Bias/Fairness detection and mitigation in ML models

-Srimugunthan Dhandapani

Outline

1. **Introduction of Model risk**
   Case studies of Faulty AI and machine learning models, what constitutes a Model risk

2. **Model Bias/Fairness: Definition and Detection**
   What is model bias and fairness, tradeoffs, Bias detection

3. **Bias Mitigation techniques**
   Adversarial debiasing, Reweighing etc

4. **Causal approach to Fairness**
   Causality applied to model fairness

Introduction to Model risk

Apple credit card case

WIRED SECURITY POLITICS GEAR BACKCHANNEL BUSINESS SCIENCE CULTURE IDEAS MERCH

The Apple Card Didn't 'See' Gender—and That's the Problem

The way its algorithm determines credit lines makes the risk of bias more acute.

DesignNews SIGN UP TODAY Sig

Automation Automotive Electronics Design Materials 3DP Industry CEC DN Direct

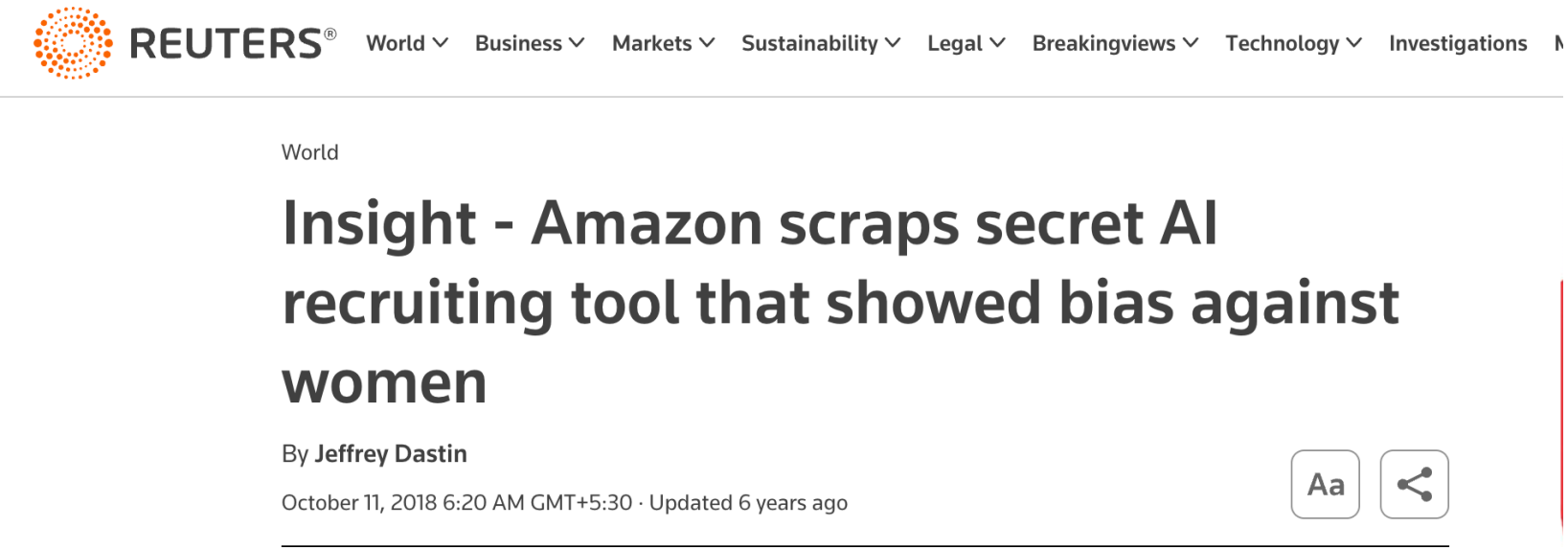
The Apple Card Is the Most High-Profile Case of AI Bias Yet

Apple Card users have alleged that its credit decision algorithm discriminates against women.

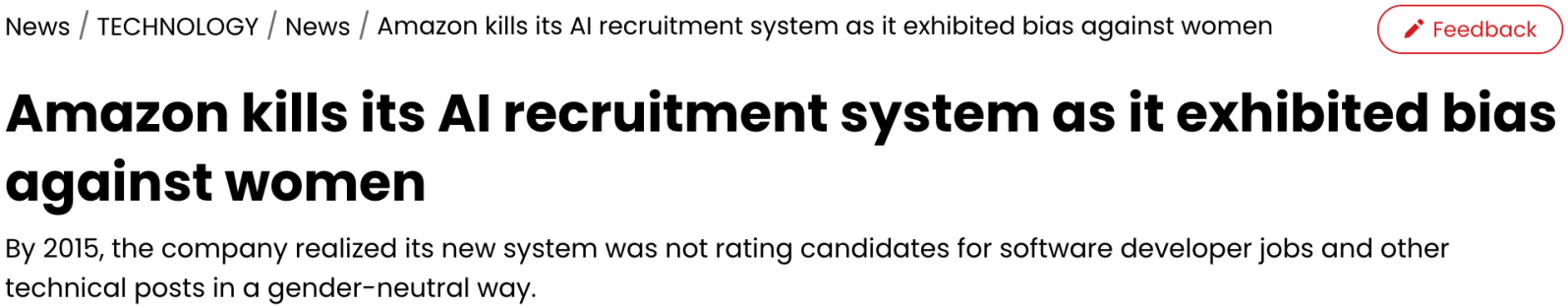
- The algorithm responsible for credit decisions for the Apple Card is giving females lower credit limits than equally qualified males.
- Provoked investigation from Department of Financial Services (DFS) to determine whether New York law was violated and ensure all consumers are treated equally regardless of sex.

