Tortoise· NOTES

Governments and AI
Will AI be fair?

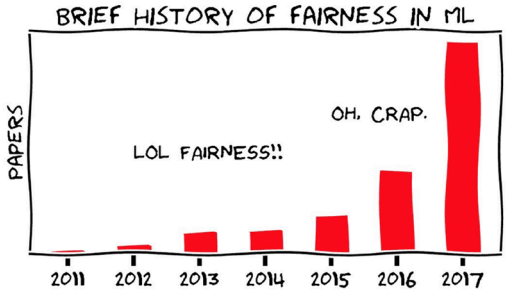
Headline partner

HENKEL^x VENTURES

A Tortoise ThinkIn • February 2020
Notes by Luke Gbedemah

Artificial intelligence and machine learning promise a more automated, interconnected and smarter future for all. Yet the algorithms that underlie these new technologies are prone to bias, and pose a significant risk. One that could undermine their entire purpose.

The number of publications on fairness from 2011 to 2017



Source: Towards data science

Bias

noun

Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.

Issue of bias

Skewed bias

Some initial bias in a set of data can intensify, rather than abate over time. Future observations confirm the premise of the predictive model more than they contradict it.

Example: A model identifies a local area as at a higher risk of crime

- Local police dispatch more officers to that area
- The increased presence of police turns out more arrests and stops in that area
- The model predicts more crime and police attention to that area intensifies
- The area becomes 'criminalised'.

Therefore: the predictive model's system is trained to have a positive bias to regions where there are fewer police, not where there is less crime.

Native bias

Any system using machine learning is prone to maintain existing bias. This means replicating existing bias rather than establishing a new, or fair, way of making the decision.

Example: A model identifies that throughout a company, managers have hired a disproportionate number of males into leadership roles

- The model labels the 'successful' applicants based on this hiring practice
- The system is trained to hire more males in replicating and automating manager decisions
- The company's model does not select applicants on the basis of capability.

Therefore: The screening model's systems is trained to extrapolate from the manager's bias at a much larger scale and with similar discrimination.

"Another example is that word embeddings trained on Google News articles "exhibit female/male gender stereotypes to a disturbing extent" e.g. the relationship between "man" and "computer programmers" was found to be highly similar to that between "woman" and "homemaker" *Bolukbasi et al. 2016*

Limited features

Data may be less reliable, and information when collected from minority groups, highlighting only a limited number of features. This can effect the reliability of labelling, and make a system tend towards lower accuracy when making predictions about minority groups.

Example: A model identifies that a person from a minority religious community should not receive insurance for their car

- The data collected about this individual is limited, and their religious identity is a core component of this set
- The model is applied to all applicants, including counterparts from a majority group
- The system tends towards a much lower accuracy for minority groups and treats religious identity as disproportionately impactful.

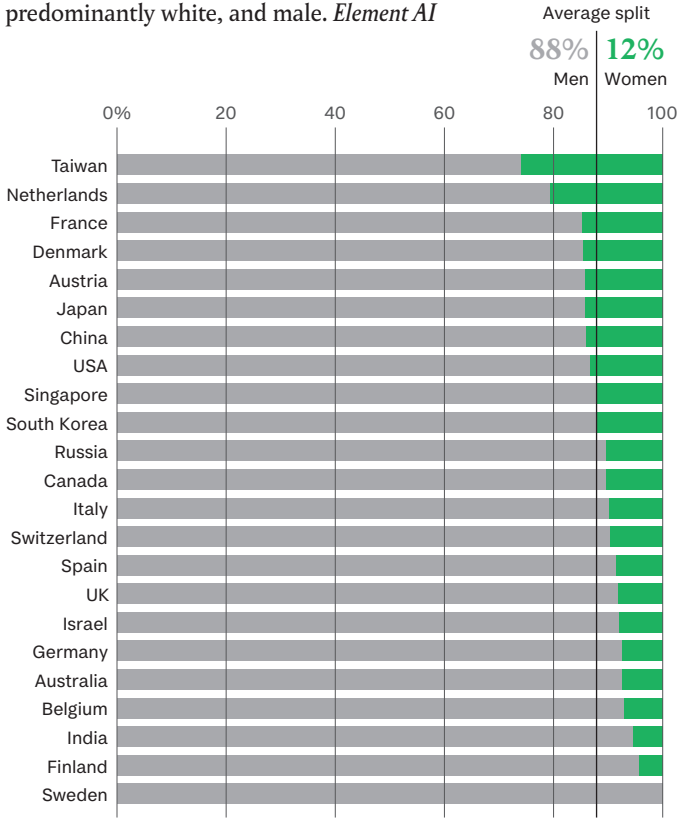
Therefore: The model's system is trained to make decisions based on much less informative or reliable data about minorities.

Towards Data Science

In a survey of 5,250 people age 13 and over, Samsung sought to gain an insight into the country's views of AI. 39% of the public voiced concerns about AI bias, rising to 43% among those who described themselves as interested in AI. *Verdict*

Case studies in AI bias

There is growing attention around the issue of diversity in the global technology industry; which remains predominantly white, and male. *Element AI*



Among 4,000 researchers who have been published at the leading conferences NIPS, ICML, or ICLR in 2017. Source: Element AI

Healthy Data Needed

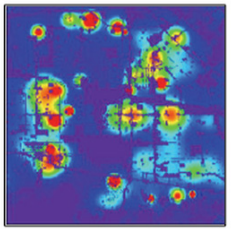
In 2011, only 4 per cent of all genome-wide association studies had been conducted using samples of non-European descent. The consequences of this bias are a problem.

Genetics research's diversity problem has led to algorithms that told African Americans they were more likely to carry a mutation putting them at risk of hypertrophic cardiomyopathy, when in fact they did not. *23andMe Blog*

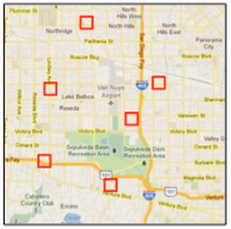
"The lack of diversity in genetics research is reflective of the general lack of diversity in all aspects of health research"

Your COMPAS is off

"The canonical example of biased, untrustworthy AI is the COMPAS system, used in Florida and other states in the US. The COMPAS system used a regression model to predict whether or not a perpetrator was likely to recidivate. Though optimized for overall accuracy, the model predicted double the number of false positives for recidivism for African American ethnicities than for Caucasian ethnicities."



Tactical ambiguity Rear-view mirror heat map



Tactical clarity Forward looking PredPol boxes

PredPol Predictive Policing Blog

The PrePol model points tactical units at specific street corners, household and bulevards.

"My point is that police make choices about where they direct their attention. Today they focus almost exclusively on the poor. That's their heritage, and their mission, as they understand it."

Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy

Why the wrong face?

Facial recognition software is being deployed more often and in more places. Policing through algorithms is getting more common.

It is a potential source of both racial and gender bias.

"In February this year, Joy Buolamwini at the Massachusetts Institute of Technology found that three of the latest gender-recognition AIs, from IBM Microsoft and Chinese company Megvii, could correctly identify a person's gender from a photograph 99 per cent of the time – but only for white men. For dark-skinned women, accuracy dropped to just 35 per cent."

The risk of false identification amongst women and ethnic minorities is high.

IBM quickly announced that it had retrained its system on a new data set, and Microsoft said it has taken steps to improve accuracy. *New Scientist*

Notes

Reading time: 4-minute read
Published: February 2020