Special Topic: Fairness in Machine Learning

Chao Lan
Machine learning is increasingly applied in sensitive areas.
Fairness matters in recidivism prediction.

Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.

- ProPublica, 2016.
Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer et al. find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

Science, this issue p. 447; see also p. 421
Schalkwijk is one of a fast-growing cohort of human resources executives relying on artificial intelligence to recruit, assess, hire, and manage their staff. In a 2018 Deloitte survey, 82% of business and technology executives said they were deploying A.I. for “workforce management.” That share is almost certainly higher today—and it’s spreading to encompass some of the world’s largest companies.

In a LinkedIn survey of hiring managers and recruiters who use A.I., 67% said they embraced the tech because it helped them save time. And a smaller cohort, 43%, cited an arguably more important motivation: A.I., they said, would help them combat bias in their decision-making. “People are inherently biased,” says Schalkwijk. “I wanted less biased hiring decisions and more data-driven hiring decisions.”

Amazon scraps secret AI recruiting tool that showed bias against women.

Other bias: commercial facial recognition.

Other bias: GoogleVoice.
Other bias: drug use prediction.

Other bias: word embedding techniques.
Algorithmic fairness is a priority in the US AI R&D strategic plan (2016).

“To avoid exacerbating biases by encoding them into technology systems, we need to develop a principle of ‘equal opportunity by design’ -- designing data systems that promote fairness and safeguard against discrimination from the first step of the engineering process and continuing throughout their lifespan.”
And it remains a priority in the 2019 update.

“Beyond purely data-related issues, however, larger questions arise about the design of AI to be inherently just, fair, transparent, and accountable.

…

Scientists must also study to what extent justice and fairness considerations can be designed into the system, and how to accomplish this within the bounds of current engineering techniques.”
Many initiatives on fairness research.

many workshops, papers, articles...
What is “fairness” in algorithmic prediction?

**Statistical Disparity** means big gap between
- $\Pr\{ f(x) = \text{hire} \mid x \text{ is male} \}$
- $\Pr\{ f(x) = \text{hire} \mid x \text{ is female} \}$

**Equalized error rates** means similar
- $\Pr\{ f(x) = \text{hire} \mid x \text{ is male} \& y = \text{not hired} \}$
- $\Pr\{ f(x) = \text{hire} \mid x \text{ is female} \& y = \text{not hired} \}$

**Individual Fairness** means
- $f(x) = f(z)$ if $x$ and $z$ are equally qualified

Amazon scraps secret AI recruiting tool that showed bias against women.

Where does algorithmic bias come from?

Women were 25% of the group and spoke 13% of the time.
Does hiding sensitive attribute help?
Let’s design a fair learner for linear model.