https://www.designnews.com/artificial-intelligence/the-apple-card-is-the-most-high-profile-case-of-ai-bias-yet

Amazon recruiting tool case



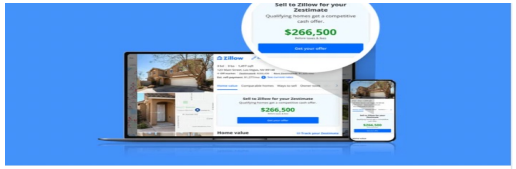
- experimental hiring tool used artificial intelligence to give job candidates scores
- Screen resumes and output top 5 relevant resumes
- Problem was models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men



- downgraded graduates of two all-women's colleges

https://www.businesstoday.in/technology/news/story/amazon-killed-its-ai-recruitment-system-as-it-exhibited-bias-against-women-152336-2018-10-11

Flawed AI model cases



Zillow to exit its home buying business, cut 25% of staff

cnn.com · 2021

Zillow is getting out of the iBuying business and will shut down its Zillow Offers division, resulting in a 25% reduction in its staff.

In its quarterly earnings report on Tuesday, the company said it will see a total write-down of more than $540 million as a result of its exit from the business, which buys homes and resells them.

As a result of shutting down Zillow Offers, the company said it wil...

Robot Stabs A Man To Death At A Factory In Haryana's Manesar!

indiatimes.com · 2015

This one's straight out of a Terminator film. Sharp welding sticks jutting out of the robotic arm of a machine pierced a worker killing him at a factory here on Wednesday. The worker had apparently moved too close to the robot while adjusting a metal sheet that had come unstuck.

Waze 'directed tourists to drive into lake'

dailymail.co.uk · 2018

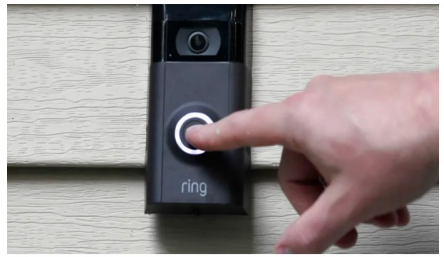
Waze has yet to give an explanation as to why the app directed a group of tourists to drive into Lake Champlain earlier this month.

UK passport photo checker shows bias against dark-skinned women

bbc.co.uk · 2020

Women with darker skin are more than twice as likely to be told their photos fail UK passport rules when they submit them online than lighter-skinned men, according to a BBC investigation.

One black student said she was wrongly told her mouth looked open each time she uploaded five different photos to the government website.



FTC charges Amazon with privacy violations over Alexa and Ring cameras

11alive.com · 2023

WASHINGTON — Amazon will pay more than $30 million to settle alleged privacy violations involving its voice assistant Alexa and its doorbell camera Ring.

https://incidentdatabase.ai/apps/discover/

Legal and Regulatory Landscape

EU

- EU AI act
- GDPR

United States

- AI bill of rights
- The National AI initiative act
- The California Privacy rights act
- Local law 144
- National Institute of Standards and Technology Risk Management framework

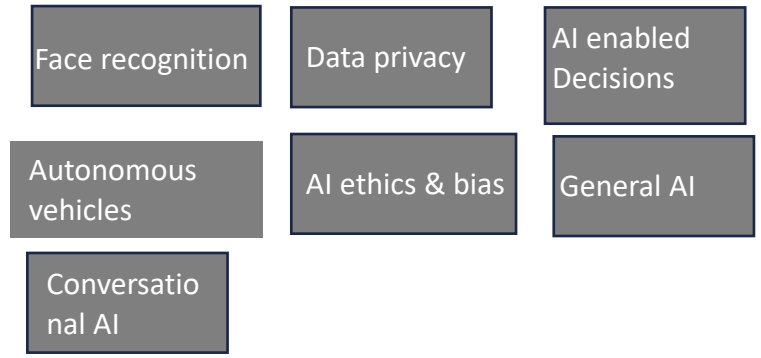
China

- Cyber security law of people's republic of China
- Data security law
- Personal Information protection law

India

- Digital Person Data Protection Bill of 2022
- Data security Law
- Personal Information protection law
- draft National Data Governance Framework Policy (NDGFP)
- New AI advisory Mar 2024

