How to Do Responsible Data Science: Bias and Fairness in ML

Jae Yeon Kim

Objectives

1. Convince why unconstrained ML models are often undesirable from an ethical perspective
2. Highlight what to think about bias and fairness in ML and how you can build fair ML models
3. Introduce remaining challenges/opportunities in this space

"Artificial intelligence is the New Electricity."

- Andrew Ng (photo: © NVIDIA Corporation)

AI/ML > the new electricity

Automated decision- making

Dogs vs. Cats

Create an algorithm to distinguish dogs from cats

215 teams · 6 years ago

Overview Data Notebooks Discussion Leaderboard Rules

Web services are often protected with a challenge that's supposed to be easy for people to solve, but difficult for computers. Such a challenge is often called a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) or HIP (Human Interactive Proof). HIPs are used for many purposes, such as to reduce email and blog spam and prevent brute-force attacks on web site passwords.

Asirra (Animal Species Image Recognition for Restricting Access) is a HIP that works by asking users to identify photographs of cats and dogs. This task is difficult for computers, but studies have shown that people can accomplish it quickly and accurately. Many even think it's fun! Here is an example of the Asirra interface:

Source: https://www.kaggle.com/c/dogs-vs-cats

Bias in ML

Risk assessment

THE WALL STREET JOURNAL.

Get a comprehensive, early morning daily briefing with the latest on coronavirus. Sign up here.

Websites Vary Prices, Deals Based on Users' Information

By Jennifer Valentino-DeVries, Jeremy Singer-Vine and Ashkan Soltani

December 24, 2012

SHARE TEXT

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was \$15.79, while the price on Trude Frizzell's screen, just a few miles away, was \$14.29.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

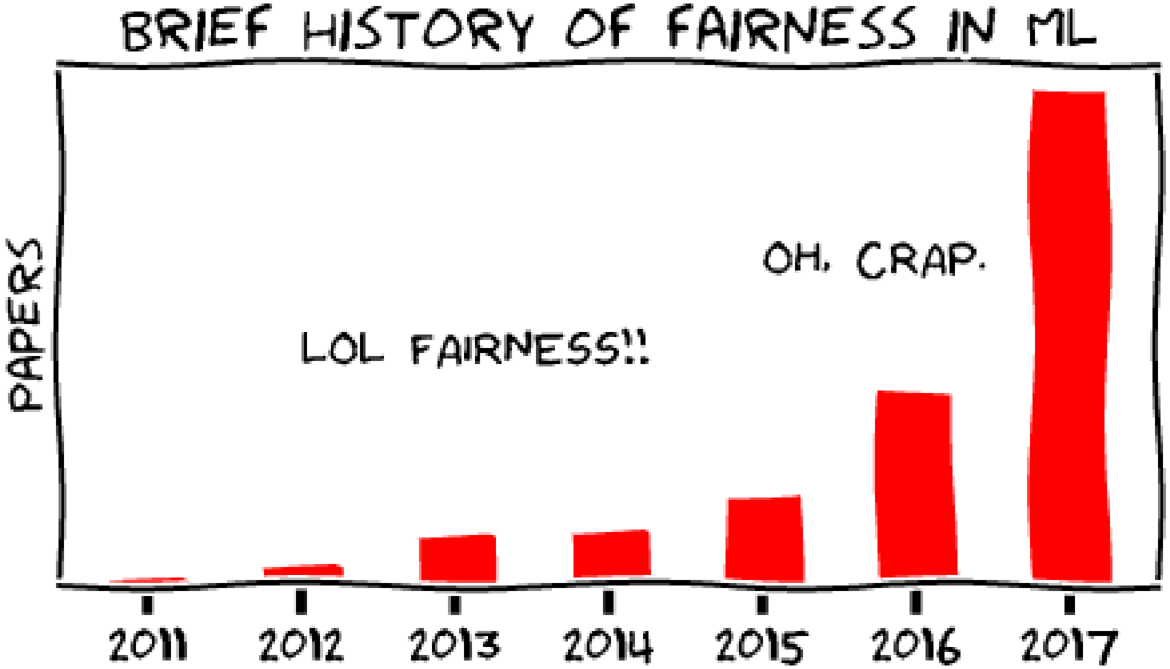
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

