Vox

How artificial intelligence learns to be racist

Simple: It's mimicking us.

Updated by Brian Resnick | @B_resnick | brian@vox.com | Apr 17, 2017, 2:10pm EDT

Noctiluxx / Getty Creative Images

Open up the photo app on your phone and search "dog," and all the pictures you have of dogs will come up. This was no easy feat. Your phone knows what a dog "looks" like.

This and other modern-day marvels are the result of machine learning. These are programs that comb through millions of pieces of data and start making correlations and predictions about the world. The appeal of these programs is immense: These machines can use cold, hard data to make decisions that are sometimes more accurate than a human's.

But know: Machine learning has a dark side. "Many people think machines are not biased," Princeton computer scientist Aylin Caliskan says. "But machines are trained on human data. And humans are biased."

Computers learn how to be racist, sexist, and prejudiced in a similar way that a child does, Caliskan explains: from their creators.

We think artificial intelligence is impartial. Often, it's not.

Nearly all new consumer technologies use machine learning in some way. Like Google Translate: No person instructed the software to learn how to translate Greek to French and then to English. It combed through countless reams of text and learned on its own. In other cases, machine learning programs make predictions about which résumés are likely to yield successful job candidates, or how a patient will respond to a particular drug.

Machine learning is a program that sifts through billions of data points to solve problems (such as "can you identify the animal in the photo"), but it doesn't **always make clear *how*** it has solved the problem. And it's increasingly clear these programs can develop biases and stereotypes without us noticing.

Last May, ProPublica **published** an investigation on a machine learning program that courts use to predict who is likely to commit another crime after being booked systematically. The reporters found that the software rated black people at a higher risk than whites.

"Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation," ProPublica **explained**. "They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts … to even more fundamental decisions about defendants' freedom."

The program learned about who is most likely to end up in jail from real-world incarceration data. And historically, the real-world criminal justice system has been unfair to black Americans.

This story reveals a deep irony about machine learning. The appeal of these systems is they can make impartial decisions, free of human bias. "If computers could accurately predict which defendants were likely to commit new crimes, the criminal justice system could be fairer and more selective about who is incarcerated and for how long," ProPublica wrote.

But what happened was that machine learning programs perpetuated our biases on a large scale. So instead of a judge being prejudiced against African Americans, it was a robot.

It's stories like the ProPublica investigation that led Caliskan to research this problem. As a female computer scientist who was routinely the only woman in her graduate school classes, she's sensitive to this subject.

Caliskan has seen bias creep into machine learning in often subtle ways — for instance, in Google Translate.

Turkish, one of her native languages, has no gender pronouns. But when she uses Google Translate on Turkish phrases, it “always ends up as ‘he’s a doctor’ in a gendered language.” The Turkish sentence didn’t say whether the doctor was male or female. The computer just assumed if you’re talking about a doctor, it’s a man.

How robots learn implicit bias

Recently, Caliskan and colleagues **published** a paper in *Science,* that finds as a computer teaches itself English, it becomes prejudiced against black Americans and women.

Basically, they used a common machine learning program to crawl through the internet, look at 840 billion words, and teach itself the definitions of those words. The program accomplishes this by looking for how often certain words appear in the same sentence. Take the word “bottle.” The computer begins to understand what the word means by noticing it occurs more frequently alongside the word “container,” and also near words that connote liquids like “water” or “milk.”

This idea to teach robots English actually comes from cognitive science and its understanding of how children learn language. How frequently two words appear together is the first clue we get to deciphering their meaning.

Once the computer amassed its vocabulary, Caliskan ran it through a version of the implicit association test.

In humans, the IAT is meant to undercover subtle biases in the brain by seeing how long it takes people to associate words. A person might quickly connect the words “male” and “engineer.” But if a person lags on associating “woman” and “engineer,” it’s a demonstration that those two terms are not closely associated in the mind, implying bias. (There are some reliability issues with the IAT in humans, which you can **read about here**.)

Here, instead at looking at the lag time, Caliskan looked at how closely the computer thought two terms were related. She found that African-American names in the program were less associated with the word “pleasant” than white names. And female names were more associated with words relating to family than male names. (In a weird way, the IAT might be better suited for use on computer programs than for humans, because humans answer its questions inconsistently, while a computer will yield the same answer every single time.)

Like a child, a computer builds its vocabulary through how often terms appear together. On the internet, African-American names are more likely to be surrounded by words that connote unpleasantness. That's not because African Americans are unpleasant. It's because people on the internet say awful things. And it leaves an impression on our young AI.

This is as much as a problem as you think.

The consequences of racist, sexist AI

Increasingly, Caliskan says, job recruiters are relying on machine learning programs to take a first pass at résumés. And if left unchecked, the programs can learn and act upon gender stereotypes in their decision-making.

"Let's say a man is applying for a nurse position; he might be found less fit for that position if the machine is just making its own decisions," she says. "And this might be the same for a women applying for a software developer or programmer position. … Almost all of these programs are not open source, and we're not able to see what's exactly going on. So we have a big responsibility about trying to uncover if they are being unfair or biased."

And that will be a challenge in the future. Already AI is making its way into the health care system, helping doctors find the right course of treatment for their patients. (There's early research on whether it can help **predict mental health crises**.)

But health data, too, is filled with historical bias. It's long been known that women get **surgery at lower rates than men**. (One reason is that women, as primary caregivers, have fewer people to take care of them post-surgery.)

Might AI then recommend surgery at a lower rate for women? It's something to watch out for.

So are these programs useless?

Inevitably, machine learning programs are going to encounter historical patterns that reflect racial or gender bias. And it can be hard to draw the line between what is bias and what is just a fact about the world.

Machine learning programs will pick up on the fact that most nurses throughout history have been women. They'll realize most computer programmers are male. "We're not suggesting you should remove this information," Caliskan says. It might actually break the software completely.

Caliskan thinks there need to be more safeguards. Humans using these programs need to constantly ask, "Why am I getting these results?" and check the output of these programs for bias. They need to think hard on whether the data they are combing is reflective of historical prejudices. Caliskan admits the best practices of how to combat bias in AI is still being worked out. "It requires a long-term research agenda for computer scientists, ethicist, sociologists, and psychologists," she says.

But at the very least, the people who use these programs should be aware of these problems, and not take for granted that a computer can produce a less biased result than a human.

And overall, it's important to remember: AI learns about how the world *has been.* It picks up on status quo trends. It doesn't know how the world *ought* to be. That's up to humans to decide.

Was this article helpful? 👍 👎

THE LATEST

9 questions for Mark Cuban

by Sean Illing

Trump is dangerous. But he's not Nixon — yet

by Sean Illing

8 can't-miss pieces of pop culture to catch up on this weekend

by Vox Culture

Why couldn’t the Civil War have been worked out? Some smart people take the question seriously.

by William Black

Dick, the 1999 teen Watergate comedy, captures the enduring weirdness of political scandals

by Alissa Wilkinson

I was scared of the FBI with Comey in charge. I’m more terrified now that he’s been fired.

by Maria McFarland Sánchez-Moreno