Top Areas coming under regulations:

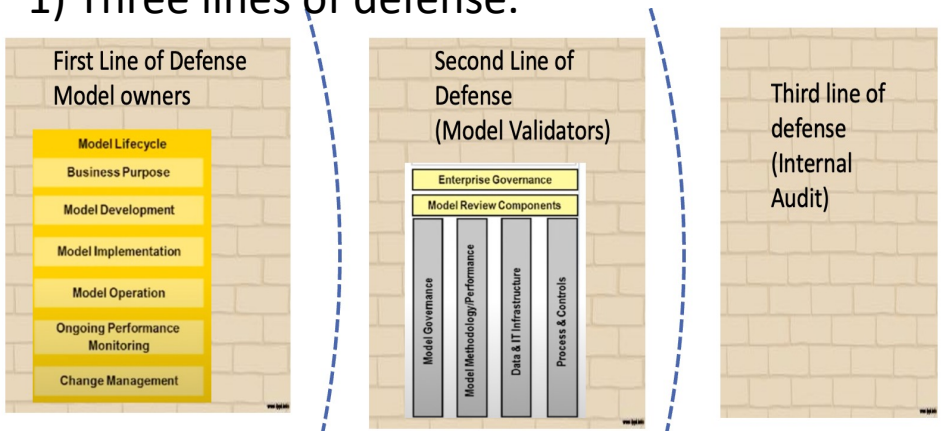


EU AI act:

- Takes a risk based approach
- Assign different risk categories and groups system from "BANNED" to "UNREGULATED"
- Classification for AI systems based on their risk level and specifies limitations and obligations on the level of risk

Model risk management

1) Three lines of defense.



2) Model governance framework

3) Model risk built into procedures of Model lifecycle. Considers compliance risk, data risk, Technology risks, third party risks, business process risks

4) Start with Model identification and risk ranking before development
5) Maintain inventory of models, Model versioning, guidelines for best practices
6) Procedures and checks before adopt or using model
5) Ongoing monitoring, Production checks
6) Periodic review and Audit

- **Popular Risk Management Frameworks:**
  - NIST AI Risk Management Framework .
  - ISO 23894 - Recently released standard for AI risk management.
  - FRB. SR11-7 Supervisory guidance on Model risk management

Example of Governance framework for model risk

	Strategy	Design	Model	Evaluate	Deploy & Evolve
Integrity Understand & Track Lineage Protect Reputation	— Alignment with strategy and business requirements — Available corporate policies and guidelines	— Use of allowed data sources and inputs — Test for data quality — Qualified SMP's involved — Data provenance check	— Evaluate training methodology or procedures — Feature provenance	— Experiment setup and configuration — Quality control model experiments and evaluation reports — Model accuracy and precision	— Runtime model metrics detection — Continuous training governance and assessment — Implementation control
Explainability Achieve Transparency Gain Confidence	— Adherence to data usage guidelines — Explanatory feature names	— Explainability requirements and schema/template defined	— Check for model metadata including attributes	— Explainability testing and acceptance	— Model improvement / change log — System Documentation — Report Generation
Fairness Be Inclusive & Ethical Ensure Appropriate Use	— Published list of allowed features	— Validation and quality check of ground truth — Bias verification and mitigation of ground truth (train, test & evaluation)	— Features compliance with policies, business requirements and regulations	— Model and concept drift evaluation — Inclusiveness testing — Model risk scoring	— Setup for continuous monitoring of fairness & accuracy — Escalation process
Resilience Serve & continuously monitor Prevent Attacks	— Model usage guidelines, restrictions, and specifications — Defined model SLA's — Required skills & support to manage and maintain	— Data usage guidelines; data privacy and protection	— Training data access protection and traceability — Model deployment / serving interoperability	— Use of approved frameworks and runtimes — Security vulnerability and adversarial attack testing — Model and concept drift detection	— Program execution — Model access and ACL — Continuous monitoring, protection and testing *(recalibration, incident response, BCP)* — Usage & feedback data protection — Model breach / Incident response plan
The Value of the Framework	Mapping to business needs Ethics and policy adherence Model measurement metrics	Understanding data lineage Detect imbalances in data Feature Analysis Bias detection and mitigation	Check feature compliance Modeling assumptions Audit Logging Hyperparameter changelog	Business Operational indicators Model explainability Evidence profiles Adversarial and security testing Continuous model monitoring	Production readiness Interoperability and serving Continuous protection Monitoring for metrics and drift Model and data governance

Source: "Model risk management" Hakan Gogtas

"Model risk management" Hakan gogtas, https://www.youtube.com/watch?v=UXv5P66f6PY

Model Bias/Fairness: Definition and Detection

What is model bias

- AI model bias refers to the systematic error that occurs when a machine learning algorithm produces results that are unfairly skewed in favour of or against certain groups of people.

