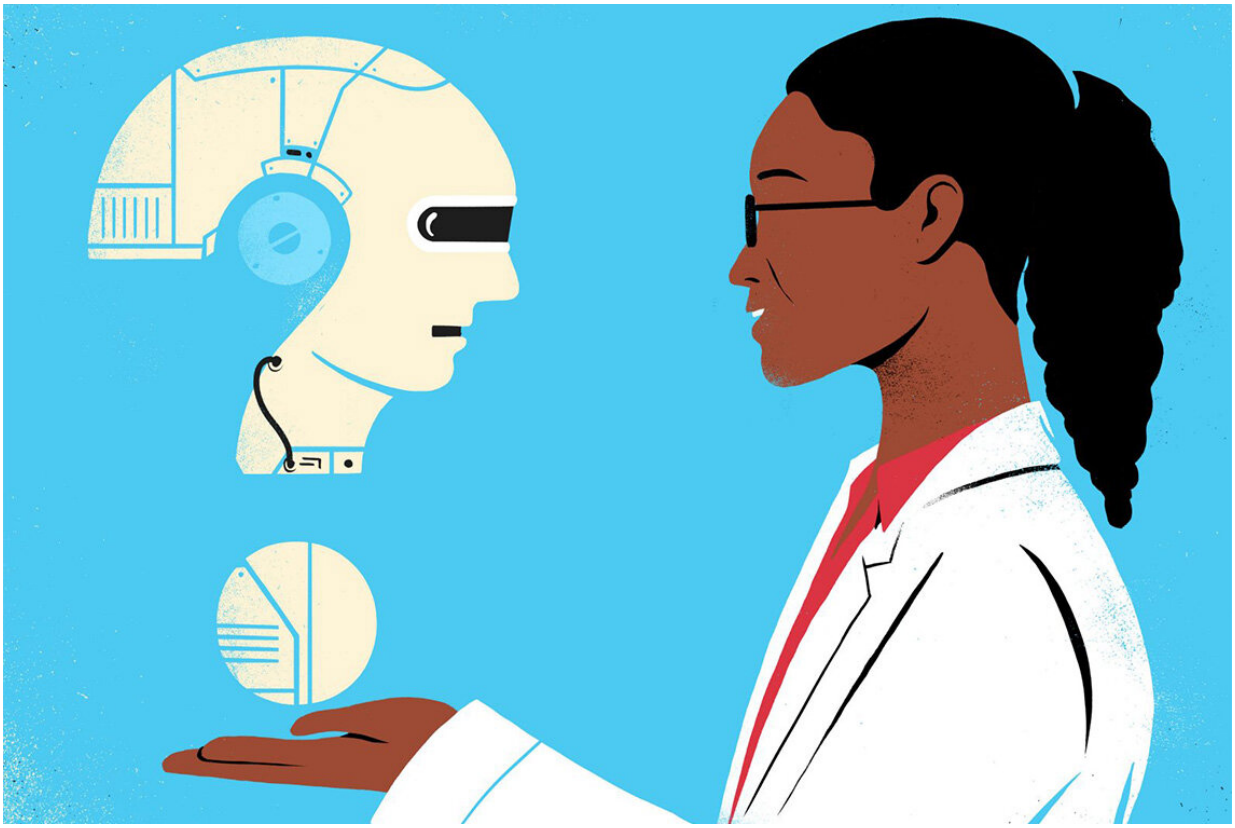
# Eliminating bias in AI

July 25 2019, by Patchen Barss



Credit: Sébastien Thibault

In human beings, intelligence is no inoculation against bias and bigotry. The same holds true for computers. Intelligent machines learn about the world through the filters of human language and historical behaviour—meaning they can just as easily absorb humanity's worst values as they can its best.

Researchers who aim to develop ever-smarter machines have their work cut out for them to ensure that they're not inadvertently imbuing computers with misogyny, racism or other forms of bigotry.

"It's a huge risk," says Marzyeh Ghassemi, an assistant professor in the University of Toronto's department of computer science and Faculty of Medicine who focuses on health-care applications for artificial intelligence (AI). "Like all advances that leapfrog societies forward, there are large risks that we must decide to accept or not to accept."

Bias can creep into algorithms in many ways. In a highly influential branch of AI known as "natural language processing," problems can arise from the "text corpus"—the source material the algorithm uses to learn about the relationships between different words.

Natural language processing, or "NLP," allows a computer to understand human-style speech—informal, conversational and contextual. NLP algorithms comb through billions of words of training text—the corpus might be, say, the entirety of Wikipedia. One algorithm works by assigning to each word a set of numbers that reflects different aspects of its meaning—"king" and "queen" for instance, would have similar scores relating to the idea of royalty, but opposite scores relating to gender. NLP is a powerful system that allows machines to learn about relationships between words—in some cases, without direct human involvement.

