CENG7880
Trustworthy and Responsible AI

Instructor: Sinan Kalkan

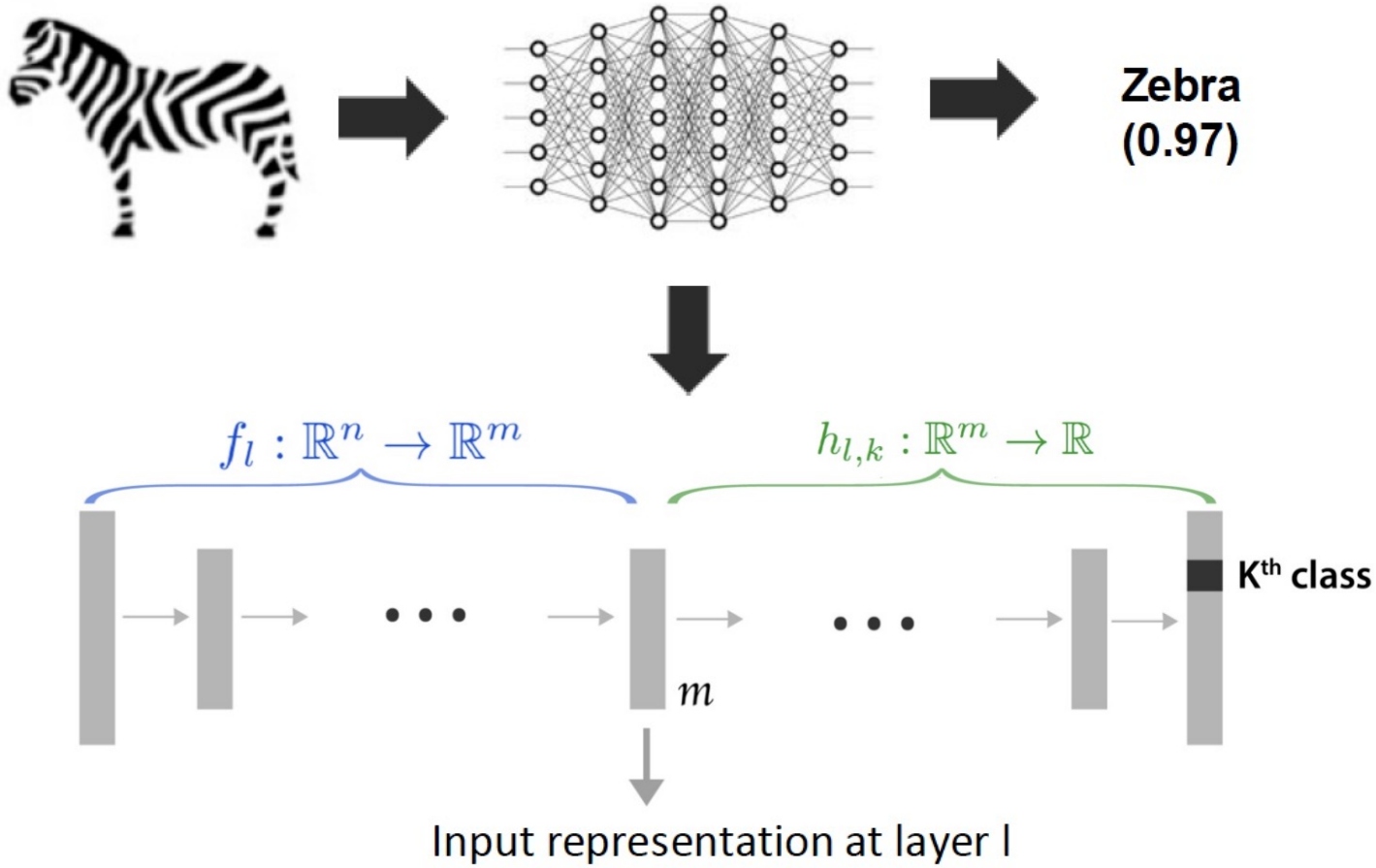
(https://ceng.metu.edu.tr/~skalkan)