Protected attributes/Sensitive attributes

https://arize.com/blog/evaluating-model-fairness/

Defining model fairness

- Multiple definition of fairness
  - **Individual Fairness**: Based on the criterion that similar individuals would receive similar predictions,
  - **Group Fairness**: Based on criterion that emphasizes equal outcomes for different groups.
  - **Predictive Parity**: Focuses on equal positive and negative predictive power across groups.
  - **Equalized Odds**: the probability of receiving a positive outcome (e.g., loan approval) given a true positive condition (e.g., deserving applicant) is equal across groups.
  - **Process Fairness** How fair is the process? (In contrast with how fair are outputs)
  - **Counterfactual Fairness** – Had any individual been of a different race, sex, etc. the prediction would not change
  - **Equal opportunity fairness** - an algorithm gives the same prediction given the outcome regardless of the protected attribute.
  - …

"21 definitions of fairness" : https://www.youtube.com/watch?v=jIXIuYdnyyk
https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/

Fairness measures

- **Statistical Parity Difference**: Difference in rate of favourable outcomes received by unprivileged group to privileged group
- **Selection rate**: Based on criterion that emphasizes equal outcomes for different groups.
- **Average Odds difference**: The average difference of false positive rate and true positive rate between groups
- **Disparate impact** The ratio of favourable outcome for the unprivileged group to that of privileged group
- **False Positive rate parity:** Difference between the false positive rate between privileged and unprivileged group
- **Equal opportunity difference**: Difference between true positive rates between privileged and unprivileged group
- **Error rate parity difference:** Difference between error rate between two groups
- **Predictive parity difference:** Predictive rate parity is calculated by the division of true positives with all observations predicted positives. Difference between predictive rate parity between two groups

https://medium.com/data-science-at-microsoft/measuring-fairness-in-machine-learning-3211b62340b

Fairness Constraints and Tradeoffs

- **The fairness impossibility theorem :**
  - is a fundamental result in algorithmic fairness that states that it's not possible to simultaneously satisfy all three common definitions of fairness when fitting statistical models: demographic parity, equalized odds, and predictive rate parity
- Other tradeoffs like Fairness and Accuracy
- Although, recent papers show it is possible to meet approximate fairness constraints with a margin of error

Inherent tradeoffs in Fairness: https://www.youtube.com/watch?v=p5yY2MyTJXA
https://www.youtube.com/watch?v=TJJxidzmCsY

Bias detection

How to detect bias

1.Data Analysis:

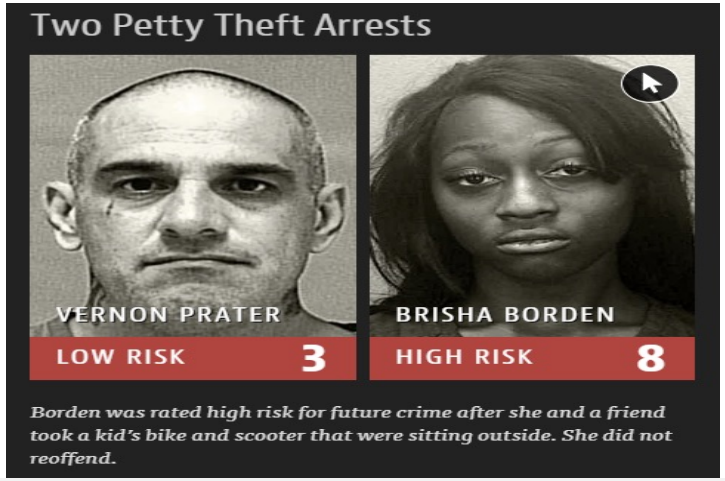
1. Conduct a comprehensive analysis of the dataset to identify biases in data collection or labeling.
2. Examine distributions of different groups to uncover any disparities or under-representations.

2.Metrics for Fairness:

1. Define appropriate fairness metrics depending on the context, such as demographic parity, equalized odds, or disparate impact.
2. Evaluate the model's performance across different groups based on these metrics.

Example

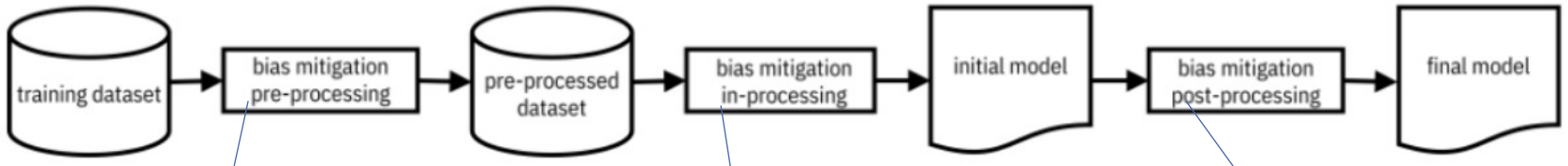
- COMPAS DATASET: The dataset for this case study is the *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) dataset, which is a collection of criminal offenders
- Predict recidivism (whether or not a person will reoffend and get re-arrested), given certain characteristics about an individual



https://learning.oreilly.com/library/view/feature-engineering-bookcamp/9781617299797/OEBPS/Text/CH04.htm
https://www.rdocumentation.org/packages/fairml/versions/0.8/topics/compas

Bias/Fairness mitigation

Mitigation Strategy



Pre-processing

- Disparate Impact Remover
- Reweighing
- Learning fair representation
- ..

