Stanford eCorner

Hope in Artificial Intelligence

30-04-2018

URL: https://stvp.stanford.edu/clips/hope-in-artificial-intelligence

Software engineer Tracy Chou observes how bias is easier to remove from machine-learning models than from human behavior. But she warns that bias can still creep in through the use of proxy data: for instance, excluding data on race but including closely related information like zip codes. An early employee of several Silicon Valley startups, Chou was a founding advisor to Project Include and continues to push for more diversity in tech.

Transcript

we're gonna come out on that.. One thing that is also promising about AI, machine learning, is that we can also know how to use it to solve bias.. So, there are companies like Text Geo which helps you to look at your job postings and see if they're bias towards male or female candidates, and help you to remove those and suggest alternative words.. There's some interesting research I was reading from a couple years ago is, some models over text.. They did that where having gendered relationships between things like, doctor and men, and nurse and women, when they could identify the gender bias, they could also go into the models and actively de-bias them and remove that bias.. Which is pretty powerful if you think about it.. With humans, even if you identify bias, you can't really get rid of it very easily.. People will acknowledge that they have biases and they need to work on it, but you can't just go up and like zero out your vector, and make sure that all future decisions are unbiased.. But, you can do that with AI models potentially, depending on which ones you're building, if yo know that you want to intentionally remove the biases here.. There's other research happening, and how do you decorrelate different things? Say you're building models, financial models and you're not supposed to use race, this is actually regulated in some industries..

You're forbidden from actually using race as an input to your models, but there are proxy variables, so looking at zip code, for example, you could end up basically getting the same biases embedded.. There are people working on how do we do post processing around the output of these models to actively decorrelate.. So, there's a lot of really interesting work happening.. I'm pretty excited that the technology industry is just changing so much...