Fairness in Machine Learning

Virginie Do
Université Paris Dauphine - PSL

May 12th, 2021

Outline

1. **Biases in AI systems**
2. Fairness in machine learning: binary decisions
3. Beyond fair learning

Biases in AI systems

Figure 2: An image search result page for the query "CEO" showing a disproportionate number of male CEOs.

Gender Classifier	Darker Subjects Accuracy	Lighter Subjects Accuracy	Error Rate Diff.
Microsoft	87.1%	99.3%	12.2%
FACE++	83.5%	95.3%	11.8%
IBM	77.6%	96.8%	19.2%

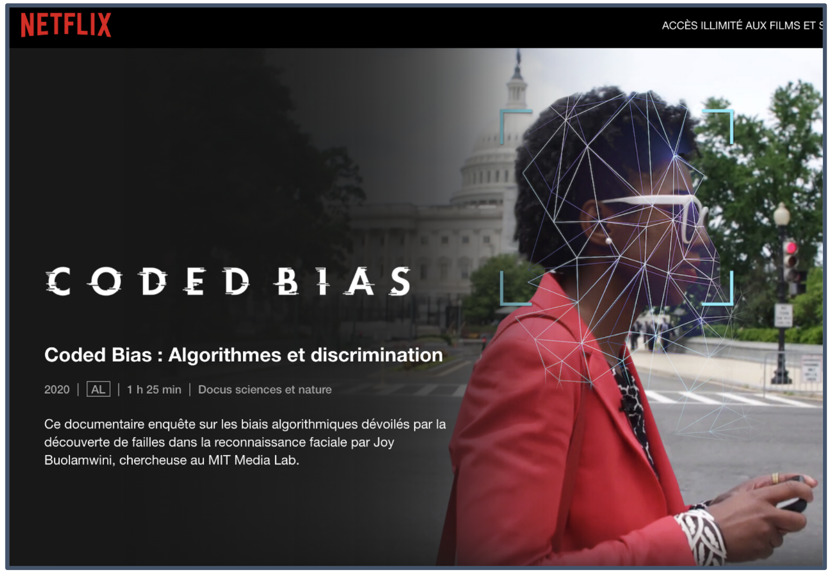
RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

HUNGARIAN - DETECTED | POLISH | PO | ENGLISH | POLISH | PORTUGUESE

Ő szép. Ő okos. Ő olvas. Ő mosogat. Ő épít. Ő varr. Ő tanít. Ő főz. Ő kutat. Ő gyereket nevel. Ő zenél. Ő takarító. Ő politikus. Ő sok pénzt keres. Ő süteményt süt. Ő professzor. Ő asszisztens.

She is beautiful. He is clever. He reads. She washes the dishes. He builds. She sews. He teaches. She cooks. He's researching. She is raising a child. He plays music. She's a cleaner. He is a politician. He makes a lot of money. She is baking a cake. He's a professor. She's an assistant.



Biases in AI systems

PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models

Sachit Menon*, Alexandru Damian*, Shijia Hu, Nikhil Ravi, Cynthia Rudin
Duke University
Durham, NC
{sachit.menon,alexandru.damian,shijia.hu,nikhil.ravi,cynthia.rudin}@duke.edu