In-processing

- Adversarial De-biasing
- Exponentiated Gradient reduction
- Gerry-Fair classifier
- ..

Post-processing

- Equalized odds postprocessing
- Reject option classification
- ..

https://aif360.readthedocs.io/en/v0.4.0/modules/algorithms.html#id15

Pre-processing Techniques

- Algorithms correcting bias at pre-processing remove information about dataset variables which might result in unfair decisions.

Re-weighing technique

- Reweighing is an example of a pre-processing algorithm.
- Weigh each observation in the training dataset by the expected probability of the observation ignoring the protected attribute.

$$W(X) = \frac{P_{obs}(X)}{P_{exp}(X_{i \neq A})}$$

- Fit a weighted logistic regression

```
lr = LogisticRegression()
lr.fit(X_train, y_train, sample_weight = weights)
```

Sex	Ethnicity	Highest degree	Job type	Cl.	Weight
M	Native	H. school	Board	+	0.75
M	Native	Univ.	Board	+	0.75
M	Native	H. school	Board	+	0.75
M	Non-nat.	H. school	Healthcare	+	0.75
M	Non-nat.	Univ.	Healthcare	–	2
F	Non-nat.	Univ.	Education	–	0.67
F	Native	H. school	Education	–	0.67
F	Native	None	Healthcare	+	1.5
F	Non-nat.	Univ.	Education	–	0.67
F	Native	H. school	Board	+	1.5

$$\frac{Pr(Sex = female) \times Pr(Y = +)}{Pr(Sex = female, Y = +)} = \frac{\frac{5}{10} \times \frac{6}{10}}{\frac{2}{10}} = 1.5$$

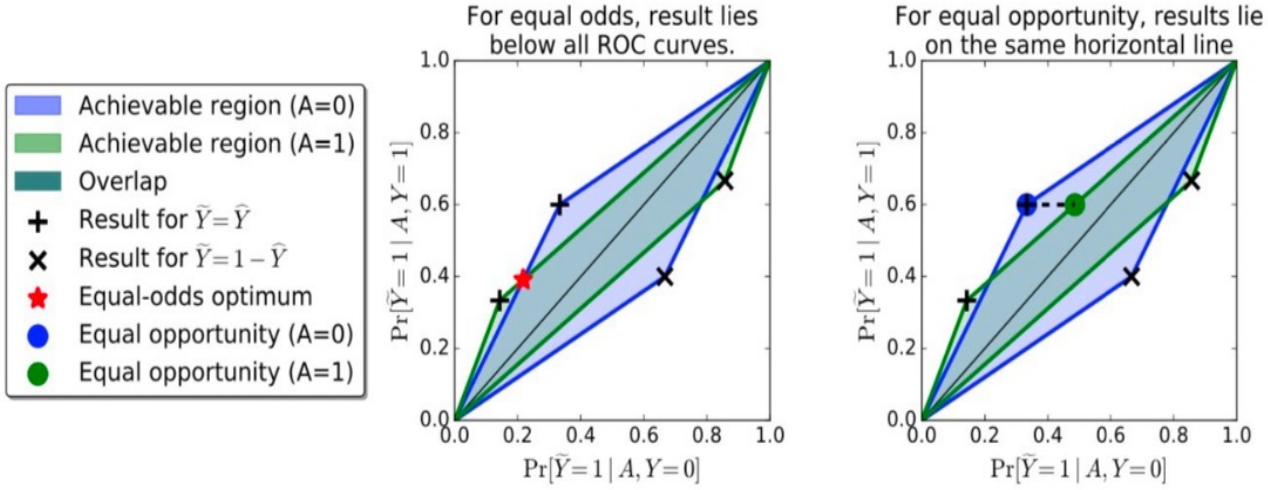
https://learning.oreilly.com/library/view/practical-fairness/9781492075721/ch04.html#idm46523157974632
https://github.com/PracticalFairness/BookRepo/blob/main/Ch04/Ch04.ipynb

Post-processing techniques

- Post-processing bias mitigation techniques adjusts the outcomes of a model to mitigate bias in predictions.

