Society | DOI:10.1145/2983270 Keith Kirkpatrick

Battling Algorithmic Bias

How do we ensure algorithms treat us fairly?

Computerized algorithms have become an integral part of everyday life. Algorithms are able to process a far greater range of inputs and variables to make decisions, and can do so with speed and reliability that far exceed human capabilities. From the ads we are served, to the products we are offered, and to the results we are presented with after searching online, algorithms, rather than humans sitting behind the scenes, are making these decisions.

However, because algorithms simply present the results of calculations defined by humans using data that may be provided by humans, machines, or a combination of the two (at some point during the process), they often inadvertently pick up the human biases that are incorporated when the algorithm is programmed, or when humans interact with that algorithm. Moreover, algorithms simply grind out their results, and it is up to humans to review and address how that data is presented to users, to ensure the proper context and application of that data.

A key example is the use of risk scores used by the criminal justice system to predict the likelihood of an individual committing a future crime, which can be used to determine whether a defendant should be allowed to post bond and in what amount, and may also be used to inform sentencing if the defendant is convicted of a crime.

Pro Publica, a nonprofit investigative journalism organization, early this year conducted a study of risk scores assigned to more than 7,000 people arrested in Broward County, FL, during 2013 and 2014, to see how many arrestees were charged with new crimes over the next two years.

The risk scores were created by Northpointe, a company whose software algorithm is used widely within the U.S. criminal justice system. The scores were the result of 137 questions either answered by defendants or

pulled from criminal records, though the defendant's race is not one of the questions. Nonetheless, some of the questions highlighted by *Pro Publica*—"Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?"—may be seen as being disproportionately impacting blacks.

Northpointe's founder, Tim Brennan, told *Pro Publica* it is challenging to develop a score that does not include items that can be correlated with race, such as poverty, joblessness, and social marginalization, since such negative traits that may indicate a propensity for criminal activity are correlated with race.

Still, according to *Pro Publica*, the risk scores examined across 2013 and 2014 proved unreliable in forecasting violent crimes, with just 20% of those predicted to commit such crimes actually doing so within two years. *Pro Publica* also claimed the algorithm falsely flagged black defendants as future criminals, wrongly labeling them this way at almost twice the rate of white defendants.

For its part, Northpointe disputed *Pro Publica*'s analysis, and the publication admitted the algorithm proved to be more accurate at predicting overall recidivism, with 61% percent of defendants being rearrested for committing a crime within two years.

It is not only the criminal justice system that is using such algorithmic assessments. Algorithms also are used to serve up job listings or credit offers that can be viewed as inadvertently biased, as they sometimes utilize end-user characteristics like household income and postal codes that can be proxies for race, given the correlation between ethnicity, household income, and geographic settling patterns.

The New York Times in July 2015 highlighted several instances of algorithmic unfairness, or outright discrimination. It cited research conducted by Carnegie Mellon University in 2015 that found Google's ad-serving system showed an ad for high-paying jobs to men much more often than it did for women. Similarly, a study conducted at the University of Washington in 2015 found that despite women holding 27% of CEO posts in the U.S., a search for "CEO" using Google's Image Search tool returned results of which just 11% depicted women. A 2012 Harvard University study published in the *Journal of Social Issues* indicated advertisements for services that allow searching for people's arrest records were more likely to come up when searches were conducted on traditionally African-American names.

For their part, programmers seem to recognize the need to address these issues of unfairness, particularly with respect to algorithms that have the potential to adversely impact protected groups, such as those in specific ethic groups, religious minorities, and others that might be subject to inadvertent or deliberate discrimination.

"Machine learning engineers care deeply about measuring accuracy of their models," explains Moritz Hardt, a senior research scientist at Google. "What they additionally need to do is to measure accuracy within different subgroups. Wildly differing performance across different groups of the population can indicate a problem. In the context of fairness, it can actually help to make models more com-

IMAGE FROM SHUTTERSTOCK.COM

plex to account for cultural differences within a population."

Tal Zarsky, a law professor at the University of Haifa, notes in a 2014 paper published in the *Washington Law Review* that identifying and eliminating cases of both explicit discrimination (cases in which the algorithm is specifically designed to treat some groups unfairly) and implicit discrimination (where the results of the algorithm wind up treating protected groups unfairly) may be challenging, but ultimately achievable. "While setting forth rules which ban such practices might be relatively easy, enforcing such a ban in a world in which the nature of the algorithm used is secret might prove to be a challenge," Zarsky wrote.