Outline

1. Biases in AI systems
2. **Fairness in machine learning: binary decisions**
3. Beyond fair learning

Fairness in Machine Learning – Binary predictions

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

The Apple Card Is the Most High-Profile Case of AI Bias Yet

Binary decisions: Good vs. Bad outcome

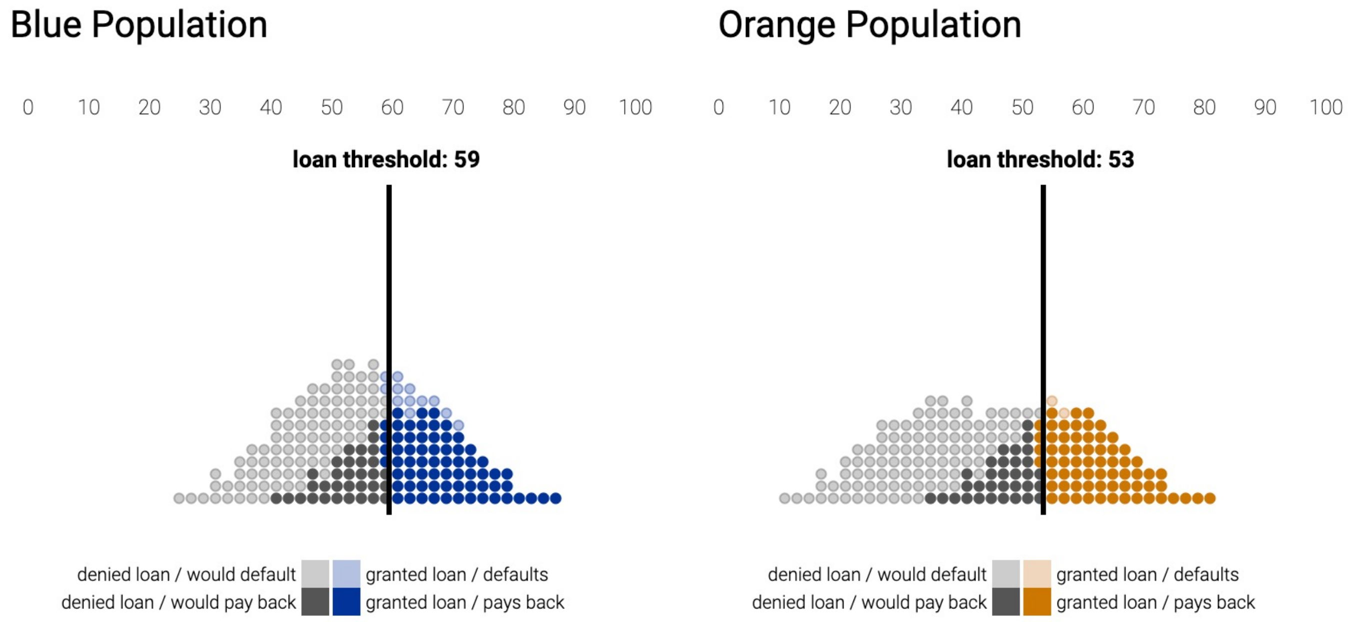
Applications: Recidivism prediction, Loan approval, Job application

Fairness in Machine Learning – Typical setup

	Example: Lending
A sensitive attribute	Gender (Men/Women)
X "relevant" features	Salary, Debt history
Y actual outcome	Repaid / Default
$\hat{Y}$ = $\mathrm{f}(X, A)$ predictor	Classifier
$\hat{S} = g(X, A)$ score function (can be turned into binary decision)	Credit score

Fairness criteria in Machine Learning	
Demographic parity	$\mathrm{P}(\hat{Y} = 1 \mid A = 0) = \mathrm{P}(\hat{Y} = 1 \mid A = 1)$
Equal opportunity	$\mathrm{P}(\hat{Y} = 1 \mid Y = 1, A = 0) = \mathrm{P}(\hat{Y} = 1 \mid Y = 1, A = 1)$
Calibration within groups	$P\left(Y = 1 \middle\vert \hat{S} = s, A = 0\right) = P\left(Y = 1 \middle\vert \hat{S} = s, A = 1\right)$

→ Incompatibility



S. Corbett-Davies et al. ’17
J. Kleinberg et al. ‘16, A. Chouldechova ’16
https://research.google.com/bigpicture

Trade-offs

Many more definitions…

- More parity measures
- Individual metric-based fairness
- Counterfactual fairness

And trade-offs:

- Between different measures of group fairness
- Between group fairness and individual fairness
- *Between group fairness and group fairness*
- Between fairness and utility

Dwork et al., *Individual fairness*, 2012
Kusner et al., *Counterfactual fairness*, 2017
Kearns et al., *Preventing fairness gerrymandering*, 2017