For course logistics and materials:
https://metu-trai.github.io

MIDDLE EAST TECHNICAL UNIVERSITY
DEPARTMENT OF COMPUTER ENGINEERING

Previously on CENG7880

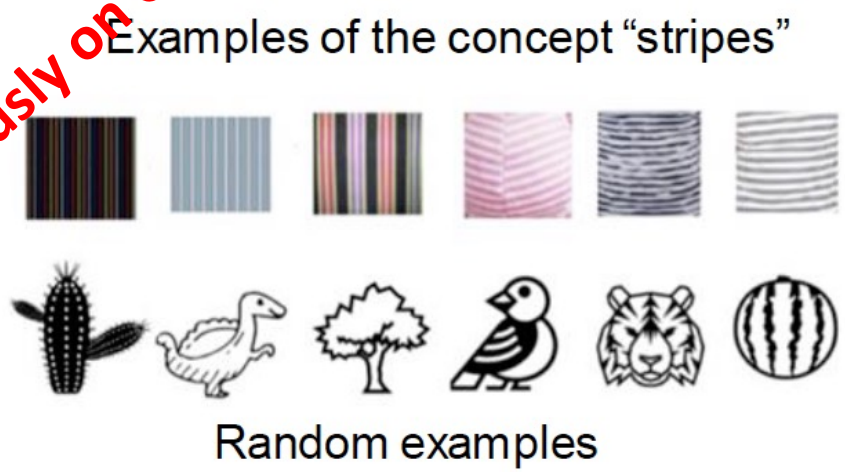
Internal Layers in DNN

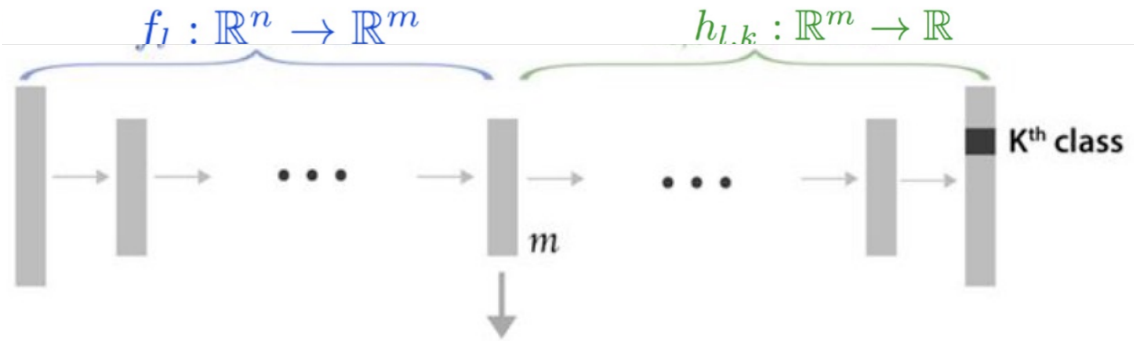


25

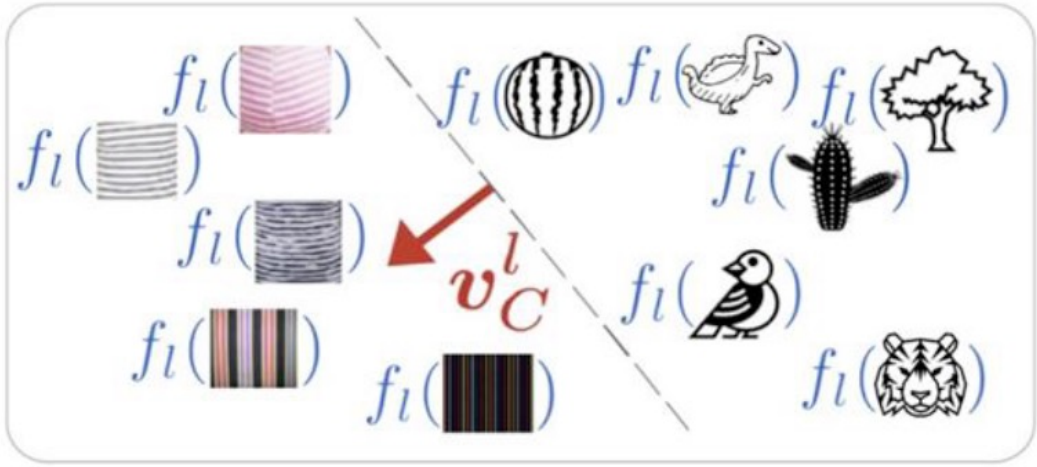
Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Step 2: Compute Concept Activation Vector (CAV)





- Consider representations at layer l of all the positive and negative concept examples
- Train a linear classifier to separate positive from negative
- CAV: Vector orthogonal to the decision boundary



26

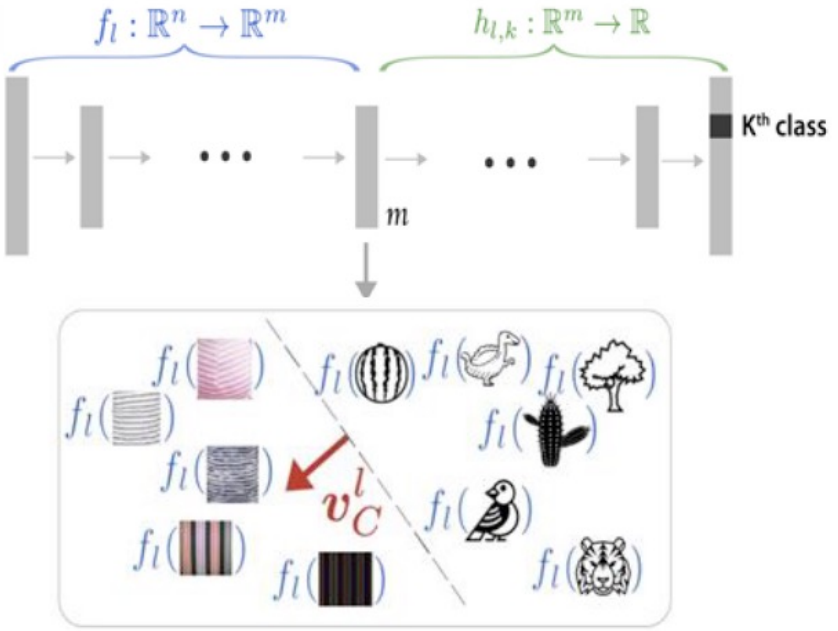
Previously on CENG7880

Step 4: Testing with CAV (TCAV)

- Let X_k be the set of all training inputs with label k

- Goal: Understand how a model f's prediction for class k is sensitive to a given concept C

- TCAV score: Fraction of k-class training inputs whose l-layer activation vector was positively influenced by concept C



$$\text{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}$$

$$\begin{aligned} S_{C,k,l}(\boldsymbol{x}) &= \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(\boldsymbol{x}) + \epsilon \boldsymbol{v}_C^l) - h_{l,k}(f_l(\boldsymbol{x}))}{\epsilon} \\ &= \nabla h_{l,k}(f_l(\boldsymbol{x})) \cdot \boldsymbol{v}_C^l, \end{aligned} \quad (1)$$

