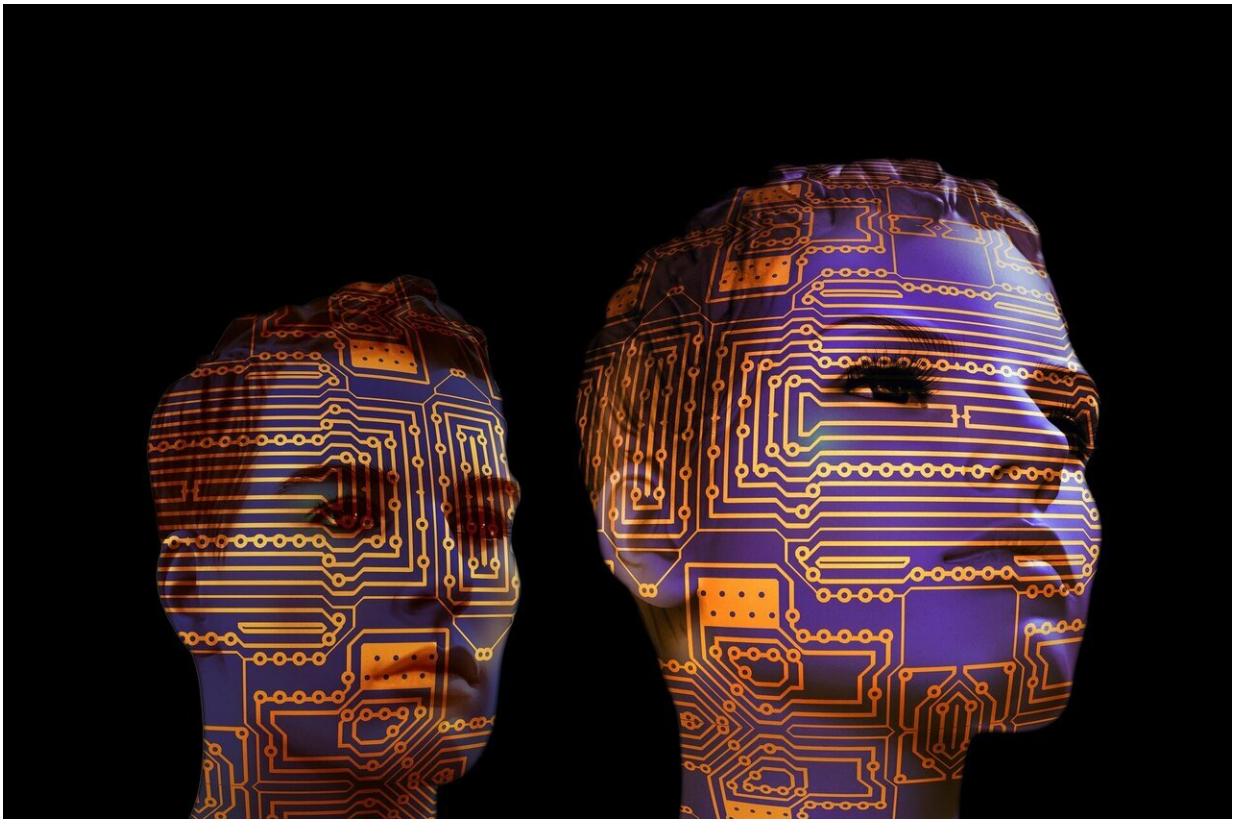
Using artificial intelligence to detect discrimination

July 10 2019



Credit: CC0 Public Domain

A new artificial intelligence (AI) tool for detecting unfair discrimination—such as on the basis of race or gender—has been created by researchers at Penn State and Columbia University.

Preventing unfair treatment of individuals on the basis of race, gender or ethnicity, for example, been a long-standing concern of civilized societies. However, detecting such discrimination resulting from decisions, whether by human decision makers or automated AI systems, can be extremely challenging. This challenge is further exacerbated by the wide adoption of AI systems to automate decisions in many domains—including policing, consumer finance, higher education and business.

"Artificial intelligence systems—such as those involved in selecting candidates for a job or for admission to a university—are trained on large amounts of data," said Vasant Honavar, Professor and Edward Frymoyer Chair of Information Sciences and Technology, Penn State. "But if these data are biased, they can affect the recommendations of AI systems."

For example, he said, if a company historically has never hired a woman for a particular type of job, then an AI system trained on this historical data will not recommend a woman for a new job.

"There's nothing wrong with the machine learning algorithm itself," said Honavar. "It's doing what it's supposed to do, which is to identify good job candidates based on certain desirable characteristics. But since it was trained on historical, biased data it has the potential to make unfair recommendations."

The team created an AI tool for detecting discrimination with respect to a protected attribute, such as race or gender, by human decision makers or AI systems that is based on the concept of causality in which one thing—a cause—causes another thing—an effect.

"For example, the question, 'Is there gender-based discrimination in salaries?' can be reframed as, 'Does gender have a causal effect on salary

?,' or in other words, 'Would a woman be paid more if she was a man?' said Aria Khademi, graduate student in information sciences and technology, Penn State.

Since it is not possible to directly know the answer to such a hypothetical question, the team's tool uses sophisticated counterfactual inference algorithms to arrive at a best guess.

"For instance," said Khademi, "one intuitive way of arriving at a best guess as to what a fair salary would be for a female employee is to find a male employee who is similar to the woman with respect to qualifications, productivity and experience. We can minimize gender-based discrimination in salary if we ensure that similar men and women receive similar salaries."

The researchers tested their method using various types of available data, such as income data from the U.S. Census Bureau to determine whether there is gender-based discrimination in salaries. They also tested their method using the New York City Police Department's stop-and-frisk program data to determine whether there is discrimination against people of color in arrests made after stops. The results appeared in May in Proceedings of The Web Conference 2019.

"We analyzed an adult income data set containing salary, demographic and employment-related information for close to 50,000 individuals," said Honavar. "We found evidence of gender-based discrimination in salary. Specifically, we found that the odds of a woman having a salary greater than $50,000 per year is only one-third that for a man. This would suggest that employers should look for and correct, when appropriate, gender bias in salaries."

Although the team's analysis of the New York stop-and-frisk dataset—which contains demographic and other information about

drivers stopped by the New York City police force—revealed evidence of possible racial bias against Hispanics and African American individuals, it found no evidence of discrimination against them on average as a group.

"You cannot correct for a problem if you don't know that the problem exists," said Honavar. "To avoid discrimination on the basis of race, gender or other attributes you need effective tools for detecting discrimination. Our tool can help with that."

Honavar added that as data-driven artificial intelligence systems increasingly determine how businesses target advertisements to consumers, how police departments monitor individuals or groups for criminal activity, how banks decide who gets a loan, who employers decide to hire, and how colleges and universities decide who gets admitted or receives financial aid, there is an urgent need for tools such as the one he and his colleagues developed.

"Our tool," he said, "can help ensure that such systems do not become instruments of discrimination, barriers to equality, threats to social justice and sources of unfairness."

Provided by Pennsylvania State University