Fair algorithms

1. Pre-processing

   Learning fair representations

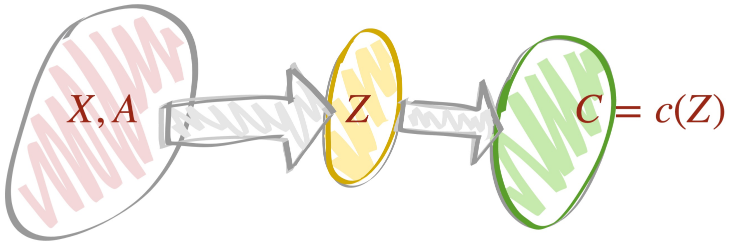
2. Optimization at training time

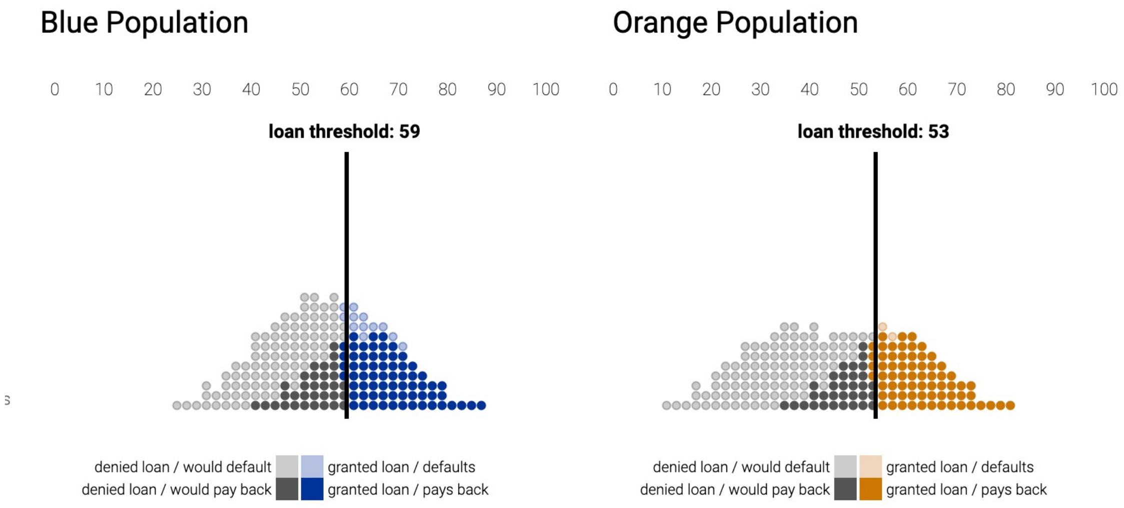
   Empirical risk minimization with constraint, regularization term