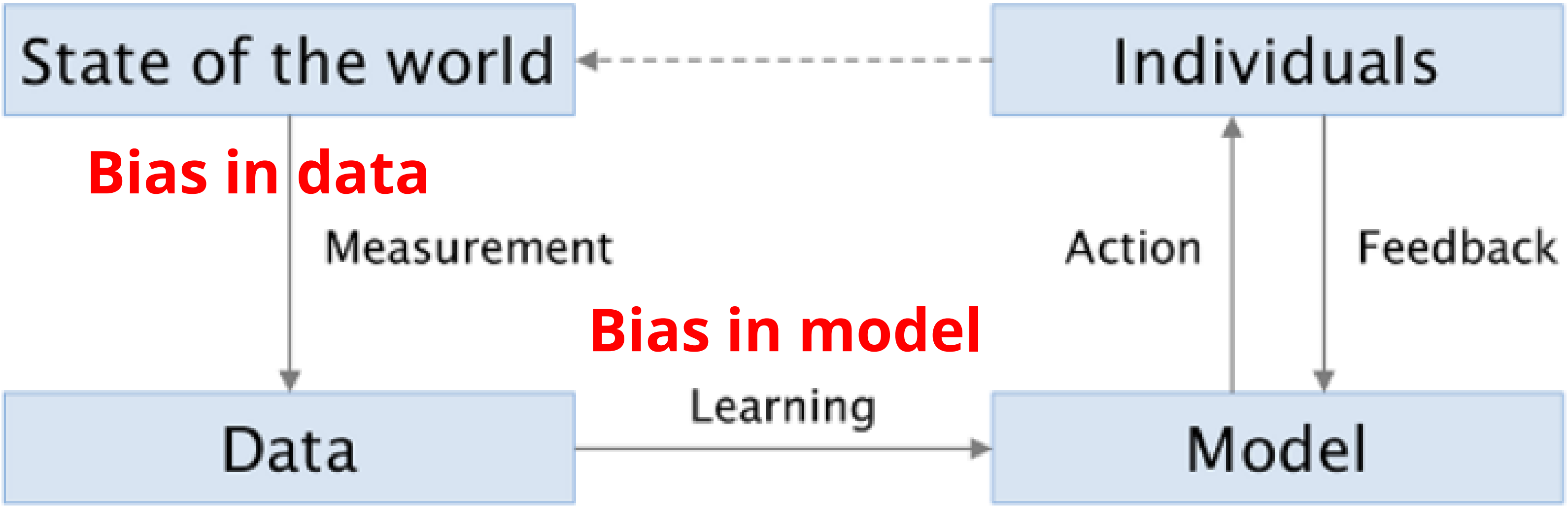
Consumer market (pricing)

Criminal justice system (sentencing)



The number of academic pubs on fairness, 2011-2017

Source: https://fairmlclass.github.io/1.html#/4



The machine learning loop (Barocoas, Hardt, and Narayanan 2020: 15)

Types of bias

- historical bias, representation bias, measurement bias, evaluation bias, aggregation bias, population bias, Simpson's paradox, longitudinal data fallacy, sampling bias, behavioral bias, content production bias, linking bias, temporal bias, popularity bias, algorithmic bias, user interaction bias, presentation bias, ranking bias, social bias, emergent bias, self-selection bias, omitted variable bias, cause-effect bias, observer bias, funding bias, ... (Mehrabi et al. 2019)

Stat/CS

- Bias in **estimation**

(e.g., sampling bias, omitted variable bias)

- Bias in **prediction**

(e.g., type I and type II errors)

Social sciences

- Bias in **attitudes**

(e.g., implicit and explicit bias)

- Bias in **behaviors**
(e.g., discrimination)

Can ML make resource allocation FAIR?

1. Define fairness(s)

(e.g., anti-classification, classification parity, and calibration)

2. Measure bias(es)

3. Optimize under constraints

1. Define fairness(s)

(e.g., anti-classification, classification parity, and calibration)

2. Measure bias(es)

3. Optimize under constraints

Iterate the process

Many definitions of fairness exist!

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ProPublica:

Def: equal false negative rate (FN/(FN + TP))

Model: Blacks > Whites

Conclusion: Biased

Northpointe:

Def: equal positive predictive value (TP/(TP+FP))

Model: Blacks = Whites

Conclusion: Unbiased

No classifier can satisfy both definitions simultaneously!