Equalized odds post processing technique:

The first method comes from a much-cited 2016 paper titled, “Equality of Opportunity in Supervised Learning” by Hardt et al



https://learning.oreilly.com/library/view/practical-fairness/9781492075721/ch06.html#idm46523155493144
https://github.com/gpleiss/equalized_odds_and_calibration/blob/master/eq_odds.py
https://github.com/Trusted-AI/AIF360/blob/main/aif360/algorithms/postprocessing/eq_odds_postprocessing.py

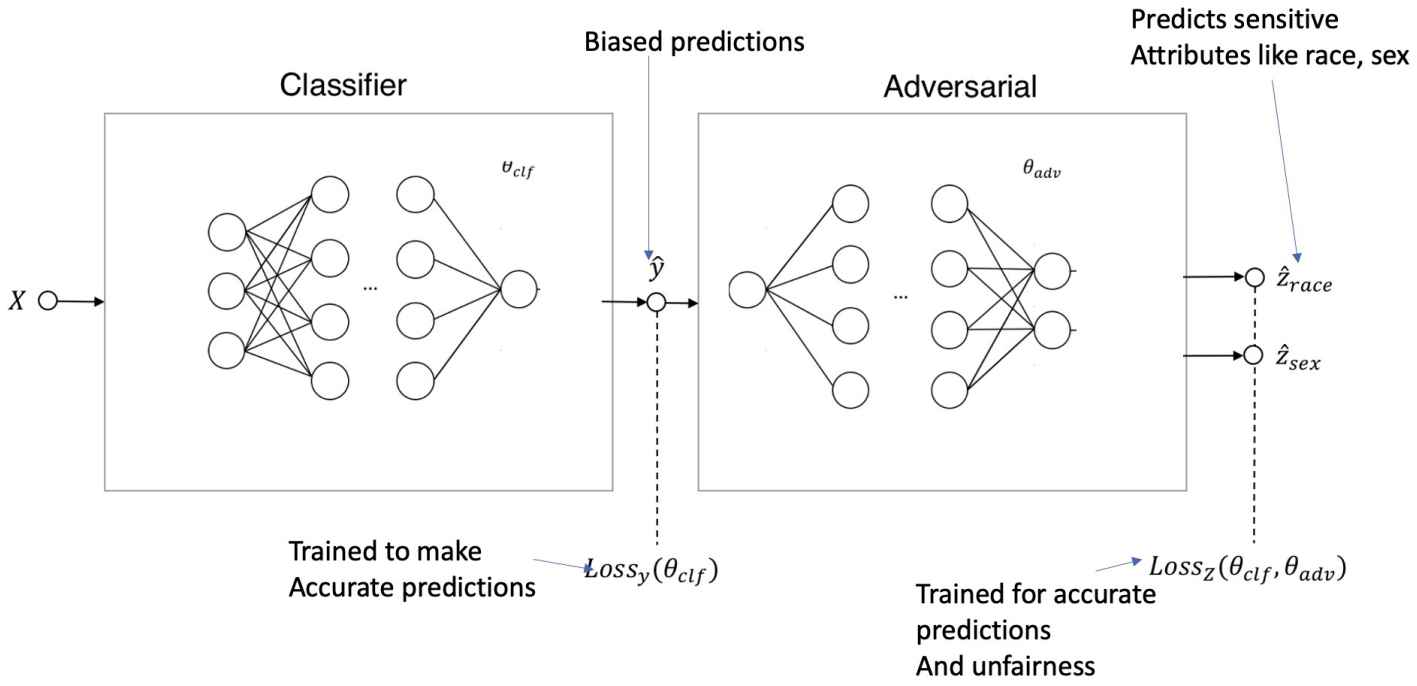
ALGORITHM

- **Input:** 1. *Trained model's predictions*: These are the initial predictions for the target variable made by the model on the data, 2. *True labels*: The actual values of the target variable for the data points, 3. *Sensitive attribute*: This is the group membership (e.g., race, gender) for which you want to achieve fairness.
- **Separate by group:** The algorithm separates the data based on the sensitive attribute into privileged and unprivileged groups.
- **Calculate true positive rates (TPR):** It calculates the TPR for each group. TPR is the proportion of individuals correctly classified as positive (e.g., not defaulting on a loan) within a group.
- **Solve linear optimization program:**. The domain for solutions lies in the overlapping region of two quadrilaterals . An optimization problem is formulated as a linear program . This program aims to find a new set of predictions that minimizes the difference between the TPRs of the privileged and unprivileged groups, subject to certain constraints.
- **Adjusted predictions:** The algorithm outputs these new, adjusted predictions that satisfy the equalized odds criteria. In simpler terms, the predictions are modified to ensure both groups have an equal chance of receiving a positive outcome (e.g., loan approval) given they truly belong to that outcome class (e.g., creditworthy borrowers).

In-processing techniques

- In-processing bias mitigation techniques corrects the bias during model training time

Adversarial debiasing technique



https://xebia.com/blog/towards-fairness-in-ml-with-adversarial-networks/

ALGORITHM

•**Train the Predictor:** The predictor model (P) is initially trained on the original data to make accurate predictions for the target variable.

•**Train the Adversary:** Then, the adversary model (A) is trained to predict the sensitive attribute based on the outputs of the predictor model (P). In essence, the adversary is trying to learn how the predictor's predictions might be influenced by the sensitive attribute.