28

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Previously on CENG7880

Fish

Dog

z_{train}

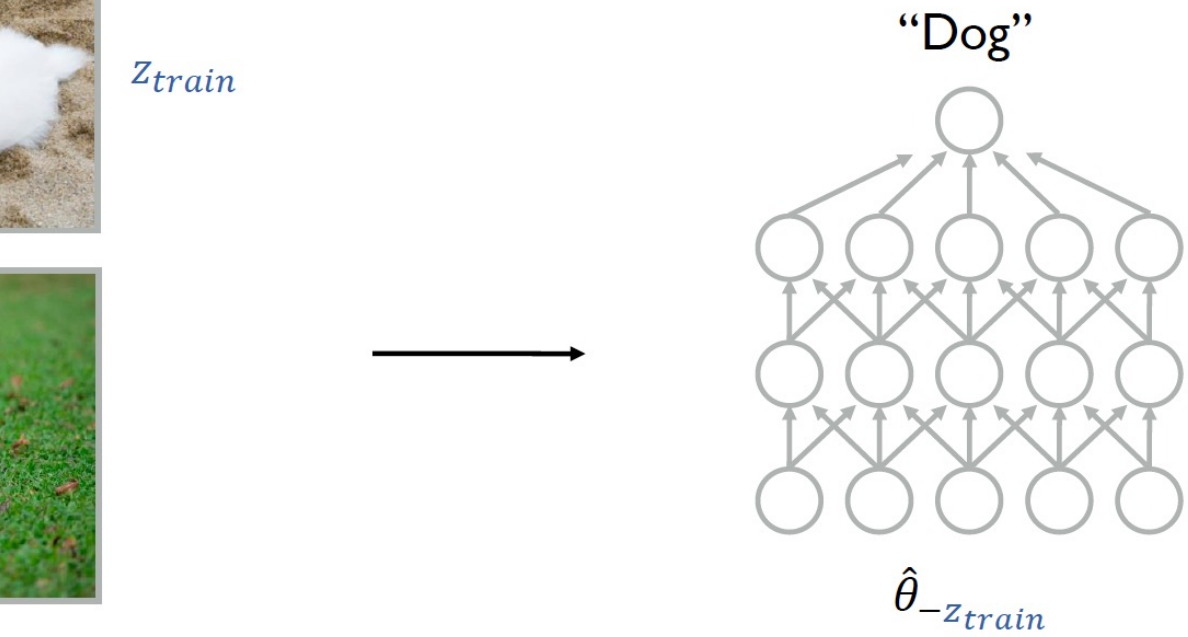
Dog

Training data $z_1, z_2, \ldots, z_n$

Pick $\hat{\theta}$ to minimize $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$

Pick $\hat{\theta}_{-z_{train}}$ to minimize $\frac{1}{n}\sum_{i \neq train}^{n} L(z_i, \theta)$

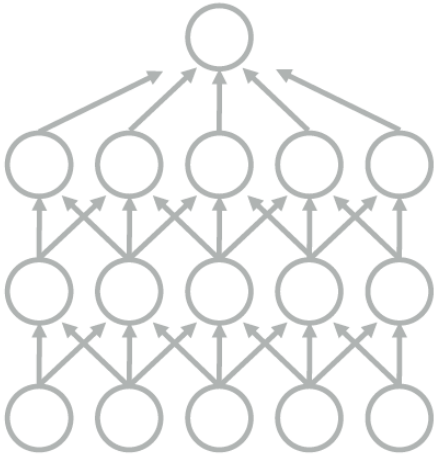
$$= \frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta) - \frac{1}{n} L(z_{train}, \theta)$$



Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Previously on CENG7880

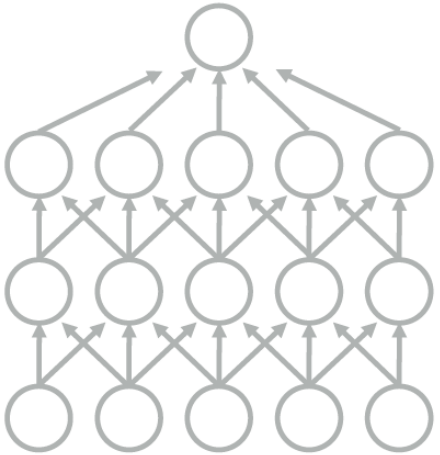
"Dog" (82% confidence)



$\hat{\theta}$

vs.

"Dog" (79% confidence)



$\hat{\theta}_{-z_{train}}$

What is $L(z_{test}, \hat{\theta}_{-z_{train}}) - L(z_{test}, \hat{\theta})$?

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Influence on the parameters

Previously on CENG7880

Calculate the impact of removing sample "z"

$$\frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) \left(-\frac{1}{n}\right) L(z,\theta)$$

Fortunately, influence functions give us an efficient approximation. The idea is to compute the parameter change if z were upweighted by some small ϵ, giving us new parameters $\hat{\theta}_{\epsilon,z} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) + \epsilon L(z,\theta)$. A classic result (Cook & Weisberg, 1982) tells us that the influence of upweighting z on the parameters $\hat{\theta}$ is given by

$$\mathcal{I}_{\text{up,params}}(z) \stackrel{\text{def}}{=} \left.\frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1} \nabla_\theta L(z,\hat{\theta}), \qquad (1)$$

where $H_{\hat{\theta}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 L(z_i,\hat{\theta})$ is the Hessian and is positive definite (PD) by assumption. In essence, we are forming a quadratic approximation to the empirical risk around $\hat{\theta}$ and take a single Newton step; see appendix A for a derivation. Since removing a point z is the same as upweighting it by $\epsilon = -\frac{1}{n}$, we can linearly approximate the parameter change due to removing z without retraining the model by computing $\hat{\theta}_{-z} - \hat{\theta} \approx -\frac{1}{n}\mathcal{I}_{\text{up,params}}(z)$.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning* (pp. 1885-1894). PMLR.

Previously on CENG7880

Influence functions

- $\hat{\theta}_{\epsilon,z_{train}} \stackrel{\text{def}}{=} \text{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) + \epsilon L(z_{train},\theta)$
- Under smoothness assumptions,

$$I_{up,loss}(z_{train}, z_{test}) \stackrel{\text{def}}{=} \left.\frac{dL\left(z_{test}, \hat{\theta}_{\epsilon,z_{train}}\right)}{d\epsilon}\right|_{\epsilon=0}$$

Influence of removing z_{train} on z_{test}

$$= \nabla_\theta L\left(z_{test}, \hat{\theta}\right)^T \left.\frac{d\hat{\theta}_{\epsilon,z_{train}}}{d\epsilon}\right|_{\epsilon=0}$$

$$= -\nabla_\theta L\left(z_{test}, \hat{\theta}\right)^T H_{\hat{\theta}}^{-1} \nabla_\theta L(z_{train}, \hat{\theta})$$

Previously on CENG7880

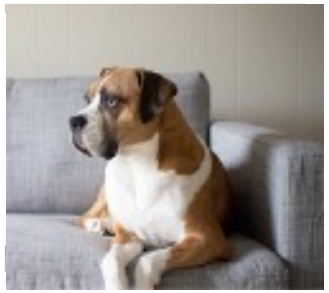
Datamodels: Data-to-Output Modeling

What we are trying to compute (model output function):