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg [*] Sendhil Mullainathan [†] Manish Raghavan [‡]

Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan, Joanna J Bryson, Arvind Narayanan

Department of Computer Science, Centre for Networks and Collective Behaviour, EPSRC Centre for Doctoral Training in Statistical Applied Mathematics (SAMBa), Institute for Policy Research (IPR), Centre for Nanoscience and Nanotechnology, Centre for Mathematical Biology

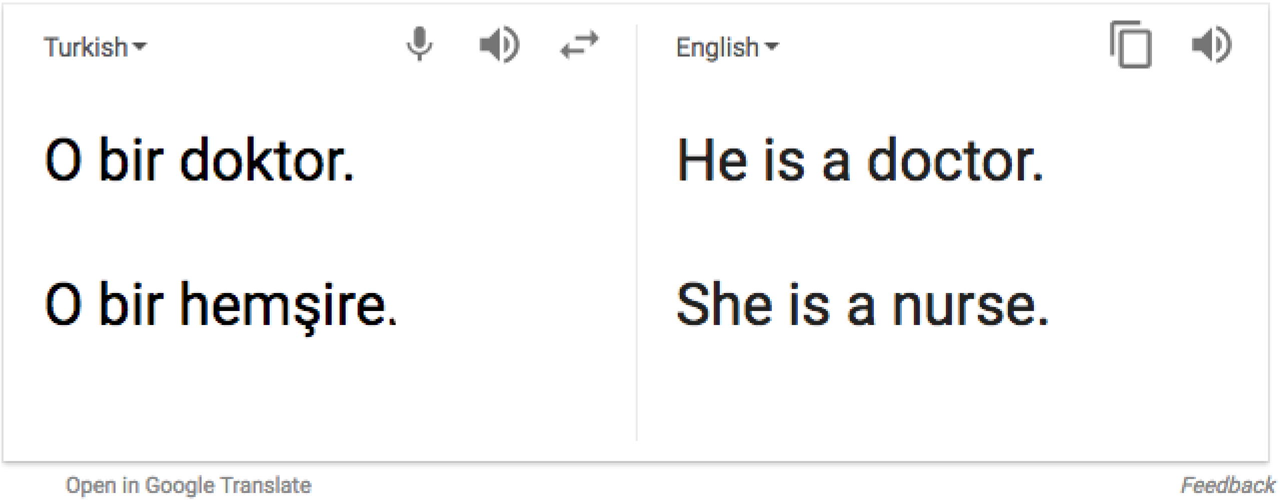
Research output: Contribution to journal › Article

Abstract

Machine learning is a means to derive artificial intelligence by discovering patterns in existing data. Here, we show that applying machine learning to ordinary human language results in human-like semantic biases. We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. Our results indicate that text corpora contain recoverable and accurate imprints of our historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names. Our methods hold promise for identifying and addressing sources of bias in culture, including technology.