• **Iterative Refinement:** Here's the adversarial part: The predictor (P) is updated to make its predictions **less informative** about the sensitive attribute. This makes it harder for the adversary (A) to predict the sensitive attribute based on the predictions. The adversary (A) is then re-trained to adapt to the changes in the predictor.

Equilibrium: This process of training and updating continues iteratively until an equilibrium is reached. Ideally, at this point, the predictor (P) maintains good accuracy for the target variable prediction, while the adversary (A) is no longer able to effectively predict the sensitive attribute from the model's outputs.

Demo

- "Adversarial debiasing git repo" https://github.com/equialgo/fairness-in-ml/tree/master

Causal approach to Fairness

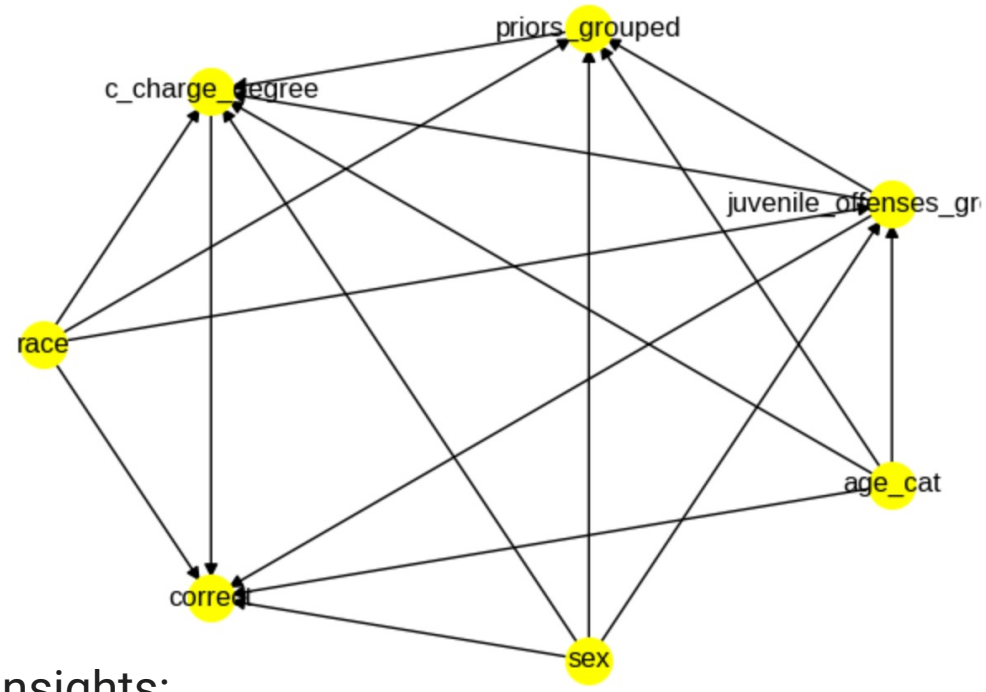
Causal approaches

- **Causal inference techniques**
  - Causal inference is a statistical method for determining cause-and-effect relationships between variables. The goal of is to **estimate the causal effect** of one variable on another.
  - These techniques account for confounding variables and other challenges to isolate the true causal effect.
- Causal modelling:
  - Causal modeling is a specific **tool** used within causal inference.
  - creating a **visual representation** (often a directed acyclic graph or DAG) of the relationships between variables. This model depicts the cause-and-effect structure, including potential confounders.
- **Structural causal modeling (SCM)** is a specific type of causal modeling that dives deeper into the causal structure between variables.
  - SCM assumes you have a complete understanding of the causal relationships between variables in a system.
  - It goes beyond just identifying correlations to represent cause-and-effect through a set of equations and a directed acyclic graph (DAG).

Fairness analysis using Causal model

Causal fairness analysis

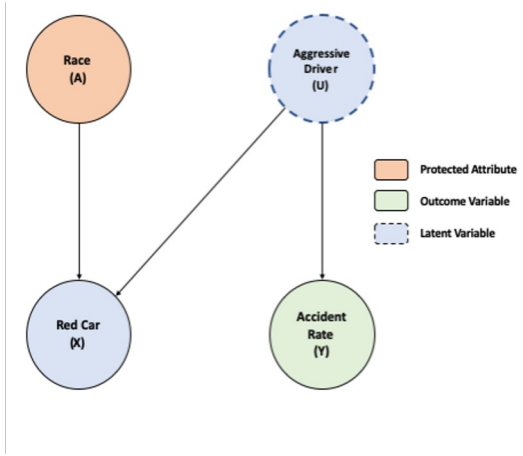
1. The creation of a causal graph formalizes assumptions about how the world works (in this case, how bias might be perpetuated by a machine learning model or system)
2. Framing fairness analysis as a causal inference task allows us to understand and directly quantify mechanisms of bias