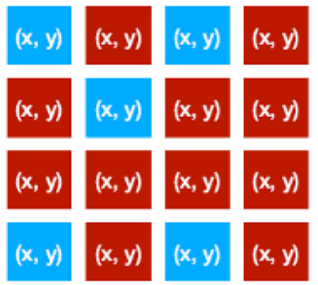
Output of interest on x
(think: margin of correct class)
after training on S'

$$f(x, S') \approx \hat{f}(x, S')$$

Datamodel for x



Specific input x



Subset S' of the training set S

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

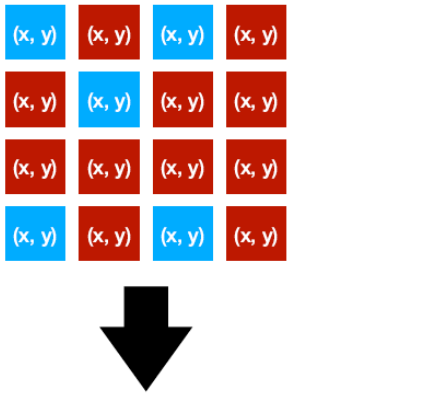
Previously on CENG7880

Model Choice: Linear

$$\hat{f}(x, S') = \theta_x^\top \mathbf{1}_{S'}$$

Learned parameter: vector of weights (one weight per training example in S)

Indicator vector of S'



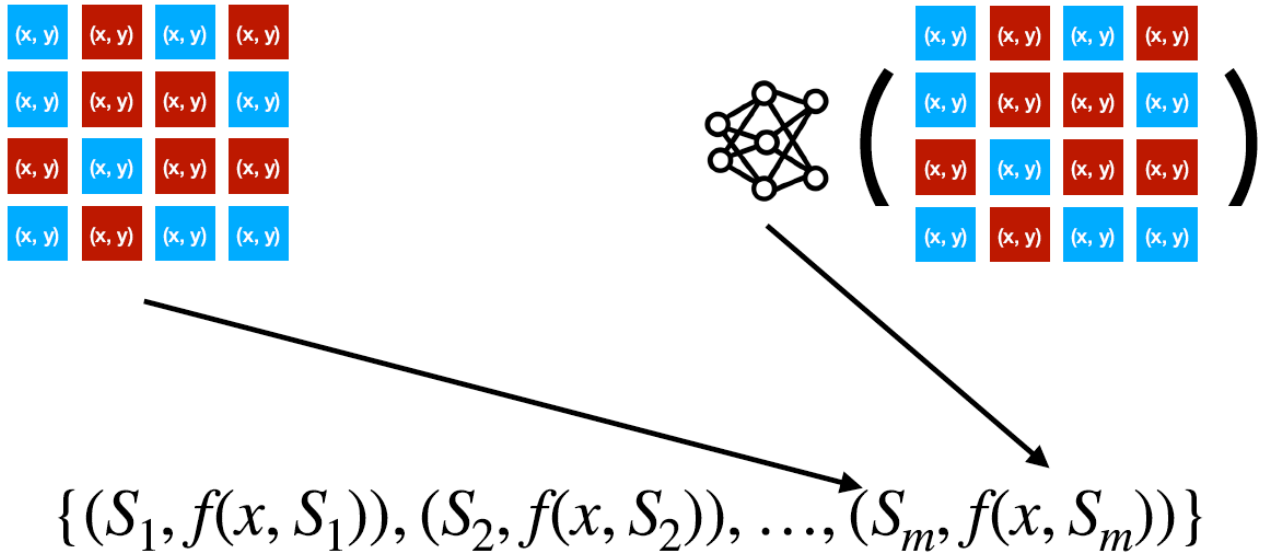
[1 0 1 0 0 1 0 0 0 0 0 0 1 0 1 0]

Remaining question: how do we fit the parameters θ_x?

Previously on CENG7880

How to fit a datamodel

Use supervised learning:



Then: Fit the linear model to this data

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Previously on CENG7880

Fitting a datamodel

(for a **specific** target example x)

$$\{(S_1, f(x, S_1)), (S_2, f(x, S_2)), \ldots, (S_m, f(x, S_m))\}$$

Minimize over all possible weights

Datamodel prediction for margin on target example x after training on S_i, i.e., $g(S_i)$

ℓ_1 regularization (for sparsity + generalization)

$$\theta_x = \min_{w \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \left(w^\top \mathbf{1}_{S_i} - f(x, S_i) \right)^2 + \lambda \|w\|_1$$

Average over all sampled subsets S_i

True (observed) margin from training on S_i and evaluating on x

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Previously on CENG7880

Putting it all together

Constructing datamodels for DNNs trained on CIFAR-10:

Requires training 1000s of models!
Made possible by FFCV (ffcv.io)

→ Repeat **500,000 times**:

→ Choose a random α-fraction of the CIFAR-10 trainset

→ Train a model (ResNet-9) on this subset

→ Measure **correct-class margin** on every test image

→ For each test image, record the pair:

(characteristic vector of the subset, vector of margins)

→ For each test image (10,000 total images):

→ Fit linear model from indicator vectors → margins

Result: **10,000 datamodels**, each parameterized by $\theta_x \in \mathbb{R}^{50,000}$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Previously on CENG7880

Datamodels: Analyzing model brittleness

"boat"
(71% confidence)