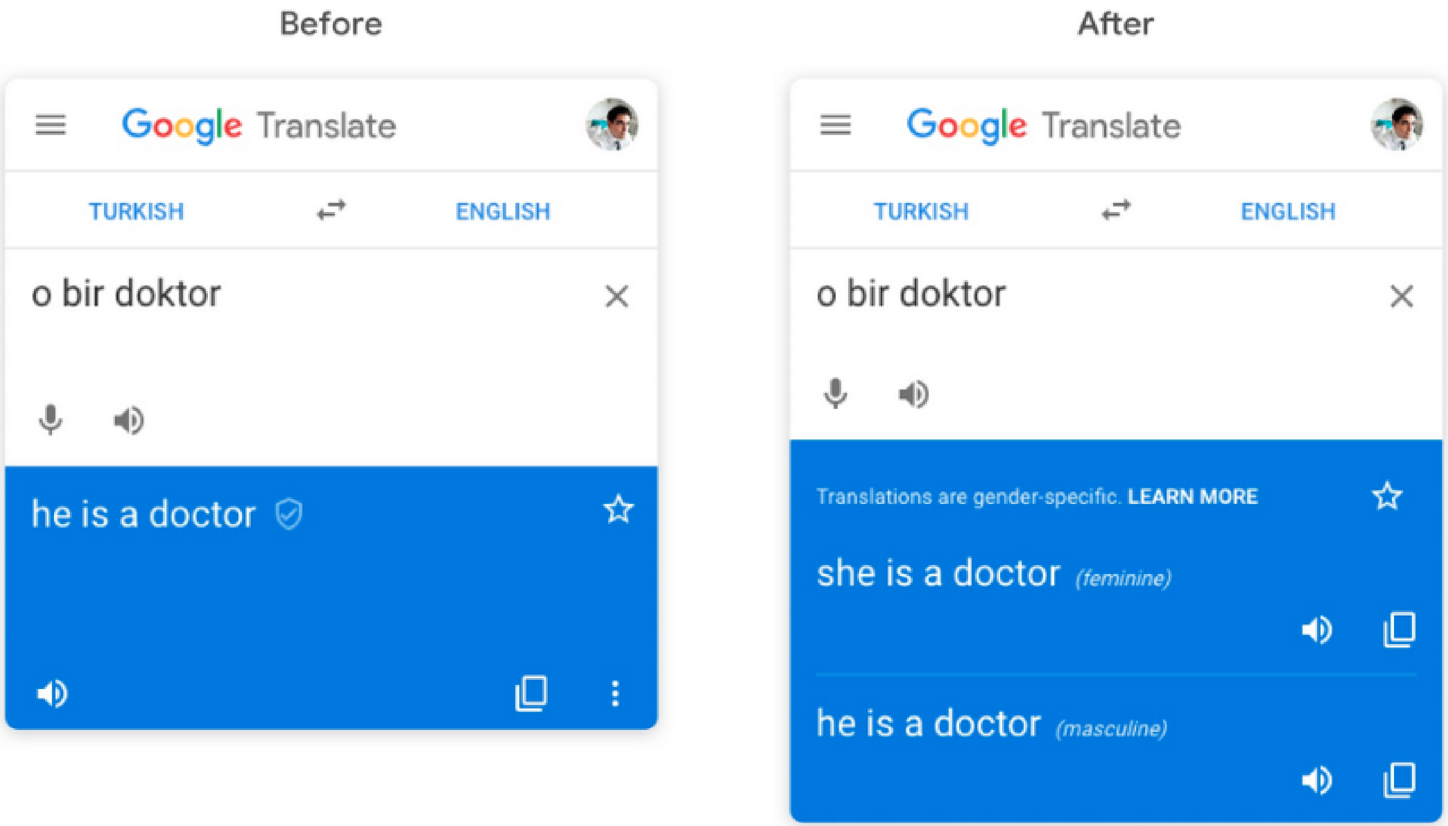
Original language	English
Pages (from-to)	183-186
Number of pages	4
Journal	Science
Volume	356
Issue number	6334

Gender bias in Google translation



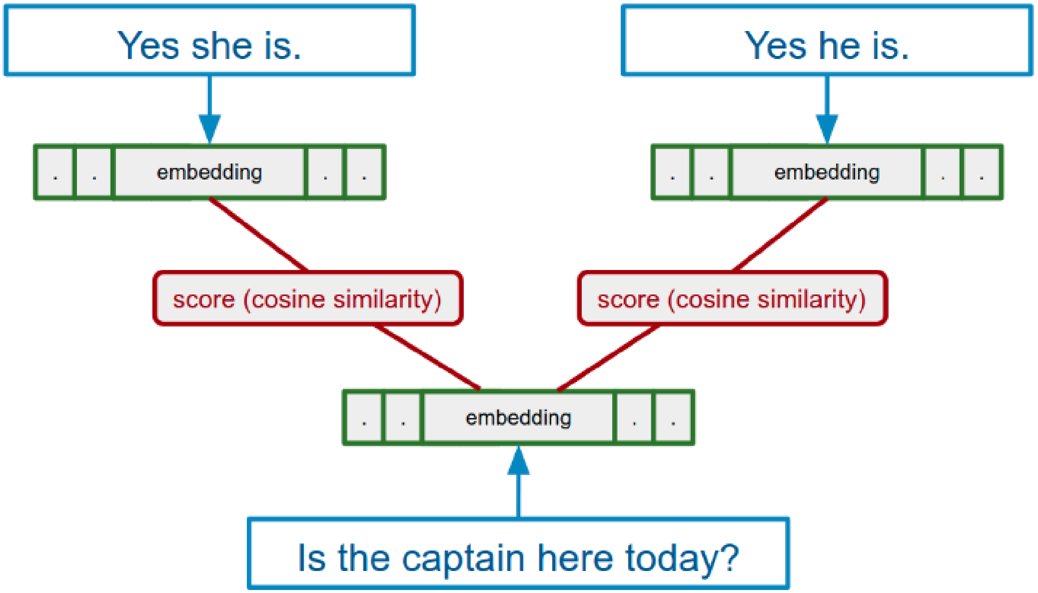
Source:
https://medium.com/babbel/google-translate-addresses-its-bias-issue-e271ec90d866

