MIT News
ON CAMPUS AND AROUND THE WORLD

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

February 11, 2018

Joy Buolamwini, a researcher in the MIT Media Lab's Civic Media group

Photo: Bryce Vickmark

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.

"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."

PRESS MENTIONS

New York Times

Graduate student Joy Buolamwini joins Kara Swisher on *The New York Times'* "Sway" podcast to discuss her crusade against bias in facial recognition technologies. "If you have a face, you have a place in this conversation," says Buolamwini.

Full story via New York Times →

NBC News

NBC News reporters Lindsay Hoffman and Caroline Kim spotlight graduate student Joy Buolamwini's work uncovering racial and gender bias in AI systems in a piece highlighting women who are "shattering ceilings, making groundbreaking discoveries, and spreading public awareness during the global pandemic." Hoffman and Kim note that Buolamwini's research "helped persuade these companies to put a hold on facial recognition technology until federal regulations were passed."

Full story via NBC News →

Fast Company

Fast Company reporter Amy Farley spotlights graduate student Joy Buolamwini and her work battling bias in artificial intelligence systems, noting that "when it comes to AI injustices, her voice resonates." Buolamwini emphasizes that "we have a voice and a choice in the kind of future we have."

Learn more →

Quartz

Quartz reporter Nicolas Rivero notes that IBM's decision to end its facial recognition program was inspired by "one influential piece of research: the Gender Shades project, from MIT Media Lab's Joy Buolamwini and Microsoft Research's Timnit Gebru." Buolamwini and Gebru found that "commercial facial recognition software was significantly less accurate for darker-skinned women than for lighter-skinned men. "

Full story via Quartz →

The Verge

Verge reporter Nick Statt notes that, “Much of the foundational work showing the flaws of modern facial recognition tech with regard to racial bias is thanks to Joy Buolamwini, a researcher at the MIT Media Lab, and Timnit Gebru, a member at Microsoft Research.”

Full story via The Verge →

Financial Times

Graduate student Joy Buolamwini has been named to the *Financial Times*’ list of change-makers, which highlights “30 of the planet’s most exciting young people.” Financial Times reporter India Ross notes that Buolamwini, “identified gender and racial biases in artificial intelligence, and her efforts have prompted technology companies such as IBM to upgrade their software accordingly.”

Full story via Financial Times →

Time Magazine

Graduate student Joy Buolamwini writes for *TIME* about the need to tackle gender and racial bias in AI systems. “By working to reduce the exclusion overhead and enabling marginalized communities to engage in the development and governance of AI, we can work toward creating systems that embrace full spectrum inclusion,” writes Buolamwini.

Full story via Time Magazine →

Fortune- CNN

Fortune reporters Aaron Pressman and Adam Lashinsky highlight graduate student Joy Buolamwini’s work aimed at eliminating bias in AI and machine learning systems.

Pressman and Lashinsky note that Buolamwini believes that “who codes matters,” as more diverse teams of programmers could help prevent algorithmic bias.

Full story via Fortune- CNN →

Bloomberg

In a recent blog post, Microsoft’s president and chief legal officer, Brad Smith, references research by MIT graduate student Joy Buolamwini while calling for government to regulate the use of facial recognition software. Buolamwini’s work “showed error rates of as much as 35% for systems classifying darker skinned women,” reports Dina Bass for Bloomberg.

Full story via Bloomberg →

co.design

Katharine Schwab of *Co.Design* highlights graduate student Joy Buolamwini and Visiting Scholar J. Nathan Matias as “design heroes” for their commitment to keep technology fair. Schwab writes that Buolamwini has forced companies “to develop better, more equitable technology” while Matias helped “reduced the prevalence of fake news.”

Full story via co.design →

New York Times

In an article for *The New York Times*, graduate student Joy Buolamwini writes about how AI systems can often reinforce existing racial biases and exclusions. Buolamwini writes that, “Everyday people should support lawmakers, activists and public-interest technologists in demanding transparency, equity and accountability in the use of artificial intelligence that governs our lives.”

Full story via New York Times →

WGBH

A recent study from Media Lab graduate student Joy Buolamwini addresses errors in facial recognition software that create concern for civil liberties. “If programmers are training artificial intelligence on a set of images primarily made up of white male faces, their systems will reflect that bias,” writes Cristina Quinn for *WGBH.*

Full story via WGBH →

Boston Magazine

Spencer Buell of *Boston Magazine* speaks with graduate student Joy Buolamwini, whose research shows that many AI programs are unable to recognize non-white faces. “‘We have blind faith in these systems,’ she says. ‘We risk perpetuating inequality in the guise of machine neutrality if we’re not paying attention.’”

Full story via Boston Magazine →

Quartz

Dave Gershgorn writes for *Quartz,* highlighting congress’ concerns around the dangers of inaccurate facial recognition programs. He cites Joy Buolamwini’s Media Lab research on facial recognition, which he says “maintains that facial recognition is still significantly worse for people of color.”

Full story via Quartz →

The Economist

An article in *The Economist* states that new research by MIT grad student Joy Buolamwini supports the suspicion that facial recognition software is better at processing white faces than those of other people. The bias probably arises “from the sets of data the firms concerned used to train their software," the article suggests.

Full story via The Economist →

Marketplace

Molly Wood *at Marketplace* speaks with Media Lab graduate student Joy Buolamwini about the findings of her recent research, which examined widespread bias in AI-supported facial recognition programs. "At the end of the day, data reflects our history, and our history has been very biased to date," Buolamwini said.

Full story via Marketplace →

co.design

Recent research from graduate student Joy Buolamwini shows that facial recognition programs, which are increasingly being used by law enforcement, are failing to identify non-white faces. "When these systems can't recognize darker faces with as much accuracy as lighter faces, there's a higher likelihood that innocent people will be targeted by law enforcement," writes Katharine Schwab for *Co. Design*.

Full story via co.design →

New Scientist

Graduate student Joy Buolamwini tested three different face-recognition systems and found that the accuracy is best when the subject is a lighter skinned man, reports Timothy Revell for *New Scientist*. With facial recognition software being used by police to identify suspects, "this means inaccuracies could have consequences, such as systematically ingraining biases in police stop and searches," writes Revell.

Full story via New Scientist →

Quartz

A study co-authored by MIT graduate student Joy Buolamwini finds that facial-recognition software is less accurate when identifying darker skin tones, especially those of women, writes Josh Horwitz of *Quartz*. According to the study, these errors could cause AI services to “treat individuals differently based on factors such as skin color or gender,” explains Horwitz.

Full story via Quartz →

Gizmodo

Writing for *Gizmodo*, Sidney Fussell explains that a new Media Lab study finds facial-recognition software is most accurate when identifying men with lighter skin and least accurate for women with darker skin. The software analyzed by graduate student Joy Buolamwini “misidentified the gender of dark-skinned females 35 percent of the time,” explains Fussell.

Full story via Gizmodo →

The New York Times

Steve Lohr writes for the *New York Times* about graduate student Joy Buolamwini’s findings on the biases of artificial intelligence in facial recognition. “You can’t have ethical A.I. that’s not inclusive,” Buolamwini said. “And whoever is creating the technology is setting the standards.”

Full story via The New York Times →