Society | DOI:10.1145/2875029

Neil Savage

When Computers Stand in the Schoolhouse Door

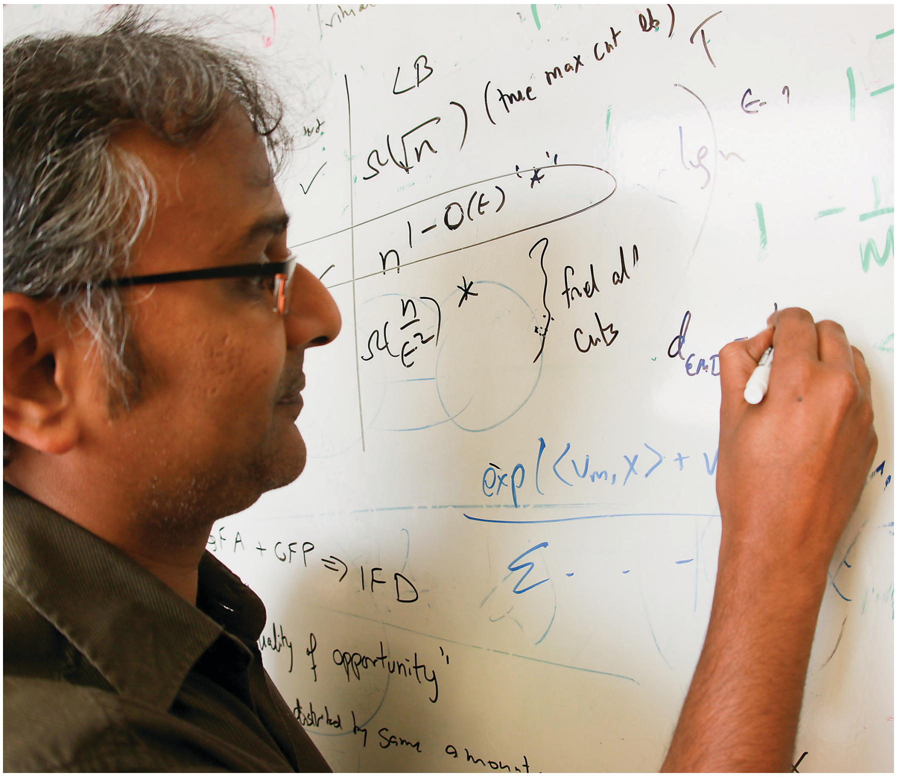
Classification algorithms can lead to biased decisions, so researchers are trying to identify such biases and root them out.

If you have ever searched for hotel rooms online, you have probably had this experience: surf over to another website to read a news story and the page fills up with ads for travel sites, offering deals on hotel rooms in the city you plan to visit. Buy something on Amazon, and ads for similar products will follow you around the Web. The practice of profiling people online means companies get more value from their advertising dollars and users are more likely to see ads that interest them.

The practice has a downside, though, when the profiling is based on sensitive attributes, such as race, sex, or sexual orientation. Algorithms that sort people by such categories risk introducing discrimination, and if they negatively affect a protected group's access to jobs, housing, or credit, they may run afoul of antidiscrimination laws. That is a growing concern as computer programs are increasingly used to help make decisions about who gets a credit card, which résumés lead to job interviews, or whether someone gets into a particular college. Even when the programs do not lead to illegal discrimination, they may still create or reinforce biases.

Computer scientists and public policy experts are beginning to pay more attention to bias in algorithms, to determine where it is showing up and what ought to be done about it. "Certainly there's a pretty hardy conversation that's begun in the research community," says Deirdre Mulligan, co-director of the Center for Law and Technology at the University of California, Berkeley.

One way algorithms may discriminate is in deciding who should see particular job-related ads. Anupam Datta, a professor of computer science at Carnegie Mellon University in Pittsburgh, PA, created AdFisher (http://bit.ly/1IRhF6P), a program that simulates browsing behavior and collects information about the ads returned after Google searches. Datta and his colleagues created 1,000 fake users and told Google half of them were men and half were women. Using AdFisher, each of the simulated users visited websites related to employment, then collected data about which ads they were shown subsequently. The tool discovered more ads related to higher-paying jobs were served to men than were presented to women.



Suresh Venkatasubramanian of the University of Utah presented a method for finding disparate impact in algorithms last year at the ACM Conference on Knowledge Discovery and Data Mining.

In particular, the top two ads shown to men were for a career-coaching service for people seeking executive positions that paid upward of $200,000 per year. Google showed that ad 1,852 times to the male group, but only 318 times to the female group. The top two ads shown to women were for a generic job-posting service and an auto dealer. "Ads for employment are a gateway to opportunity," Datta says. "If you don't make [job-seekers] aware of opportunities, you might be reinforcing biases."

The source of the differentiation be-

PHOTO BY VINCENT HORIUCHI/UNIVERSITY OF UTAH COLLEGE OF ENGINEERING

tween ads is not entirely clear. It could be the advertisers specified groups they wanted to target their ads toward. Waffles Pi Natusch, president of the Barrett Group executive coaching firm, told the *Pittsburgh Post-Gazette* last year the company does not specifically target men, but does seek out clients who have executive-level experience, are older than 45, and already earn more than $100,000 a year. Datta says there may be some correlation between those preferences and a person's gender.

How much the advertiser was willing to spend on the ad may have played a role as well. Google's algorithm presents advertisers with profiles of users and allows them to bid for placement on pages seen by those users. If a job ad paid the same whether it was targeted toward a male or a female user, but a clothing ad was willing to pay a premium to be seen by a woman, it could be that the job ad got outbid in the women's feed but not the men's feed.

It is also possible, Datta says, that Google's algorithm simply generated more interest for a particular ad from one group. "If they saw more males were clicking on this ad than females, they may have decided to serve more of these ads to male viewers," he says.