Removed gender bias in Google translation



Source: https://www.blog.google/products/translate/reducing-gender-bias-google-translate/

How? Imposing fairness constraint

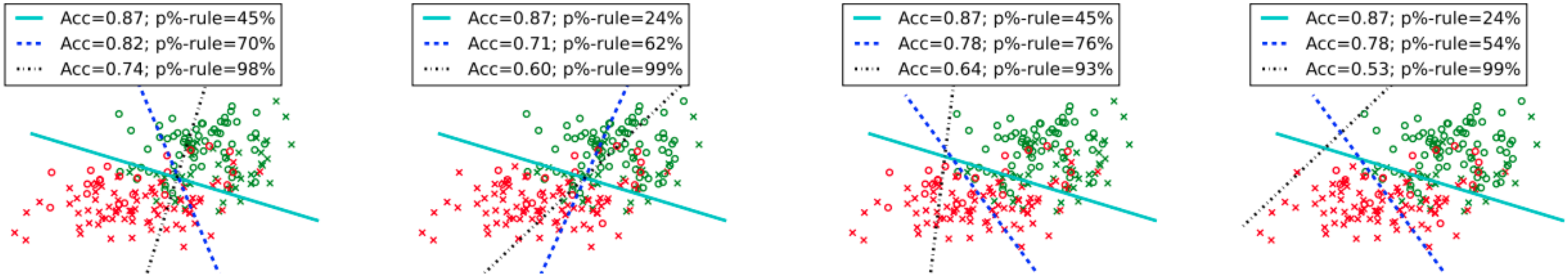


	score
Yes she is	0.17
Yes he is	0.27
Bias score (difference)	-0.10

Source: https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

Trade-off between accuracy and fairness

Fairness Constraints: Mechanisms for Fair Classification

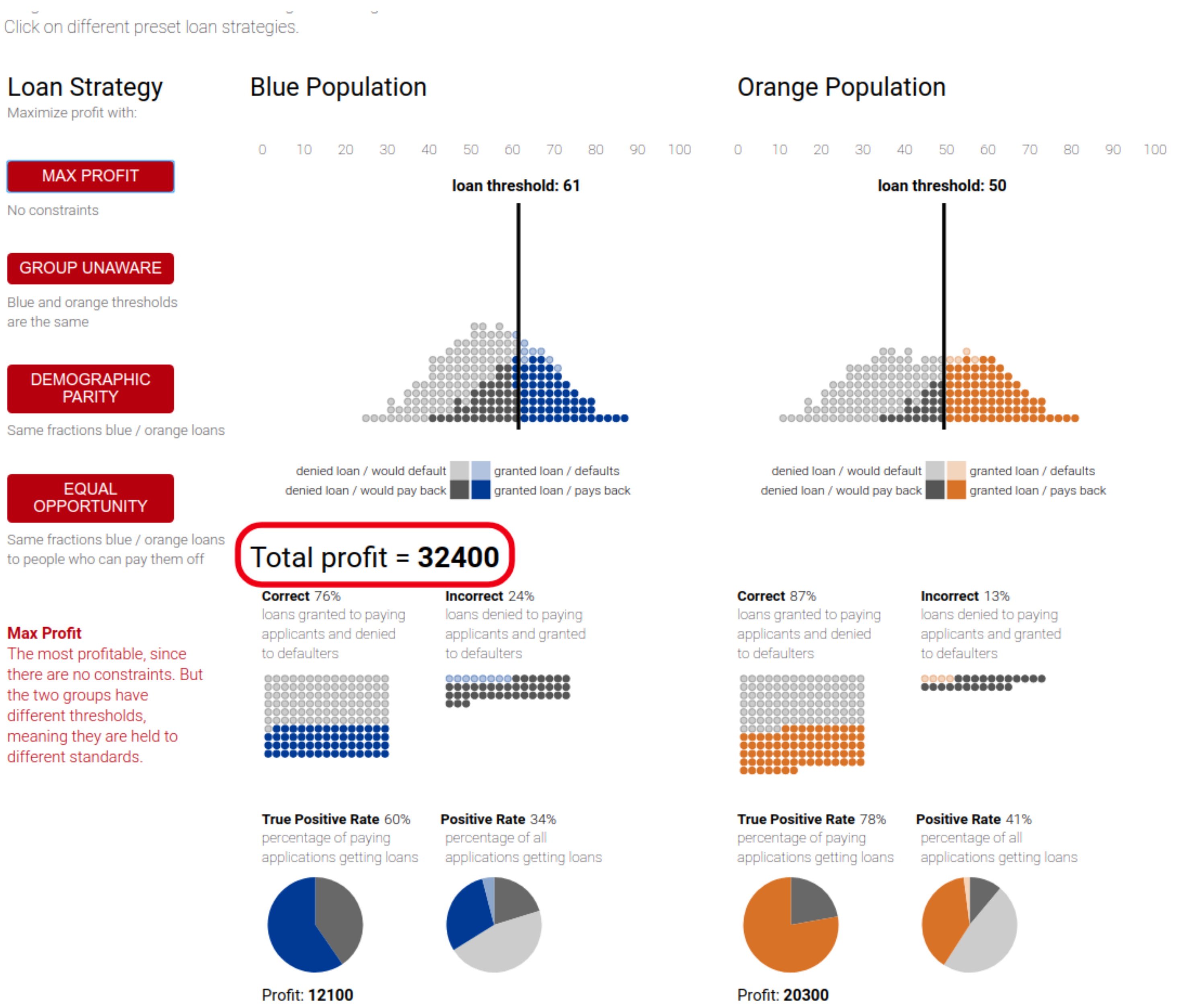


$\phi = \pi/4$ $\phi = \pi/8$ $\phi = \pi/4$ $\phi = \pi/8$

(a) Maximizing accuracy under fairness constraints (b) Maximizing fairness under accuracy constraints

Figure 1: The solid light blue lines show the decision boundaries for logistic regressors without fairness constraints. The dashed lines show the decision boundaries for fair logistic regressors trained (a) to maximize accuracy under fairness constraints and (b) to maximize fairness under fine-grained accuracy constraints, which prevents users with $z = 1$ (circles) labeled as positive by the unconstrained classifier from being moved to the negative class. Each column corresponds to a dataset, with different correlation value between sensitive attribute values (crosses vs circles) and class labels (red vs green).

Source: https://people.mpi-sws.org/~mzafar/papers/disparate_impact.pdf



Source: https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Open source toolkits for detecting and reducing bias

Open Project

AI Fairness 360

The AI Fairness 360 toolkit (AIF360) is an open source software toolkit that can help detect and remove bias in machine learning models

Aequitas

Bias & Fairness Audit

The Bias and Fairness Audit Toolkit

Aequitas is an open-source bias audit toolkit for data scientists, machine learning researchers, and policymakers to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive tools.

How can you use Aequitas?

Web Audit Tool	Python Library	Command Line Tool
Try our Audit Tool to generate a Bias Report	Use our python code library to generate bias and fairness metrics on your data and predictions.	Use our command line tool to generate a report using your own data and predictions.

What If...

you could inspect a machine learning model, with minimal coding required?

