Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

Larry Hardesty | MIT News Office

Three commercially released facial-analysis programs from major technology companies demonstrate both skin-type and gender biases, according to a new paper researchers from MIT and Stanford University will present later this month at the Conference on Fairness, Accountability, and Transparency.

In the researchers’ experiments, the three programs’ error rates in determining the gender of light-skinned men were never worse than 0.8 percent. For darker-skinned women, however, the error rates ballooned — to more than 20 percent in one case and more than 34 percent in the other two.

The findings raise questions about how today’s neural networks, which learn to perform computational tasks by looking for patterns in huge data sets, are trained and evaluated. For instance, according to the paper, researchers at a major U.S. technology company claimed an accuracy rate of more than 97 percent for a face-recognition system they’d designed. But the data set used to assess its performance was more than 77 percent male and more than 83 percent white.

LINKED VIDEO:
https://youtu.be/TWWsW1w-BVo

“What’s really important here is the method and how that method applies to other applications,” says Joy Buolamwini, a researcher in the MIT Media Lab’s Civic Media group and first author on the new paper. “The same data-centric techniques that can be used to try to determine somebody’s gender are also used to identify a person when you’re looking for a criminal suspect or to unlock your phone. And it’s not just about computer vision. I’m really hopeful that this will spur more work into looking at [other] disparities.”

Buolamwini is joined on the paper by Timnit Gebru, who was a graduate student at Stanford when the work was done and is now a postdoc at Microsoft Research.

Chance discoveries

The three programs that Buolamwini and Gebru investigated were general-purpose facial-analysis systems, which could be used to match faces in different photos as well as to assess characteristics such as gender, age, and mood. All three systems treated gender classification as a binary decision — male or female — which made their performance on that task particularly easy to assess statistically. But the same types of bias probably afflict the programs' performance on other tasks, too.

Indeed, it was the chance discovery of apparent bias in face-tracking by one of the programs that prompted Buolamwini's investigation in the first place.

Several years ago, as a graduate student at the Media Lab, Buolamwini was working on a system she called Upbeat Walls, an interactive, multimedia art installation that allowed users to control colorful patterns projected on a reflective surface by moving their heads. To track the user's movements, the system used a commercial facial-analysis program.

The team that Buolamwini assembled to work on the project was ethnically diverse, but the researchers found that, when it came time to present the device in public, they had to rely on one of the lighter-skinned team members to demonstrate it. The system just didn't seem to work reliably with darker-skinned users.

Curious, Buolamwini, who is black, began submitting photos of herself to commercial facial-recognition programs. In several cases, the programs failed to recognize the photos as featuring a human face at all. When they did, they consistently misclassified Buolamwini's gender.

Quantitative standards

To begin investigating the programs' biases systematically, Buolamwini first assembled a set of images in which women and people with dark skin are much better-represented than they are in the data sets typically used to evaluate face-analysis systems. The final set contained more than 1,200 images.

Next, she worked with a dermatologic surgeon to code the images according to the Fitzpatrick scale of skin tones, a six-point scale, from light to dark, originally developed by dermatologists as a means of assessing risk of sunburn.

Then she applied three commercial facial-analysis systems from major technology companies to her newly constructed data set. Across all three, the error rates for gender classification were consistently higher for females than they were for males, and for darker-skinned subjects than for lighter-skinned subjects.

For darker-skinned women — those assigned scores of IV, V, or VI on the Fitzpatrick scale — the error rates were 20.8 percent, 34.5 percent, and 34.7. But with two of the systems, the error rates for the darkest-skinned women in the data set — those assigned a score of VI — were worse still: 46.5 percent and 46.8 percent. Essentially, for those women, the system might as well have been guessing gender at random.

“To fail on one in three, in a commercial system, on something that’s been reduced to a binary classification task, you have to ask, would that have been permitted if those failure rates were in a different subgroup?” Buolamwini says. “The other big lesson ... is that our benchmarks, the standards by which we measure success, themselves can give us a false sense of progress.”

“This is an area where the data sets have a large influence on what happens to the model,” says Ruchir Puri, chief architect of IBM’s Watson artificial-intelligence system. “We have a new model now that we brought out that is much more balanced in terms of accuracy across the benchmark that Joy was looking at. It has a half a million images with balanced types, and we have a different underlying neural network that is much more robust.”

“It takes time for us to do these things,” he adds. “We’ve been working on this roughly eight to nine months. The model isn’t specifically a response to her paper, but we took it upon ourselves to address the questions she had raised directly, including her benchmark. She was bringing up some very important points, and we should look at how our new work stands up to them.”

Topics: ResearchSchool of Architecture and PlanningMedia LabCenter for Civic MediaArtificial intelligenceComputer science and technologyDiversity and inclusionMachine learningTechnology and society

LINKED VIDEO:
https://youtu.be/TWWsW1w-BVo