> **"Your perception of the gender balance of an occupation matters not only to how you hire, how you recruit, but it also affects the choice of people who go into the profession."**

Google would not discuss the issue beyond offering the following official statement: "Advertisers can choose to target the audience they want to reach, and we have policies that guide the type of interest-based ads that are allowed. We provide transparency to users with 'Why This Ad' notices and Ad Settings, as well as the ability to opt out of interest-based ads."

One of the challenges, researchers say, is that many of the datasets and algorithms used for classification tasks are proprietary, making it difficult to pinpoint where exactly the biases may reside. "The starting point of this work was the observation that many important decisions these days are being made inside black boxes," Datta says. "We don't have a very good sense of what types of personal information they're using to make decisions."

Look in the Mirror

Some imbalance may come from users' own biases. Sean Munson, a professor at the University of Washington's Department of Human Centered Design and Engineering in Seattle, WA, looked at the results returned by searches for images representing different jobs. In jobs that were more stereotypically male, there was a higher proportion of men in the search results than there

was in that profession in reality, while women were underrepresented. What is more, when users were asked to rate the quality of the results, they were happier with images where the gender shown matched the stereotype of the occupation—male construction workers or female nurses, say.

Some of the imbalance probably comes from which images are available. "We also play a role when we click on things in image result sets," Munson says. "My personal belief—and not knowing the Google algorithm—is that it's just reflecting our own biases back at us."

While it is unlikely image searches would violate anti-discrimination laws, Munson says skewed results could still have negative consequences. "Your perception of the gender balance of an occupation matters not only to how you hire, how you recruit, but it also affects the choice of people to go into the profession," he says.

Other algorithms, though, may run afoul of the law. A credit-scoring algorithm that winds up recommending against borrowers based on their race, whether purposefully or not, would be a problem, for instance. Anti-discrimination law uses the concept of adverse (or disparate) impact to avoid having to prove intent to discriminate; if a policy or procedure can be shown to have a disproportionately negative impact on people in a protected class or group, it will be considered discriminatory.

Joseph Domingo-Ferrer and Sara Hajian, computer scientists at Rovira i Virgili University in Tarragona, Spain, have developed a method for preventing such discrimination in data mining applications that might be used to assess credit worthiness.

One obvious approach might be to simply remove any references to race from the training data supplied to a machine learning algorithm, but that can be ineffective. "There might be other attributes which are highly correlated with this one," Hajian says. For instance, a certain neighborhood may be predominantly black, and if the algorithm winds up tagging anyone from that neighborhood a credit risk, the effect is the same as if it had used race. In addition, transforming the data too much by removing such attributes might lead to less-accurate results.

"Ethical issues can be integrated with data mining and machine learning without losing a lot of data utility."

What the researchers do instead is let the algorithm develop its classification rules, then examine its results. Using legal definitions of discrimination and protected classes, they examine the algorithm to see which rules led to the unwanted decisions; they then can decrease the number of records in the data that supports those rules. That way, Hajian says, they can transform the data enough to remove the discriminatory results while still preserving its usefulness.

"Ethical issues can be integrated with data mining and machine learning without losing a lot of data utility," she says.

At the ACM Conference on Knowledge Discovery and Data Mining in Sydney, Australia in August 2015, Suresh Venkatasubramanian, a computer scientist in the School of Computing at the University of Utah, presented a method for finding disparate impact in algorithms. Like Hajian and Domingo-Ferrer, Venkatasubramanian checks on the classification algorithm by examining its results, which does not require him to look at any proprietary code or data. If he can look at the decisions the algorithm made and use those to accurately infer protected attributes, such as race or sex, in the dataset, that means the algorithm has produced a disparate impact. He also offers a method, similar to the other researchers', that lets the data be transformed in a minimal way, to eliminate the bias while preserving the utility.

This type of approach only works when there is a clear legal standard to define bias. Other cases, such as image search results, require developing a societal consensus on whether there is a problem and what should be done about it.

"Some of it is a technological challenge, a computer science challenge, but some of it is a challenge of how these algorithms should operate in a larger social context," Datta says.

Should a company like Google be liable if it shows ads from job-coaching services more to men than to women, or does that not rise to the level of actual job discrimination? Is the company just a neutral platform delivering information? Should it tweak the algorithm to deliver the ads more proportionally, and what if that causes it to lose money because of lower click-through rates?

"There is a cost to trying to enforce fairness, and someone has to bear that cost," Datta says.

Not every instance of differentiation is discrimination, says Mulligan. Few people, for instance, object because women's magazines contain advertisements targeted specifically at women.

"Any system of classification has a bias. That's actually what makes it useful. It's curating. It's how it helps you sort stuff," Datta says.

"The question becomes, what sort of unfairness do we want to avoid?" C

Further Reading

Hajian, S., and Domingo-Ferrer, J.
A Methodology for Direct and Indirect Discrimination Prevention in Data Mining, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No. 7 (2013)

Kay, M., Matuszek, C., and Munson, S.A.
Unequal Representation and Gender Stereotypes in Image Search Results for Occupations, *CHI 2015*, Seoul, Korea (2015)

Datta, A., Tschantz, M.C., and Datta, A.
Automated Experiments on Ad Privacy Settings, *Proceedings on Privacy Enhancing Technologies, 1, (2015)*

Dwork, C., and Mulligan, D. K.
It's Not Privacy, and it's Not Fair, *Stanford Law Review Online*, Vol. 66, No. 35, (2013)

Venkatasubramanian, S.
Certifying and Removing Disparate Impact, https://www.youtube.com/watch?v=4ds9fBDtMmU

Neil Savage is a science and technology writer based in Lowell, MA.