Indeed, some observers have called on the organizations that write and use algorithms to be more transparent in terms of clearly spelling out the data collected, identifying which pieces of data are used in the algorithm, and disclosing how this data is weighted or used in the algorithm. Such insights may help to pinpoint areas of discrimination that may not be apparent otherwise.

"The blessing and the curse of being transparent is that you're really clear, and with that clarity, sometimes you find discrimination," explains Jana Eggers, CEO of Nara Logics, a Cambridge, MA-based artificial intelligence platform provider. "Because it's uncovered, we go in and fix it, even if we have a lot to fix. Before, when we had the unconscious bias of people [making decisions], it was hard, if not impossible, to track down and understand."

One solution for handling discrimination is to monitor algorithms to determine fairness, though it may be difficult to establish a common definition of fairness, due to a variety of competing interests and viewpoints. Indeed, business decisions (such as the decision to offer a mortgage or credit card) are often predicated on criteria that disproportionately impact some minority communities, while making sense for the company that wants to maximize profit and reduce risk.

"Our normative understanding of what is 'fair' is constantly changing, and therefore the models must be revisited," Zarksky says.

Fairness is not necessarily clean-cut, given the competing interests, whether looking at commercial interests (profit versus access to products and services) or within the justice system, which must balance public safety, administrative efficiency, and the rights of defendants.

It may be difficult to establish a common definition of fairness, due to a variety of competing influences and viewpoints.

That is why algorithms likely need to be reviewed and revised regularly with human input, at least for the time being, particularly with respect to their impact on protected classes. The U.S. federal government has established race, gender, and national origin as protected classes, and some states have added additional groups, such as sexual orientation, age, and disability status.

Common wisdom among programmers is to develop a pure algorithm that does not incorporate protected attributes into the model, and there are currently no regulations governing inadvertent discrimination as a result of an algorithm. However, Hardt says, "what my [research] collaborators and I realized early on is that in order to detect and prevent discrimination, it may actually help to take protected attributes into account. Conversely, blindly ignoring protected attributes can lead to undesirable outcomes."

Despite their widespread use and potential to complicate the lives of many, it may be too early to establish a regulatory body for algorithms, given their complexity.

"Even the very notion of what we're trying to regulate is delicate as many machine learning systems are complex pipelines that, unlike food ingredients, cannot be described succinctly," Hardt says. "It would be more effective right now to invest in research on fairness, accountability, and transparency in machine learning."

Indeed, the high potential costs associated with regulation may stall any regulatory activity, at least in the near term.

"Although the agency's direct costs could be relatively low, the potential costs to some regulated entities could be relatively high," says Andrew Tutt, a Washington, D.C.-based attorney and former Visiting Fellow at the Yale Law School Information Society Project. Tutt has suggested the creation of a federal regulator that would oversee certain algorithms in an effort to help prevent unfairness or discrimination, in much the way the National Highway Traffic Safety Administration (NHTSA) or the Food and Drug Administration regulate automobiles and pharmaceuticals, respectively, for safety.

"There is no doubt that in the formation of such an agency, a difficult balance will need to be struck between innovation on the one hand and other values, like safety, on the other," Tutt says. "But I think that on balance, the benefits would outweigh the costs."

Nevertheless, Tutt's proposal only recommends oversight over algorithms that directly impact human safety, such as algorithms used to direct autonomous vehicles, rather than algorithms that may result in discrimination.

Hardt is not completely opposed to regulatory oversight, given that algorithms, and the way they are used, can do significant harm to many people. "I would like to see meaningful regulation eventually," Hardt says. "However, I'm afraid that our technical understanding is still so limited that regulation at this point in time could easily do more harm than good."

Further Reading

Zarsky, T.
Understanding Discrimination in the Scored Society, *Washington Law Review*, *Vol.89, No. 4, 2014*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2550248

Big Data: Seizing Opportunities, Preserving Values, Executive Office of the President, May 2014, https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Narayanan, A.
CITP Luncheon Speaker Series: Arvind Narayanan – Algorithmic society, Center for Information Technology Policy, https://www.youtube.com/watch?v=hujgRt9AsJQ

Keith Kirkpatrick is principal of 4K Research & Consulting, LLC, based in Lynbrook, NY.

© 2016 ACM 0001-0782/16/10 $15.00