

A practical view on bias & fairness

—
Damiaan Zwietering
zwietering@nl.ibm.com

We created an AI to...

We applied an algorithm to historical data that generated a model to predict...



Statistics: summarize

Data mining: repeated, automated

Machine learning: feedback

Deep learning: discover features

Artificial intelligence: reasoning

What does it take to trust a decision made by a machine?

Apart from accuracy



Is it fair?



Is it easy to understand?



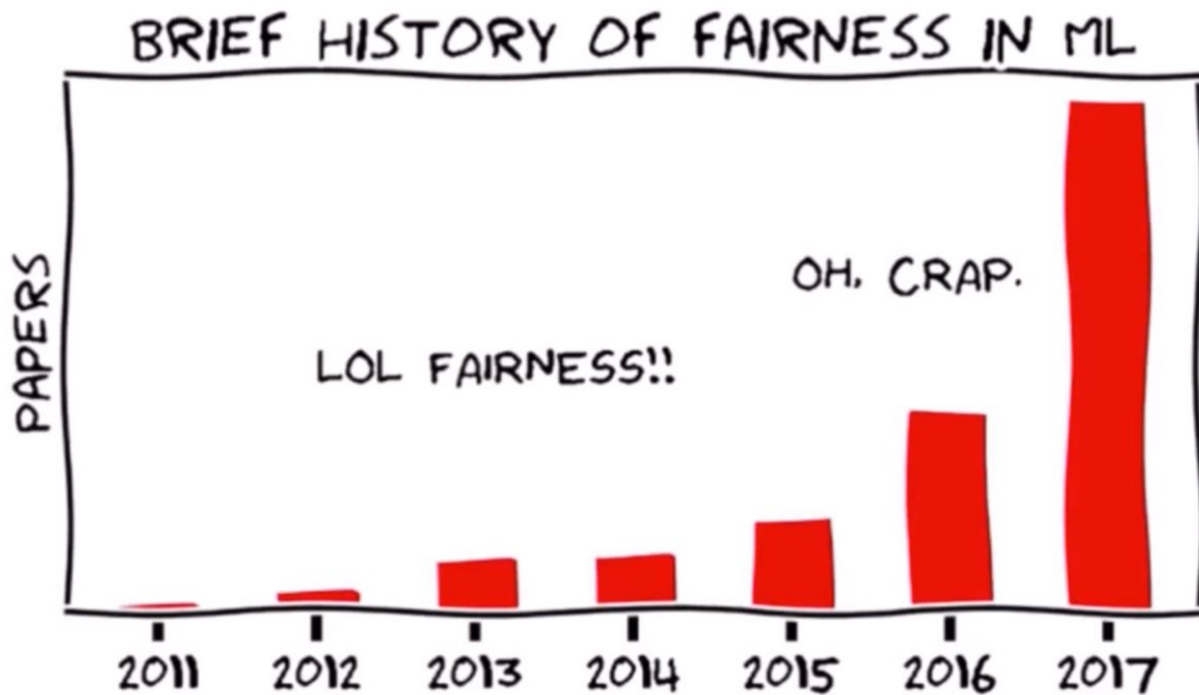
Did anyone tamper with it?



Is it accountable?

Research

Algorithmic fairness is gaining a lot of attention



(Hardt, 2017)

Do we care?

Of course we do!

TECH | AMAZON | ARTIFICIAL INTELLIGENCE

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By James Vincent | @jvincent | Oct 10, 2018, 7:09am EDT

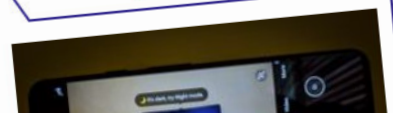
f t SHARE



21



MOST READ



Unwanted bias and algorithmic fairness

Machine learning, by its very nature, is always a form of statistical discrimination

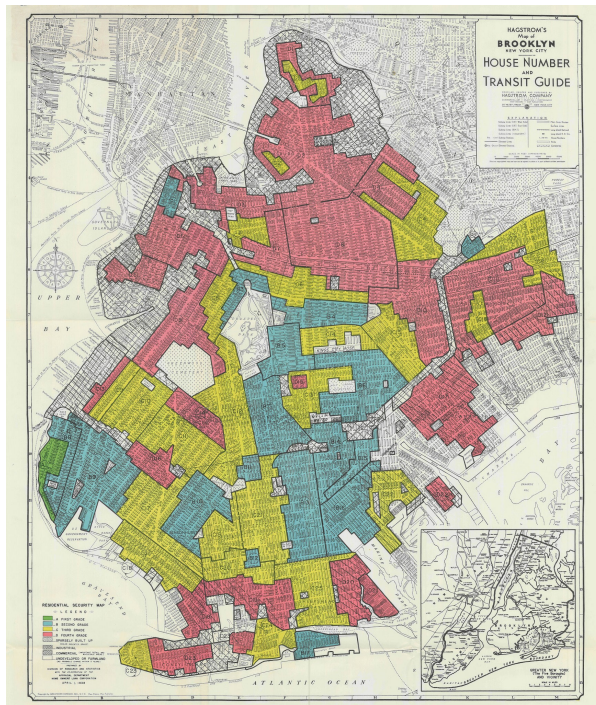


Areas of concern

- Accountability
- Value Alignment
- Explainability
- Fairness
- User Data Rights

Bias mitigation is hard

We cannot simply drop protected attributes



Proxies, correlation

Monitoring impossible

Explainability lost

“Fairness does not mean everyone gets the same. Fairness means everyone gets what they need.”

