Fairness

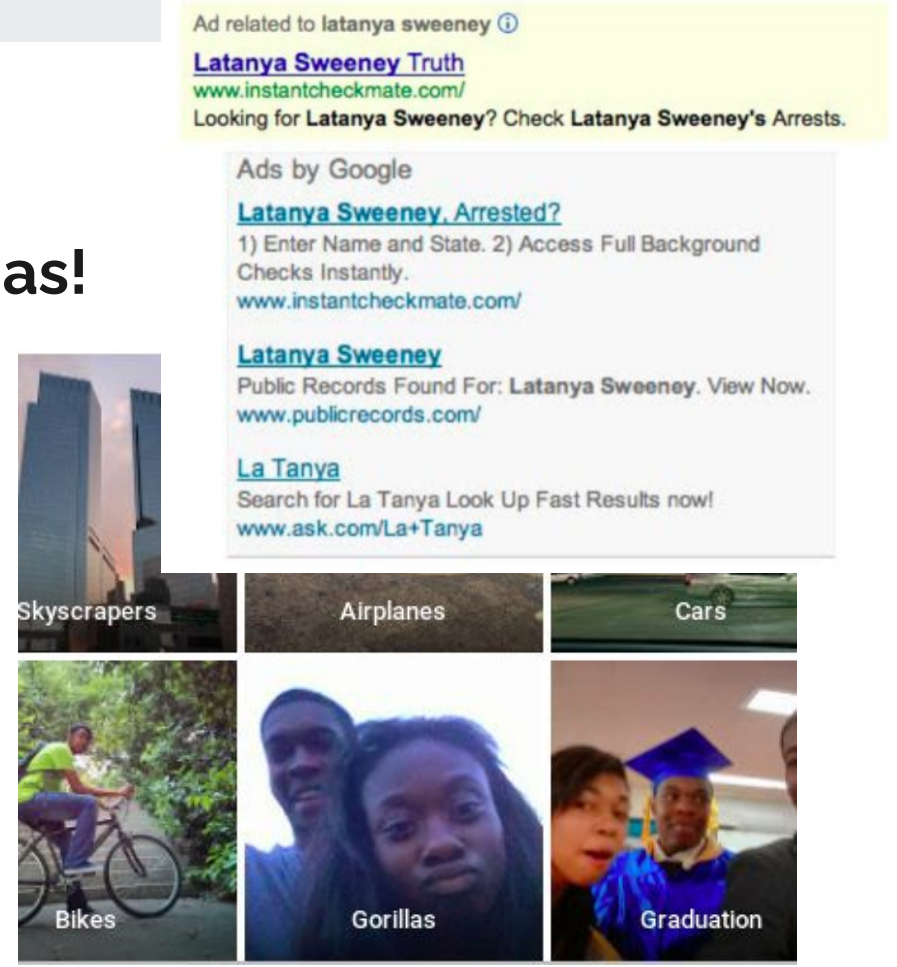
Machine Learning in Production / AI Engineering - Recitation 9

Myth: Data and ML Tools Are Neutral!

- Translating high-level goals into data is not neutral.
- Data at best reflect the current state of the world.
- Learning algorithms pick up the patterns in data.

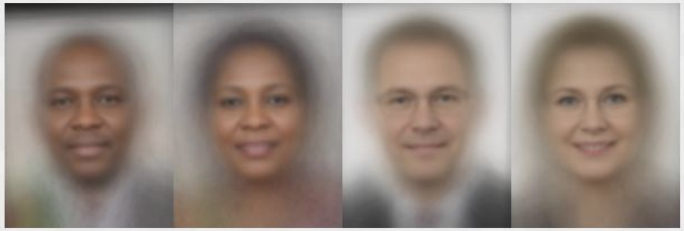
Millions of Examples of AI Bias!

Millions of Examples of AI Bias!



Millions of Examples of AI Bias!

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



jews should

- jews should **be wiped out**
- jews should **leave israel**
- jews should
- jews should **get over the holocaust**
- jews should **go back to poland**
- jews should **apologize for killing jesus**
- jews should **all die**
- jews should **be perfected**
- jews should **not have a state**

Let’s play with a visual simulation of discrimination in ML

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

Responsible AI

"More than half of the interviewees worked on their initiatives as individuals and not as part of a team (14 out of 26). About 40% of the respondents reported that they volunteer time outside of their official job function to do their work on responsible AI initiatives…"

"... interviewees described going through stress-related challenges in relation to their responsible AI work. During some of the interviews, we saw a noticeable tone change in the interviewees' voice when discussing questions related to ethical tensions, accountability, risk culture, and others."