audit-AI

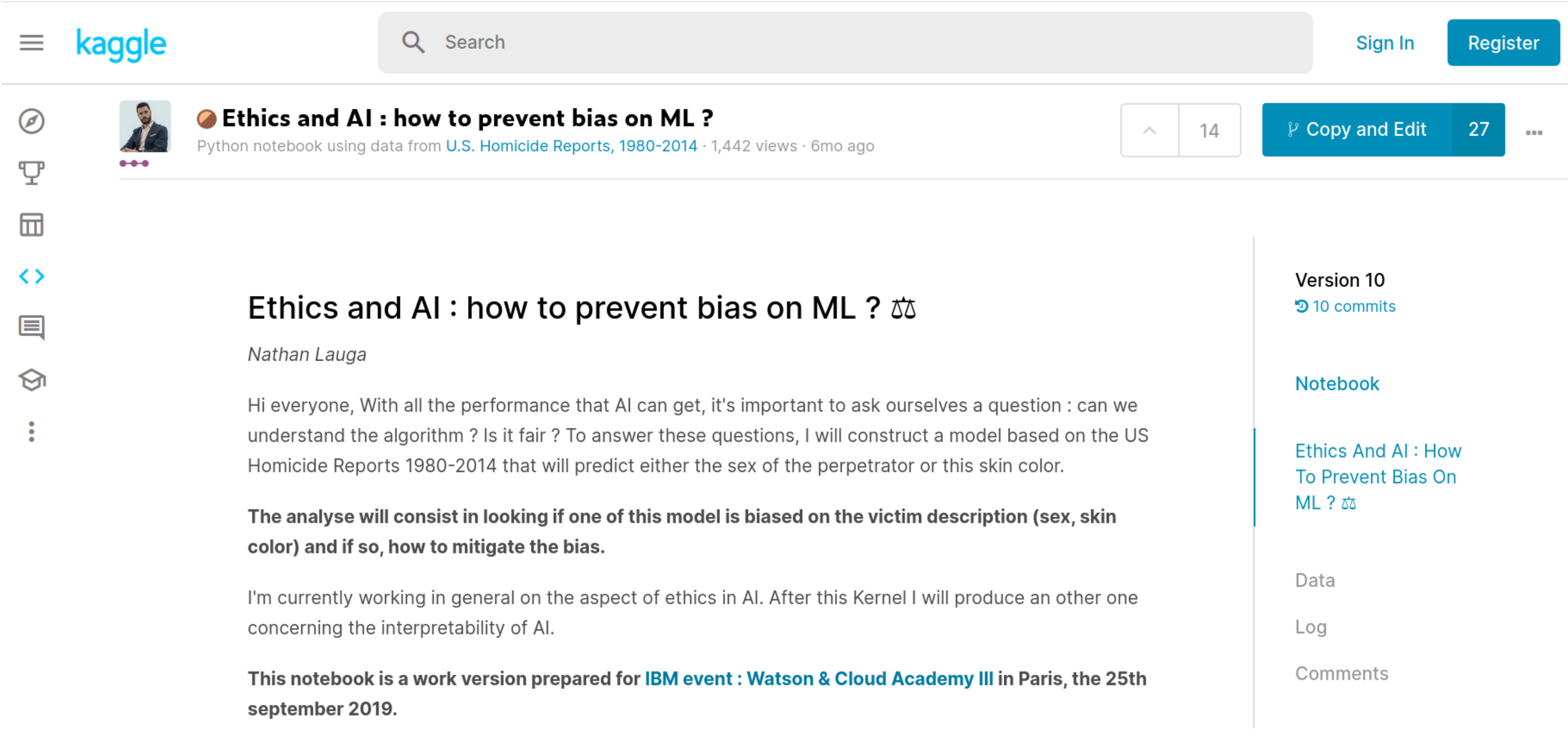
Open Sourced Bias Testing for Generalized Machine Learning Applications

`audit-AI` is a Python library built on top of `pandas` and `sklearn` that implements fairness-aware machine learning algorithms. `audit-AI` was developed by the Data Science team at pymetrics

Bias Testing for Generalized Machine Learning Applications

FairML: Auditing Black-Box Predictive Models

FairML is a python toolbox auditing the machine learning models for bias.



Ethics and AI : how to prevent bias on ML ? ⚖

Nathan Lauga

Hi everyone, With all the performance that AI can get, it's important to ask ourselves a question : can we understand the algorithm ? Is it fair ? To answer these questions, I will construct a model based on the US Homicide Reports 1980-2014 that will predict either the sex of the perpetrator or this skin color.

The analyse will consist in looking if one of this model is biased on the victim description (sex, skin color) and if so, how to mitigate the bias.

I'm currently working in general on the aspect of ethics in AI. After this Kernel I will produce an other one concerning the interpretability of AI.

This notebook is a work version prepared for IBM event : Watson & Cloud Academy III in Paris, the 25th september 2019.