Rick Riordan

The Red Pyramid, 2010

AI Fairness 360 - Demo



4. Compare original vs. mitigated results

Dataset: Adult census income

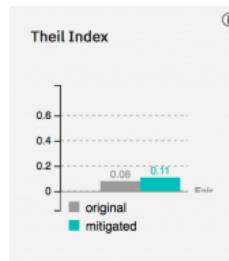
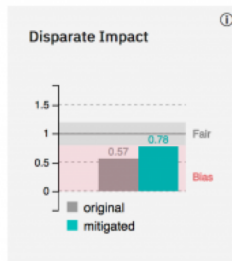
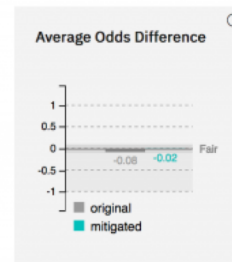
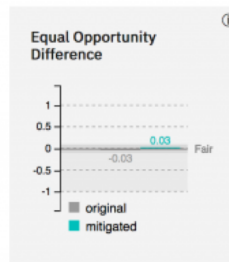
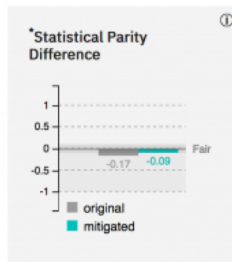
Mitigation: **Optimized Pre-processing algorithm applied**

Protected Attribute: Race

Privileged Group: **White**, Unprivileged Group: **Non-white**

Accuracy after mitigation changed from 82% to 74%

Bias against unprivileged group was reduced to acceptable levels* for 1 of 2 previously biased metrics (1 of 5 metrics still indicate bias for unprivileged group)

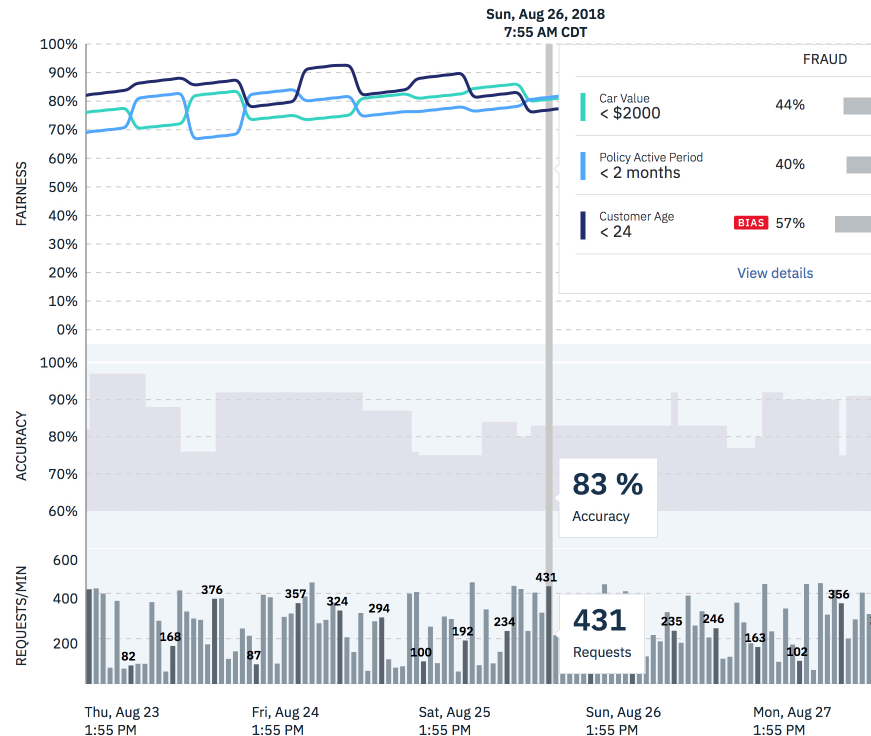


"As I grow older, I pay less attention to what men say. I just watch what they do."

Andrew Carnegie
1835-1919

Fraud Detection

Description	Suggests if a claim is fraudulent.	Date Created	September 1, 2017
Model Owner	Dinesh Kapadila	Date Retrained	May 5, 2018
Business Owner	Camilla Señor	Last Evaluated	1 hour ago



AIF360: <https://aif360.mybluemix.net/>

OpenScale: <https://www.ibm.com/cloud/watson-openscale/>

