and outcomes. Here we can look to the college admissions process. Fairness in treatment would require looking at every candidate with the same metrics in mind, regardless of situation or context (e.g., making race-blind admissions, or using only the SAT score and GPA). Fairness in outcome is adjusting decision making to accommodate for situational and contextual characteristics, such as having slightly different requirements for students with disadvantaged backgrounds based on the understanding that they may not have had the same access to education and opportunity as students from wealthy suburbs.

In some cases, the goals of different business groups may appear to be at odds. In the credit card offer example, the marketing team may want to issue as many cards as possible to increase the company's market share and boost the brand, while the risk team's goal is to minimize potential losses incurred when customers who should not have been granted credit in the first place don't pay their bills. Because there is no way to completely satisfy both teams, they have to strike a balance that everyone can live with and that ideally supports the organization's ethical standards and business goals.

Because fairness is a social construct, it will take engagement and active discussion among teams to decide what constitutes fairness in any scenario. In every case, asking the following questions can help guide the conversation and keep it focused on common goals for the AI system:

- What decision is being made with the help of the algorithmic solution?
- What decision maker or functional business group will be using the algorithmic solution?
- What are the protected attributes to which we want to be fair?
- Who are the individuals and groups that will

experience the ramifications of the treatment and outcome?

- Can we clearly communicate the steps we've taken in designing the system to treat those affected by it fairly?
- How will we explain the decision-making process to key stakeholders?

Any decision — especially those humans make — can be construed as unfair in some way to one or more people affected by it. But exploring these issues can bring you closer to achieving responsible AI that strikes a balance between business goals and ethical concerns, while cultivating the trust of customers and other stakeholders. Given that respondents to PwC's 2019 AI Predictions Survey reported that their top challenge for 2019 was ensuring AI systems are trustworthy, it's time to stop blaming a biased algorithm and start talking about what is fair. ✚

strategy+business magazine
is published by certain member firms
of the PwC network.
To subscribe, visit strategy-business.com
or call 1-855-869-4862.

- strategy-business.com
- facebook.com/strategybusiness
- linkedin.com/company/strategy-business
- twitter.com/stratandbiz

Articles published in *strategy+business* do not necessarily represent the views of the member firms of the PwC network. Reviews and mentions of publications, products, or services do not constitute endorsement or recommendation for purchase.

© 2019 PwC. All rights reserved. PwC refers to the PwC network and/or one or more of its member firms, each of which is a separate legal entity. Please see www.pwc.com/structure for further details. Mentions of Strategy& refer to the global team of practical strategists that is integrated within the PwC network of firms. For more about Strategy&, see www.strategyand.pwc.com. No reproduction is permitted in whole or part without written permission of PwC. "strategy+business" is a trademark of PwC.

pwc | strategy&