"Even though we're not always teaching them specifically, what they learn is incredible," says Kawin Ethayarajh, a researcher who focuses partly on fairness and justice in AI applications. "But it's also a problem. In the corpus, the relationship between 'king' and 'queen' might be similar to the relationship between 'doctor' and 'nurse.'"

But of course, all kings are men; not all doctors are men. And not all

nurses are women.

When an algorithm absorbs the sexist tropes of historical human attitudes, it can lead to real-life consequences, as happened in 2014 when Amazon developed an algorithm to vet job applicants' resumés. The company trained its machines using 10 years of hiring decisions. But in 2015, they acknowledged that, in tests, the system was giving unearned preference to resumés from male applicants. They tweaked the system to force it to ignore gender information, but ultimately shut down the project before actually putting it to use as they could not be sure their algorithm wasn't perpetrating other forms of discrimination.

Mitigating sexist source material can involve technological and methodological adjustments. "If we can understand exactly what underlying assumptions the corpus has that cause these biases to be learned, we can either select corpora without those biases or correct it during the training process," says Ethayarajh.

It's common practice for researchers to design an algorithm that corrects prejudicial assumptions automatically. By adjusting the weight of the numbers it assigns to each word, the computer can avoid making sexist or racist associations.

But what exactly are the assumptions that need correcting? What does a fair-minded AI really look like? Debates over privilege, bigotry, diversity and systemic bias are far from settled. Should a hiring algorithm have a stance on affirmative action? Should a self-driving car take special care if another vehicle has a "Baby on Board" sticker? How should an AI-driven analysis of legal documents factor in the historical treatment of Indigenous Peoples? Contentious societal issues don't disappear merely because machines take over certain recommendations or decisions.

Many people view Canada's flawed but relatively successful model of multiculturalism as a chance to lead in fair AI research.

"Canada certainly does have an opportunity," says Ronald Baecker, a professor emeritus of computer science and the author of Computers and Society: Modern Perspectives. He sees a role for government to redress the societal inequities, injustices and biases associated with AI by, for example, setting up protections for employees who choose to speak out against biased or unfair AI-driven products. "There's a need for more thinking and legislation with respect to the concept of what I would call 'conscientious objection' by high-tech employees."

He also believes that the computer scientists developing smart technologies should be required to study the societal impact of such work. "It's important that professionals who work in AI recognize their responsibility," he says. "We're dealing with life-and-death situations in increasingly important activities where AI is being used."

Algorithms that help judges set bail and sentence criminals can absorb long-standing biases in the legal system, such as treating racialized people as if they are more likely to commit additional crimes. The algorithms might flag people from certain communities as posing too high a risk to receive a bank loan. They also might be better at diagnosing skin cancer in white people than in people with darker skin, as a result of having been trained on skewed source material.

The stakes are incredibly high in health care, where inequitable algorithms could push people who have been poorly served in the past even further into the margins.

In her work at U of T and at the Vector Institute for Artificial Intelligence, Ghassemi, like other researchers, takes pains to identify potential bias and inequity in her algorithms. She compares the

recommendations and predictions of her diagnostic tools against real-world outcomes, measuring their accuracy for different genders, races, ages and socio-economic factors.

In theory, Canada offers a head start for researchers interested in health-care applications that reflect values of fairness, diversity and inclusion. Our universal health-care system creates a repository of electronic health records that provides a wealth of medical data that could be used to train AI-driven applications. This potential drew Ghassemi to Toronto. But the technology, information, formatting and rules to access these records vary from province to province, making it complicated to create the kind of data sets that can move research forward.

Ghassemi was also surprised to learn that these records only rarely include data about race. This means if she's using an algorithm to determine how well a given treatment serves different sectors of society, she could identify disparities between men and women, for example, but not between white people and racialized people. As a result, in her teaching and research, she's using publicly available American data that contains information about race.

"Auditing my own models [using American data], I can show when something has higher inaccuracy for people with different ethnicities," she says. "I can't make this assessment in Canada. There's no way for me to check."

Ghassemi is interested in creating AI applications that are fair in their own right—and that also can help human beings counteract their own biases. "If we can provide tools based on large diverse populations, we're giving doctors something that will help them make better choices," she says.

Women, for example, are significantly underdiagnosed for heart

conditions. An AI could flag such a danger for a doctor who might overlook it. "That's a place where a technological solution can help, because doctors are humans, and humans are biased," she says.

Ethayarajh concurs with Ghassemi and Baecker that Canada has an important opportunity to press its advantage on fairness and bias in artificial intelligence research.

"I think AI researchers here are very aware of the problem," Ethayarajh says. "I think a part of that is, if you look around the office, you see a lot of different faces. The people working on these models will be end-users of these models. More broadly, I think there is a very strong cultural focus on fairness that makes this an important area for researchers in this country."

Provided by University of Toronto