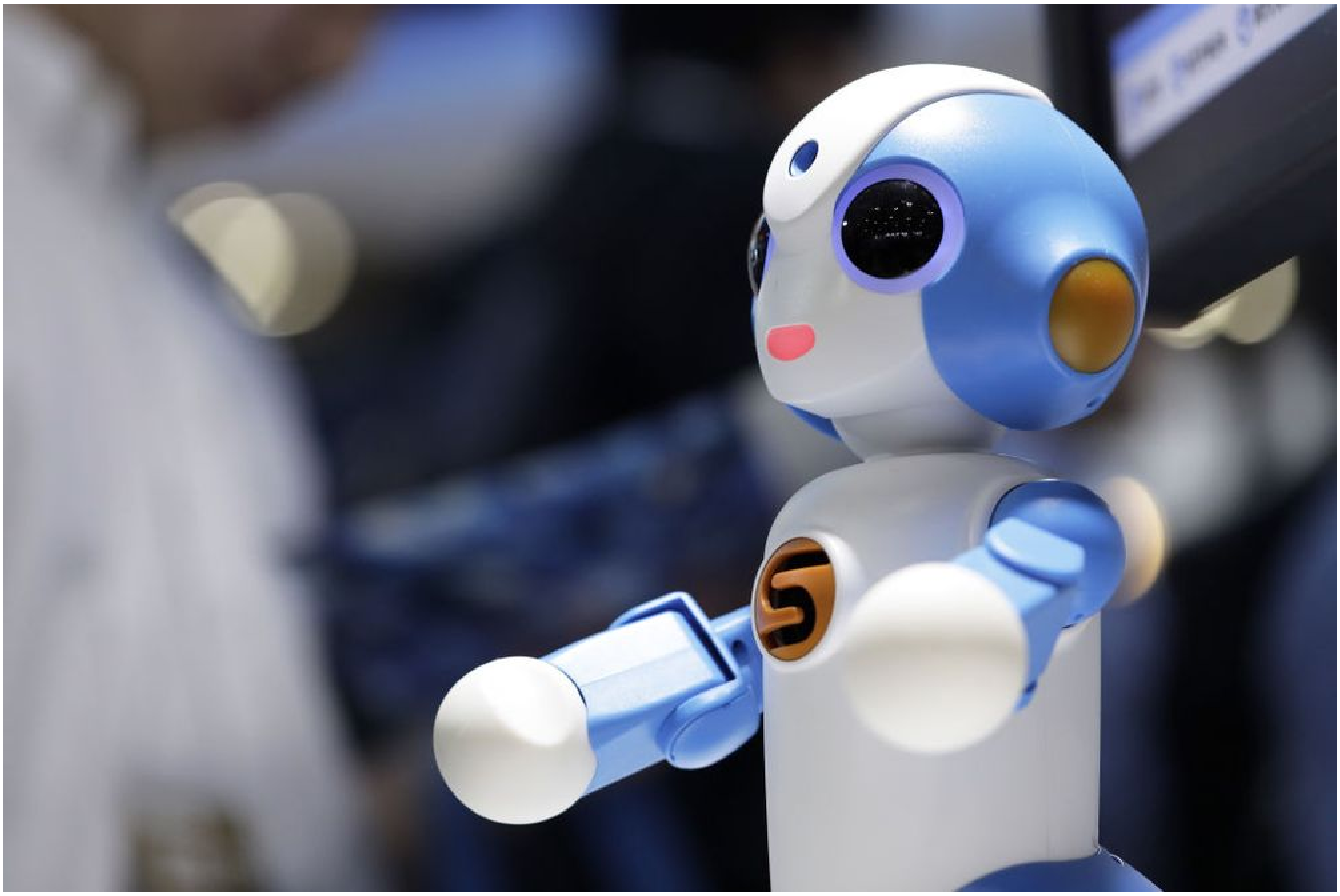
Technology & Ideas

AI Researchers Are Pushing Bias Out of Algorithms

Practical ideas for ensuring that artificial intelligence is ethical and fair are gushing from inside the tech profession.

By Ariel Procaccia

March 7, 2019, 7:30 AM EST



Not so scary. *Photographer: Photographer: Kiyoshi Ota/Bloomberg*

Artificial intelligence is getting a bad rap. People of color and women are said to be disadvantaged by the technology underlying bail recommendations, facial recognition, self-driving cars and online advertising. Some critics argue that AI can also worsen health disparities based on income and race. These examples raise the specter of a world governed by racist, sexist and plutocratic algorithms.