Rakova, B., Yang, J., Cramer, H. and Chowdhury, R., 2021. **Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices.** Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), pp.1-23.

Trends

- **Prevalent practices:** Responsible AI work perceived as a “taboo topic”
- **Emerging practices:** Flexibility and some support from the legal teams
- **Aspirational future:** Having organizational frameworks to encourage anticipatory approaches

Rakova, B., Yang, J., Cramer, H. and Chowdhury, R., 2021. **Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices.** Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), pp.1-23.

Sharing experience...

> *"When prompted, very few of our interviewees report considerations for fairness either at the product or the model level. Only two participants from model teams (P14a, P22a) reported receiving fairness requirements, whereas many others explicitly mentioned that fairness is not a concern for them yet (P4a, P5b, P6b, P11a, P15c, P20a, P21b, P25a, P26a)."*

- Model teams usually do not feel responsible for fairness

Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. **Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process.** In 44th International Conference on Software Engineering (ICSE '22), https://arxiv.org/pdf/2110.10234.pdf

Let's discuss...

- How you can make your colleagues/company care about fairness in ML?

Exercise -
Job Candidate Screening

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Goals & Stakeholders

Population groups & distribution

- Background, Country of origin, Degree , Work experience, Gender, Race, etc

Goal

- Get the best and qualified applications for the recruitment
- Promote diversity and not be unfair to certain minority groups

Assumptions

- Applicants are filling in information that reflect reality, and are authentic
- Applications aren't recruited outside of what the AI system recommends
- All and correct data is provided from the past recruitments

Harms & Biases

Harms

- Allocation / Representation?
- Stereotyping?

Biases

- Historical?
- Tainted examples?
- Skewed samples?
- Limited features?
- Sample size disparity?
- Proxies?

Harms & Biases

Allocation / Representation

- Male candidates are more often recruited
- Candidates with only CS background are selected more often
- Candidates from a only subset of countries are selected more often
- Candidates from similar background as existing employees are selected more often

Harms & Biases

Biases

- Historical
  (past biases - recruitment committee / before AI)
- Tainted examples
  (dataset biased due to improper human labeling)
- Skewed samples
  (manual recruitment biases skews future recruitments; ex: selection based on existing ones)
- Limited features

Harms & Biases

Biases

- Sample size disparity
  (insufficient data for applications from certain countries / backgrounds)
- Proxies
  (ex: type of undergraduate degree can serve as a proxy for gender - not always applicable)
- Population bias
  (training vs target demographic), Outliers, etc.

Biases in Data Collection & Preprocessing

Data

- Past applicant's application packets
- Employer's recruitment records, reasons, etc.

Data collection

- Acquisition (sufficient applications data was not recorded/kept track of until few years ago)
- Querying (the whole applicant pool data was not used to train on)
- Filtering (some application data was considered useless after company structure changed)

Data preprocessing

- Cleaning (some applicants did not fill in all data required in the packet, so they were removed)
- Enrichment (we filled in placeholder or average values based on historical data for some fields)
- Aggregation (multiple areas of study were merged into a single feature)

Fairness

- What do want to achieve - equality or equity?
- What type of fairness definition/metric is appropriate?
  - Anti-classification
    - ??
  - Group Fairness/Independence
    - ??
  - Separation w/ FPR or FNR
    - ??

Fairness

- What do want to achieve - equality or equity?
  - Equity – Depends on goal

Fairness

- What type of fairness definition/metric is appropriate?
  - Anti-classification
    - Gender/race, etc. should not be considered at all for recruitment
    - Recall: proxies
  - Group Fairness/Independence
    - Rate of positive predictions (acceptance rate) is the same across groups
    - Vary thresholds for different groups - Equity
  - Separation w/ FPR or FNR
    - False positive and negative rates are the same across groups
    - Vary thresholds for different groups - Equity
    - Ex: FNR - probability of incorrectly being rejected is equal across groups

Thank you!

Keep fairness in mind while designing ML software