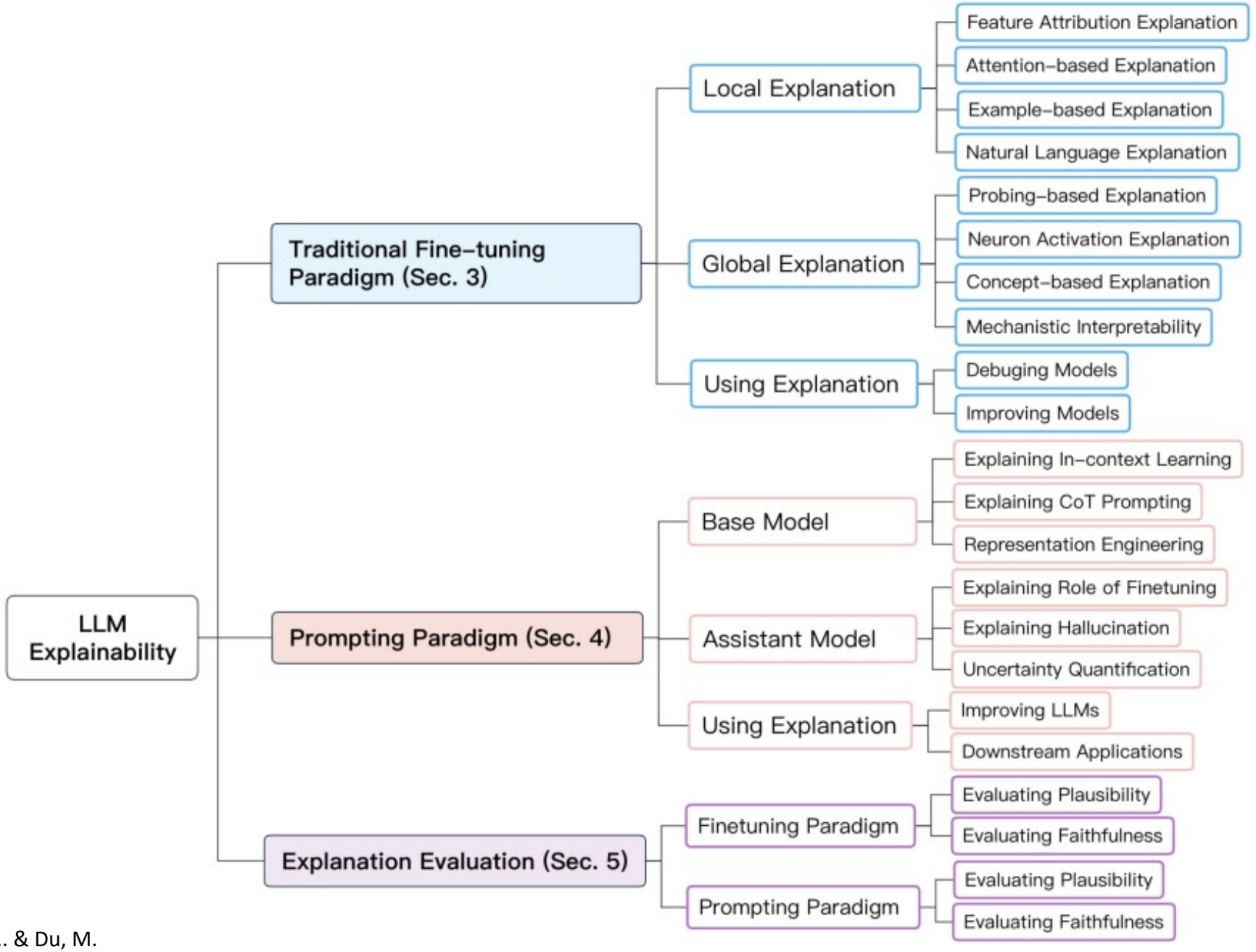
Removing
nine images

"airplane"

~25% of examples misclassified by removing
< 0.2% of training examples

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Previously on CENG7880

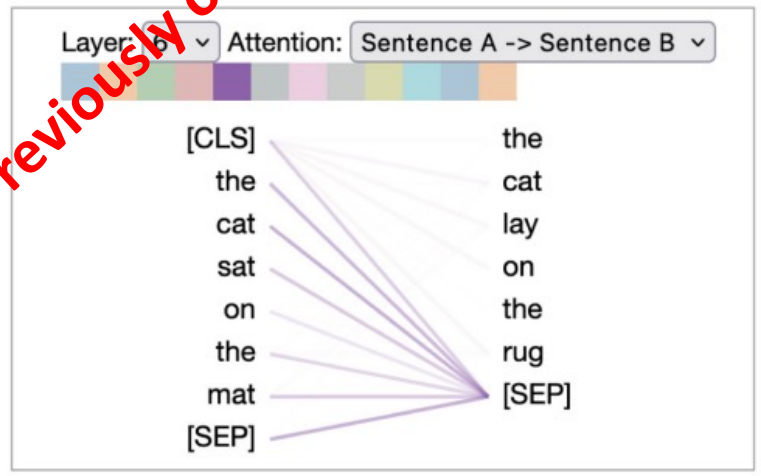


Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1-38.

Traditional Finetuning Paradigm: Local Explanations

Previously on CENG7880

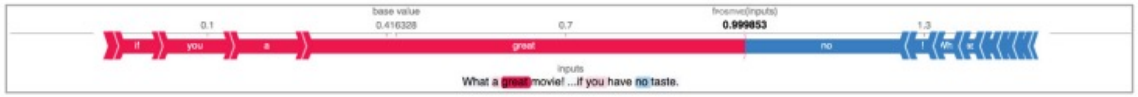
(a) Attention Visualization

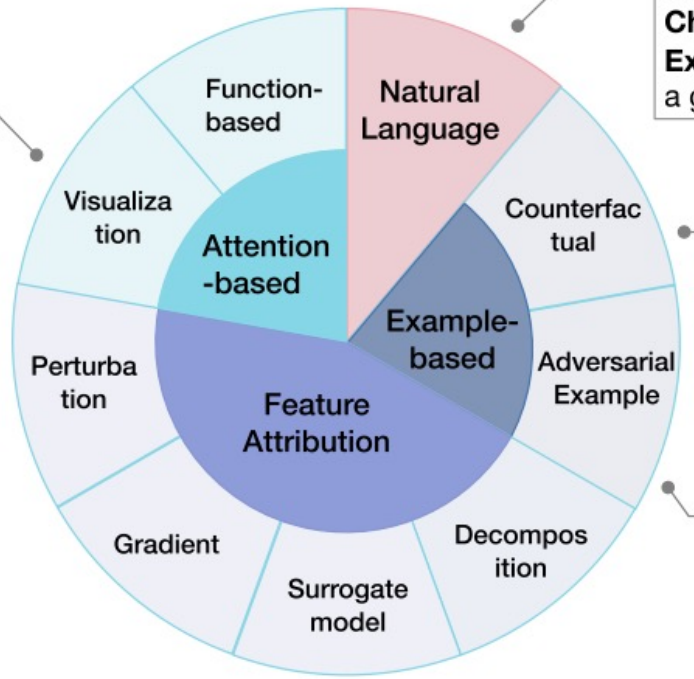


(b) Question Answering

Context: In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his **Colorado Springs experiments**.
Question: What ~~did~~ Tesla spend Astor's money on?
Confidence: 0.78 —> 0.91

(c) Sentiment Analysis





(d) Commonsense Reasoning

Question: While eating a **hamburger with friends**, what are people trying to do?.
Choices: **have fun**, tasty, or indigestion
Explanation: Usually a hamburger with friends indicates a good time.