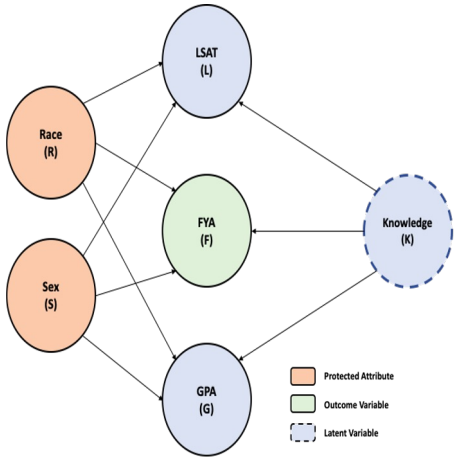


Analysis insights:
African American defendants have a **21.8% greater chance** of being classified as high risk, and that this difference is **due directly to their race**

https://appliedcausalinference.github.io/aci_book/10-fairness.html
https://colab.research.google.com/drive/1bviQfw1BWtgKw4O-XXPAqXb-t7vtMRls?usp=sharing

Counterfactual fairness



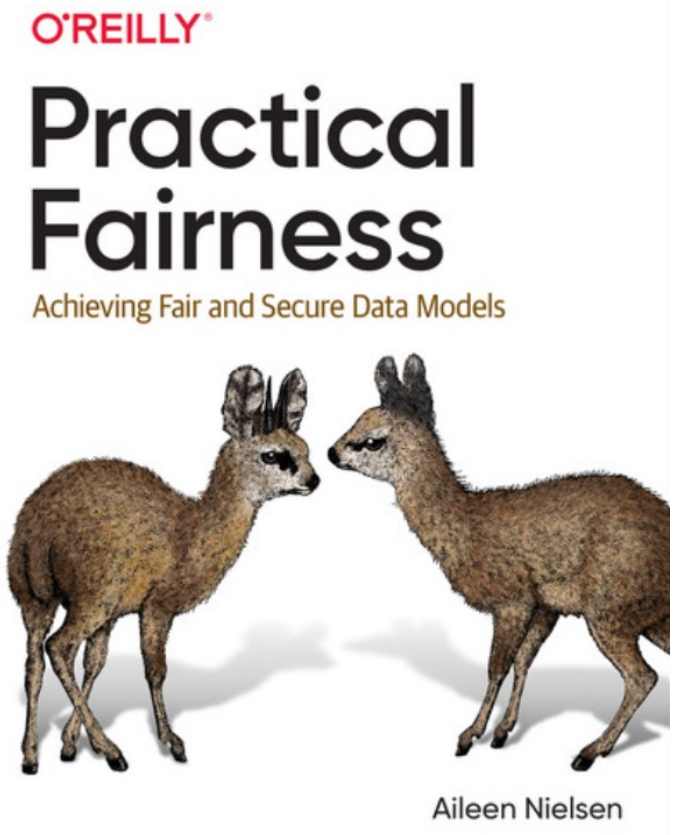


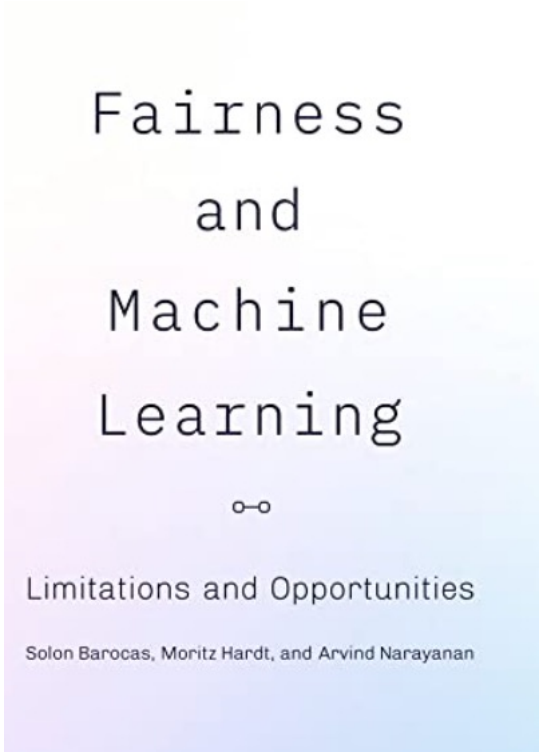
- **Counterfactual fairness paper** "Counterfactual Fairness", Kusner et al, 2017
  - explicitly modeling the causal structure of the world.
  - In simple terms we are essentially extracting the latent variable from observed variables and then using that latent variable and non-descendents of the protected attribute in the predictive model.

https://github.com/Kaaii/CS7290_Fairness_Eval_Project/blob/main/Counterfactual%20Fairness%20Final.ipynb
https://github.com/yunhao-tech/Counterfactual-Fairness

Thanks

Books to refer:





LinkedIn connect

https://www.linkedin.com/in/srimugunthan-dhandapani/