3. Post-processing

   Threshold on a score function

Representation learning approach





Additional references

Tutorials

- Hardt and Barocas, tutorial @ NeurIPS ’17
- Narayanan @ FAccT ‘18

Surveys

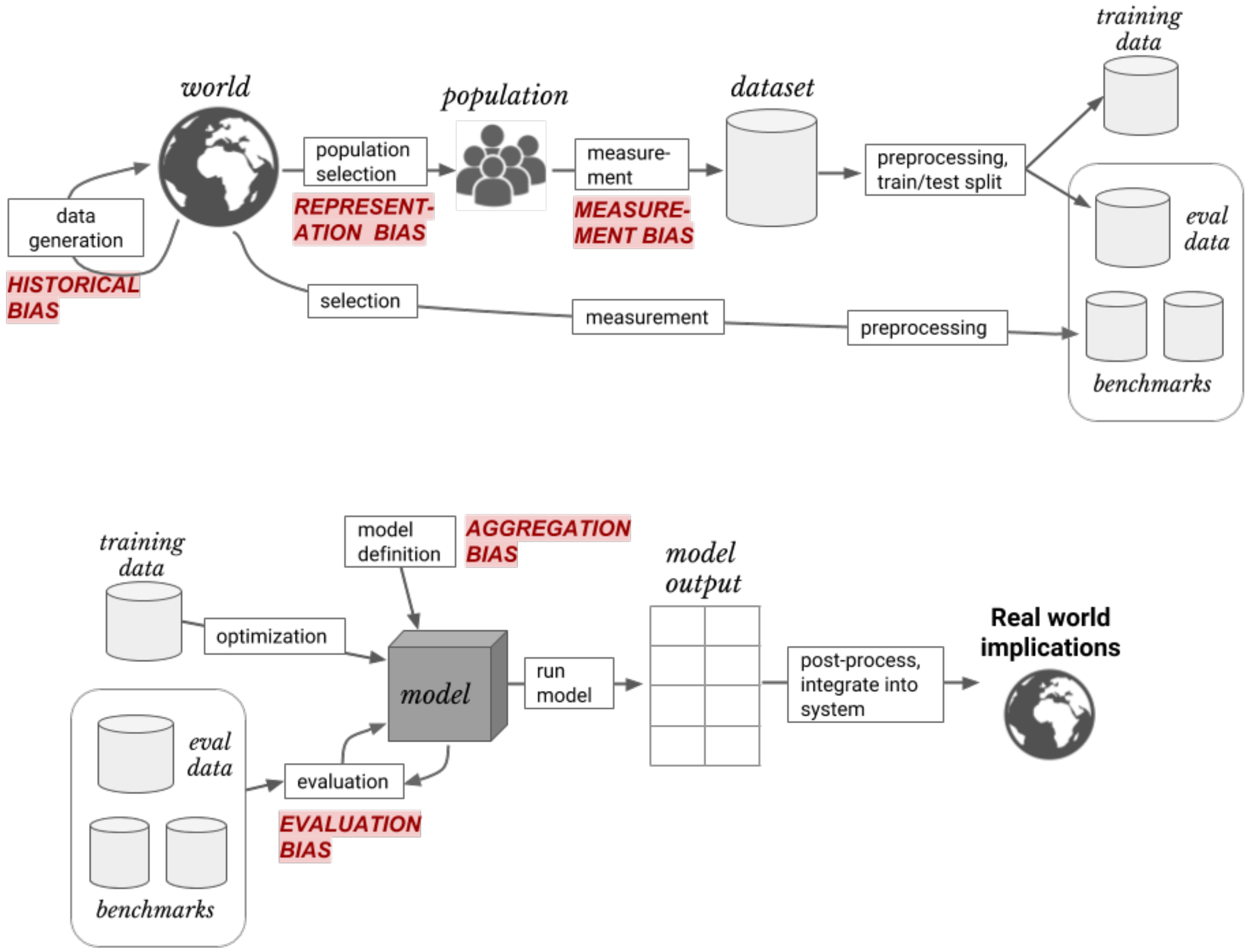
- Chouldechova and Roth, *The frontiers of fairness in machine learning*
- Corbett-Davies and Goel, *The measure and mismeasure of fairness*
- Barocas, Hardt, Narayanan, *Fairness and machine learning: limitations and opportunities.* [Book]

Outline

1. Biases in AI systems
2. Fairness in machine learning: binary decisions
3. **Beyond fair learning**

Biased data?

Potential sources of harm arise at different stages of the ML pipeline



Five potential sources of harm (Suresh and Guttag, 2019)

Parity vs. preference

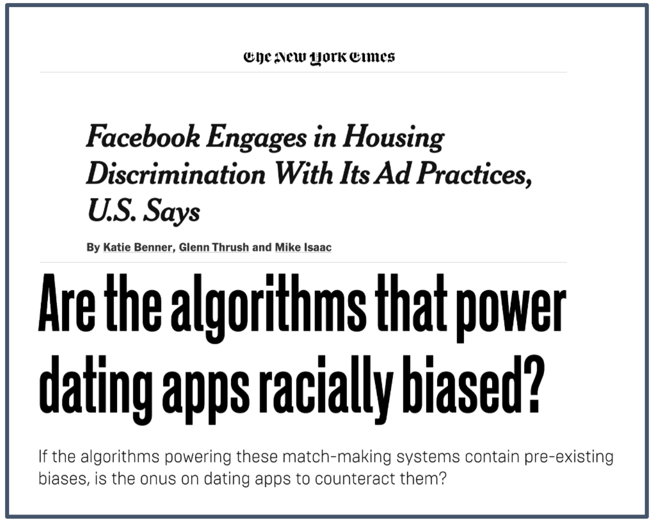
When subjects have different preferences / utilities, should they be given the same predictions?

→ Personalization

Preference guarantees

with concepts like “envy-freeness”: no one should prefer someone else’s model to their given model.

Zafar et al. ’16, Ustun et al. ’19, Balcan et al. ’19, Kim et al. ‘20


The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

Are the algorithms that power dating apps racially biased?

If the algorithms powering these match-making systems contain pre-existing biases, is the onus on dating apps to counteract them?

Discussion

- Interdisciplinarity
  - “Mathematical” fairness for computer scientists vs. fairness for ethicists, philosophers, legal scholars, economists…
- Context
  - Applications: which fairness definition for which specific context? should ML be used at all?
  - Fairness for unobserved characteristics: ethnicity, sexual orientation.
  - Complex pipelines
- Explainability