(e) Sentiment Analysis

Original text: It is great for kids (**positive**).
Negation examples: It is not great for kids (**negative**)

(f) Classification

Original text: The characters, cast in impossibly contrived situations, are totally estranged from reality (**Negative**).
Perturbed text: The characters, cast in impossibly engineered circumstances, are fully estranged from reality (**Positive**)

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1-38.

Previously on CENG7880

Traditional Finetuning Paradigm: Global Explanations

- Probing-based Explanations
- Neuron-Activation Explanations
- Concept-based Explanations
- Mechanistic Interpretability

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1-38.

Previously on CENG7880

Prompting Paradigm: Base-Model Explanation

- Explaining In-Context Learning
- Explaining Chain-of-Thought Prompting
- Representation Engineering

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1-38.

Previously on CENG7880

Prompting Paradigm: Assistant-Model Explanation

- Explaining the Role of Fine-tuning
- Explaining Hallucination
- Uncertainty Quantification

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1-38.

Agenda

- Fairness
  - Notions, Definitions and Measures
  - Fairness Algorithms

Administrative Notes

- Final Exam:
  - 13 January 16:30
- Paper selection finalized except for two projects
- Project milestones
  - **1. Milestone (November 23, midnight):**
    - Read & understand the paper
    - Download the datasets
    - Prepare the Readme file excluding the results & conclusion
  - **2. Milestone (December 7, midnight)**
    - The results of the first experiment
  - **3. Milestone (January 4, midnight)**
    - Final report (Readme file)
    - Repo with all code & trained models

Fairness

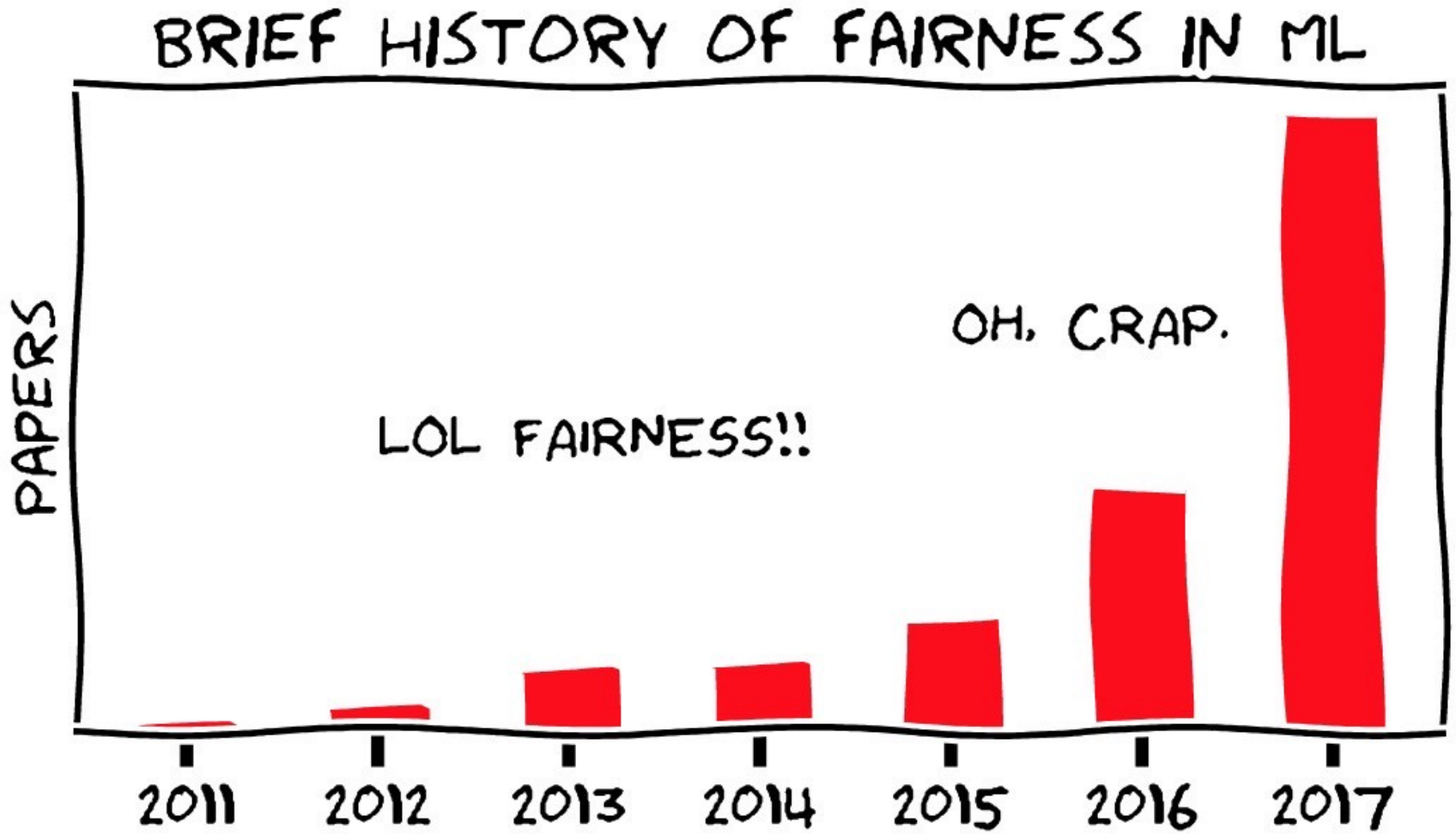
Bias and Fairness

- Oxford Dictionary:

  “[…] inclination or prejudice for or against one person or group, especially in a way considered to be unfair."