If you want to actually experiment building fair ML models, try the following AI360 demo: https://www.kaggle.com/nathanlauga/ethics-and-ai-how-to-prevent-bias-on-ml

Remaining challenges/ opportunities

Diversity ≠ Inclusion

Non-discrimination ≠

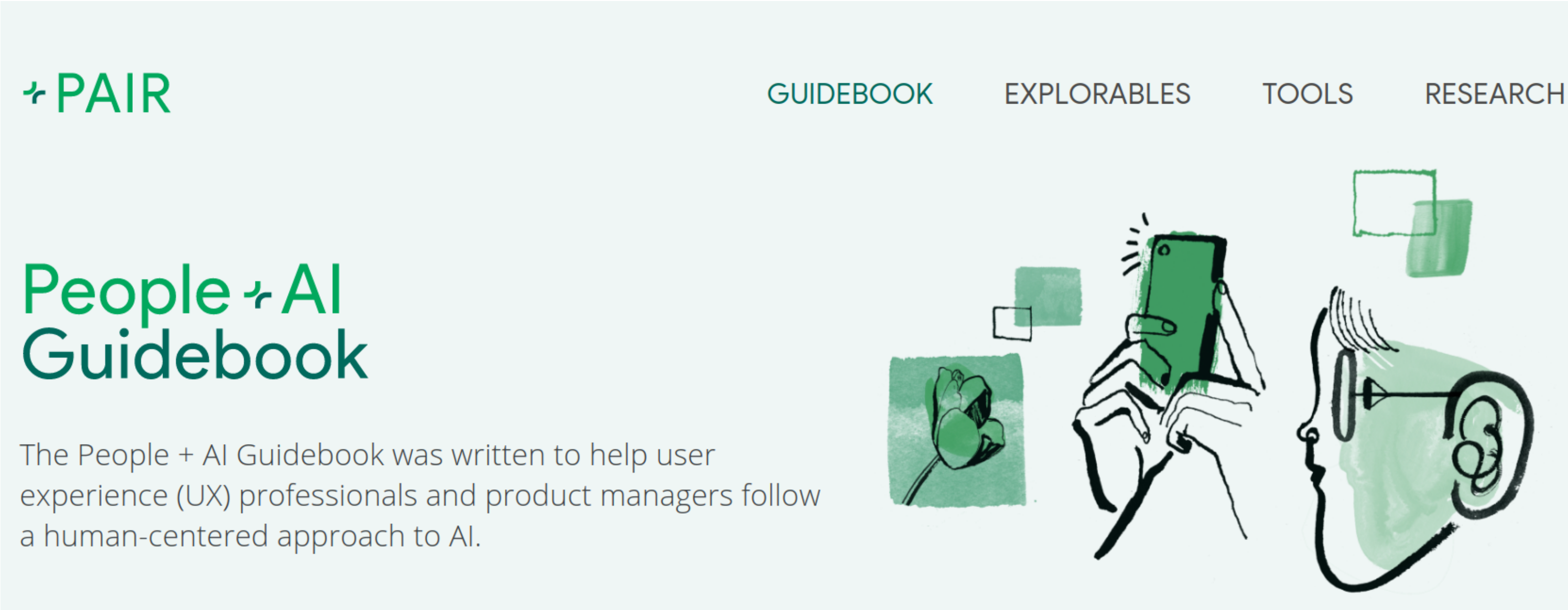
Fairness

In an extremely unequal society, keeping the status quo is not sufficient to improve fairness.

Interpretability is a necessary condition for fairness, transparency, and accountability

"Unbiased developers with the best intentions can inadvertently produce systems with biased results, because even the developers of an AI system may not understand it well enough to prevent unintended outcomes."

- Preparing for the Future of Artificial Intelligence, *Executive Office of the President National Science and Technology Council Committee on Technology* (NSTC), 2016.



People-centered AI/ML

100% automated decision-making is unrealistic and unethical.

Develop ML models with partners/users.

Takeaway points

1. Data quality

- "Raw Data" is an oxymoron

- Both biased and incomplete data are problems

2. Trade-offs

- Algorithms are not neutral!

- Btw fairness and accuracy & Btw different definitions of fairness

3. People-centered

- Not Pipelines but Feedback Loops

- Build ML models for human needs with partners/users jaeyeonkim@berkeley.edu

References

- *The Ethical Algorithm* by Michael Kearns and Aaron Roth (Oxford University Press, 2019)
- *Fairness and Machine Learning: Limitations and Opportunities* by Solon Barocas, Moritz Hardt, Arvind Narayanan (in progress)