What gets less attention, though, is the reaction of AI specialists to these issues. And that's too bad. The perception that AI researchers and developers care more about algorithms and robots than about people is misguided.

I am reminded of an essay circulated by a sociologist at a workshop last year, in which he asserted that algorithms were "once reserved for the a-melodic, eye-contactless orations of computer scientists" who focus on their "engineering-infused definition." But more and more of the orations of computer scientists, be they dry as a bone or worthy of Cicero, are devoted to AI's impact on people and society, and the effort to build ethical AI is already in full swing.

The scale of this endeavor is significant. In particular, there are two new interdisciplinary conferences dedicated to ethical AI: the Conference on Fairness, Accountability and Transparency [1] and the Conference on AI, Ethics, and Society. In 2019 they had a combined total of over 400 research papers submitted for peer review. Such a paper is typically the culmination of months or years of work by a team of scientists, and many present concrete technical approaches to the design of ethical AI.

To address issues of bias and unfairness, researchers are tackling two related challenges.

The first challenge is to spell out what "fairness" means in mathematical terms. This may seem difficult, but there's actually an embarrassment of riches: Dozens of proposals are under consideration.

The second challenge is to put those abstract fairness definitions to practical use. Again, a wide variety of approaches have been suggested to modify existing machine learning algorithms or massage their output to even the odds.

The importance given to understanding and preventing bias and unfairness is evident in AI's mainstream conferences. For example, at the 2018 International Conference on Machine Learning, five papers were selected for awards among 2,473 submissions; two of them deal with fairness.

Tellingly, one of the award-winning papers prominently references the work of John Rawls, the theorist of social equality and fairness who has replaced John Searle, a giant of the philosophy of mind, as AI's favorite philosopher. This is representative of a much larger trend. Philosophy has long played a key role in the investigation of intelligence, and the discipline is back in vogue among AI researchers grappling with ethical questions.

Philosophy and ethics have also become indispensable components of AI education. Computer scientists are now teaching courses on ethics and AI at leading academic research centers like Carnegie Mellon, Cornell and Stanford that complement those offered by philosophers. The curriculum of Carnegie Mellon's new bachelor's degree in AI, which I helped design, includes a mandatory course in ethics.

These efforts set the stage for ethical AI. But they would ultimately prove futile (barring massive regulation) if the corporate tech giants, whose AI-driven systems interact with billions of people, had an interest in preserving the status quo. Some skeptical observers believe that this is the case; I take a more optimistic view.

In addition to the obvious benefits of avoiding bad publicity and alleviating pressure from conscientious employees, there is a more subtle reason: The experts who shape AI development and policy at major tech companies are embedded in the AI community, and many are former academics. They share the same principles, passions and aspirations that have made ethical AI one of the discipline's most vibrant areas. Some of the prominent innovators in ethical AI work for Microsoft, IBM and Google.

Still, there are danger zones.

Autonomous weapons are emerging as a major concern, and the future of work appears worrisome, too, in a world where computers can do so many tasks that humans once monopolized. It has been argued that AI may boost repressive regimes at the expense of liberal democracies. Looking further into the future, the existence of a superintelligent system that doesn't have humanity's best interests at heart – a well-loved science fiction theme – is becoming a credible possibility. [2]

The AI community is eager to address these challenges, and I believe it's equal to the task. For those who devote their lives to AI, the overriding priority – much like Asimov's laws of robotics – will always be the good of humanity.

[1] Although this conference attracted roughly 600 researchers from all over the world, there were no attendees from China, a world leader in AI research alongside the U.S. This may be a worrisome indication of research priorities there.

[2] For those who find this possibility terrifying, take comfort in knowing that there's one thing experts agree on: If terminators were built, and if they spoke English, they almost surely wouldn't have a heavy Austrian accent.

This column does not necessarily reflect the opinion of the editorial board or Bloomberg LP and its owners.

To contact the author of this story:
Ariel Procaccia at arielpro@cs.cmu.edu

To contact the editor responsible for this story:
Jonathan Landman at jlandman4@bloomberg.net

Ariel Procaccia is Gordon McKay Professor of Computer Science at Harvard University. His areas of expertise include artificial intelligence, theoretical computer science and algorithmic game theory.

Read more opinion