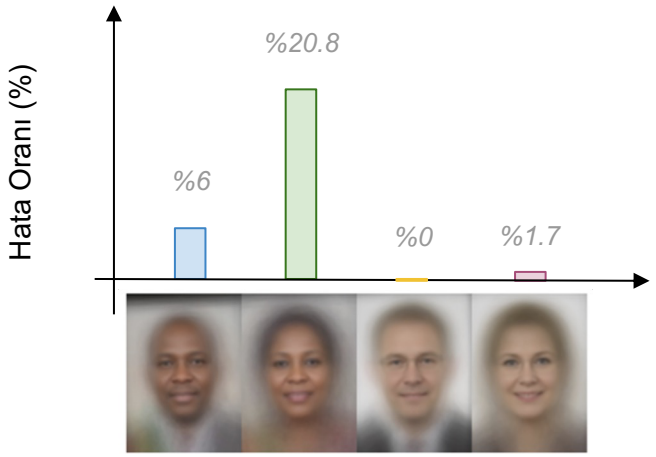
- TDK:

  “Bir kimse veya bir şeyle ilgili olarak belirli şart, olay ve görüntülere dayanarak önceden edinilmiş olumlu veya olumsuz yargı, peşin yargı, peşin hüküm, peşin fikir.”



https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

Bias and Fairness in ML:
Examples: Face Recognition

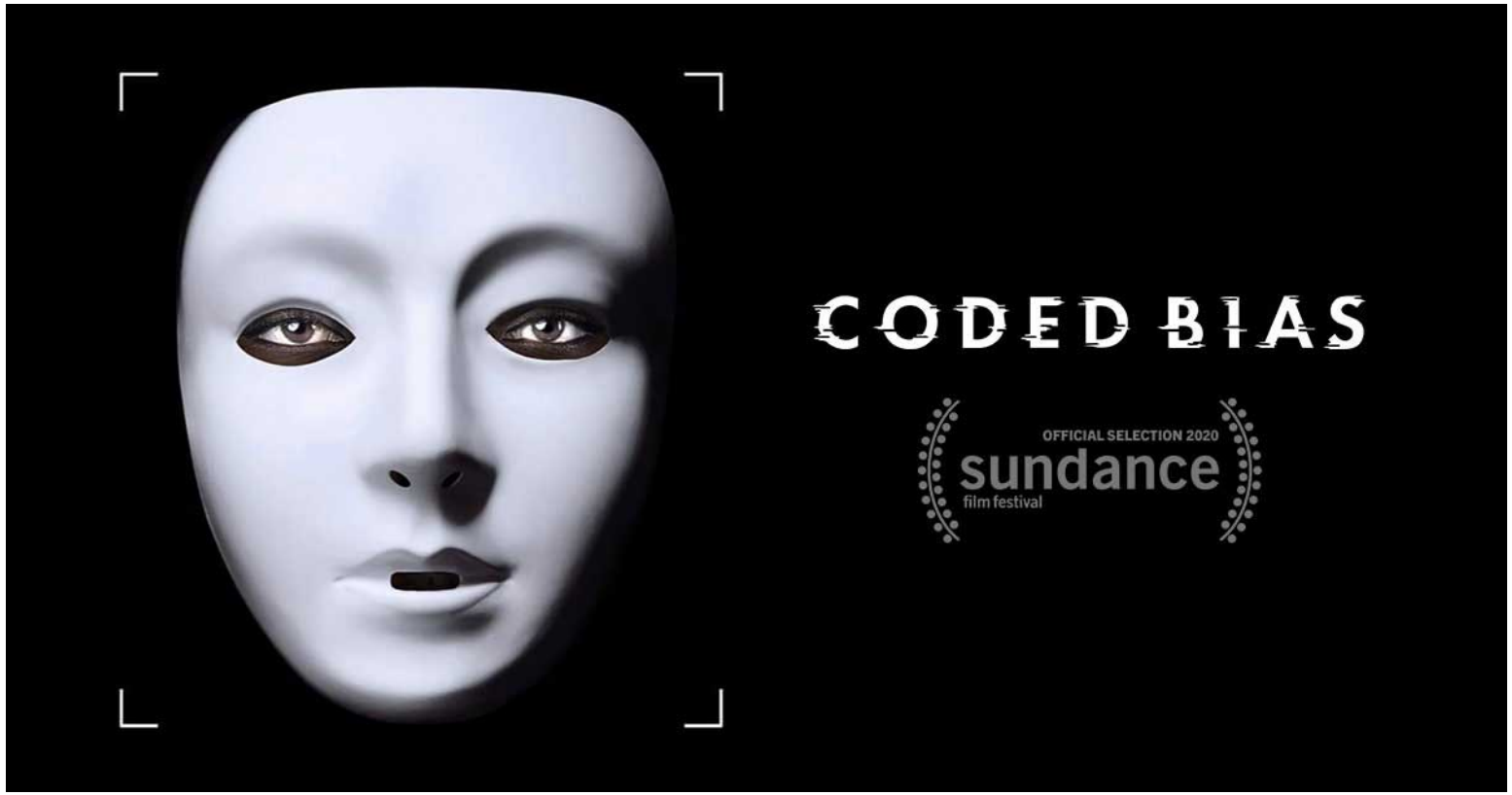


Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

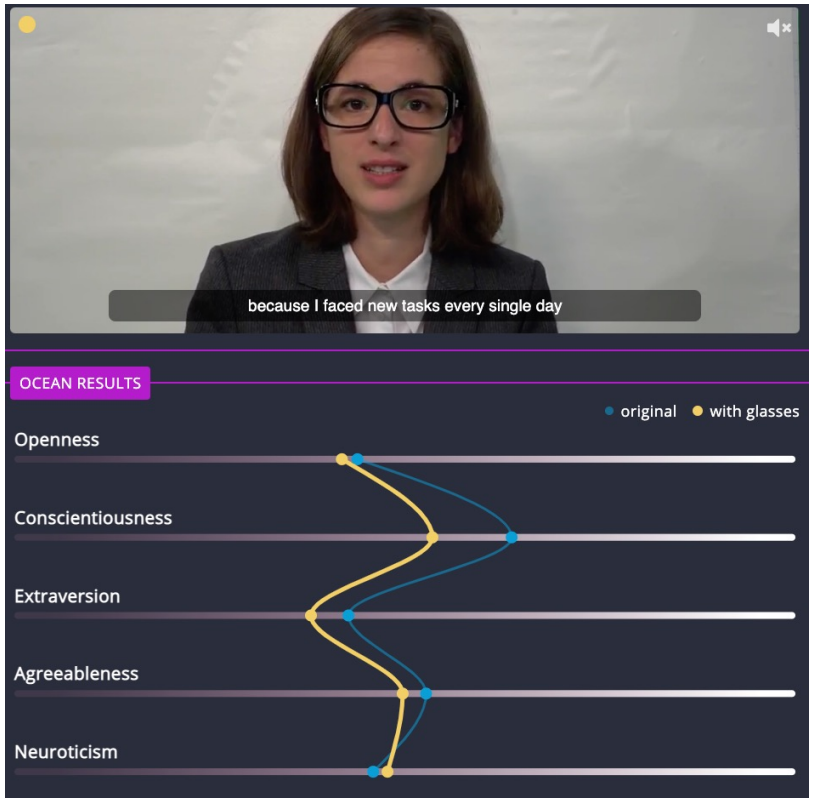
Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77-91 (2018).

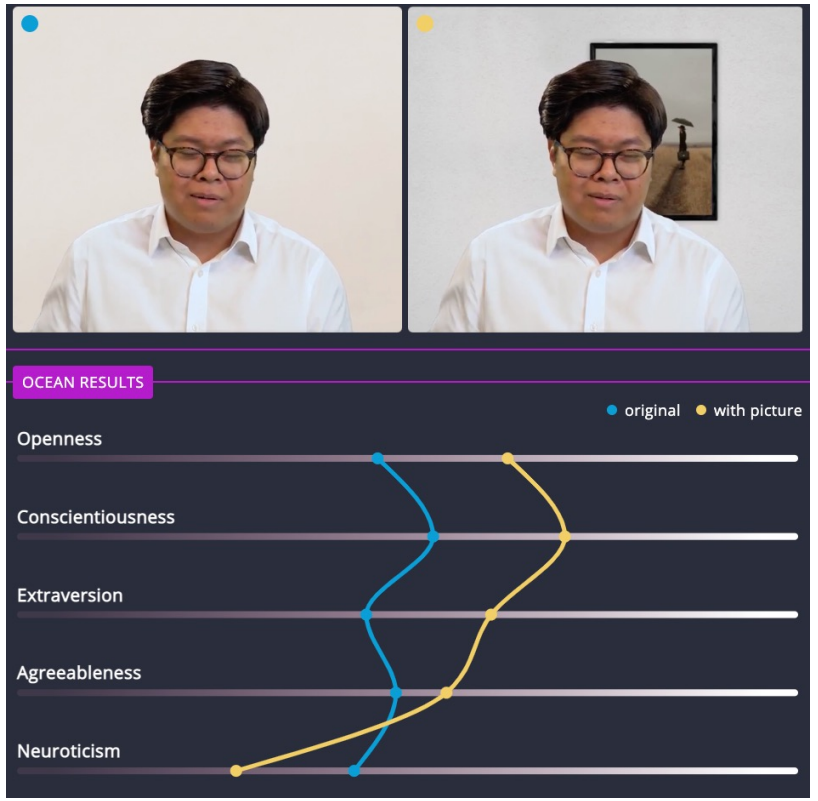
Bias and Fairness in ML: Examples

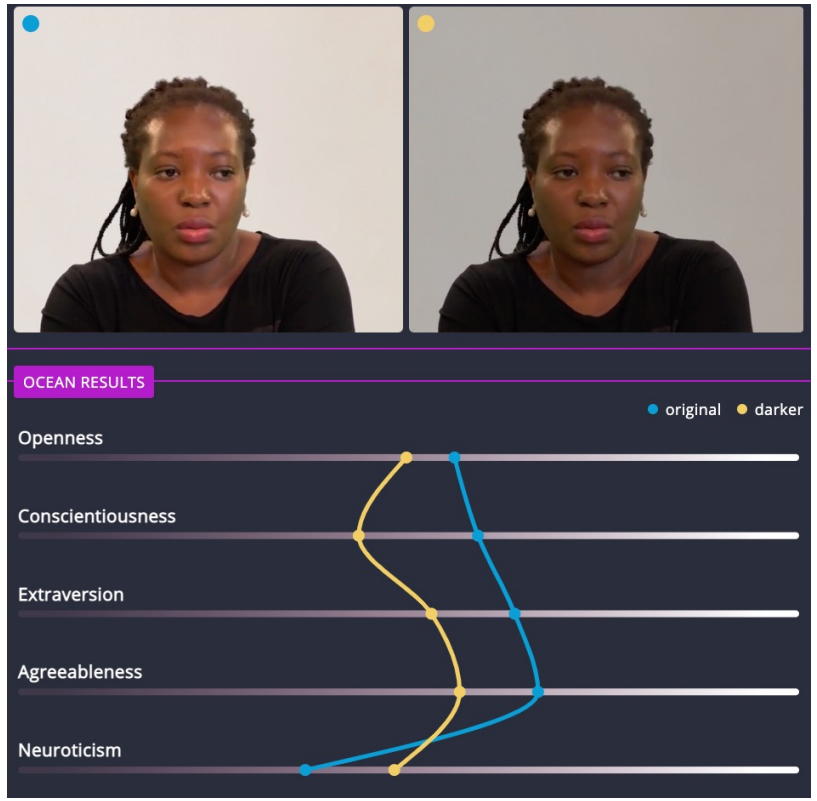
https://www.imdb.com/title/tt11394170/



Bias and Fairness in ML:
Examples: Personality Identification







https://web.br.de/interaktiv/ki-bewerbung/en/

Bias and Fairness in ML: Examples: Emotion Recognition

Happy Anger Disgust Neutral

Fear Anger Surprise Sad

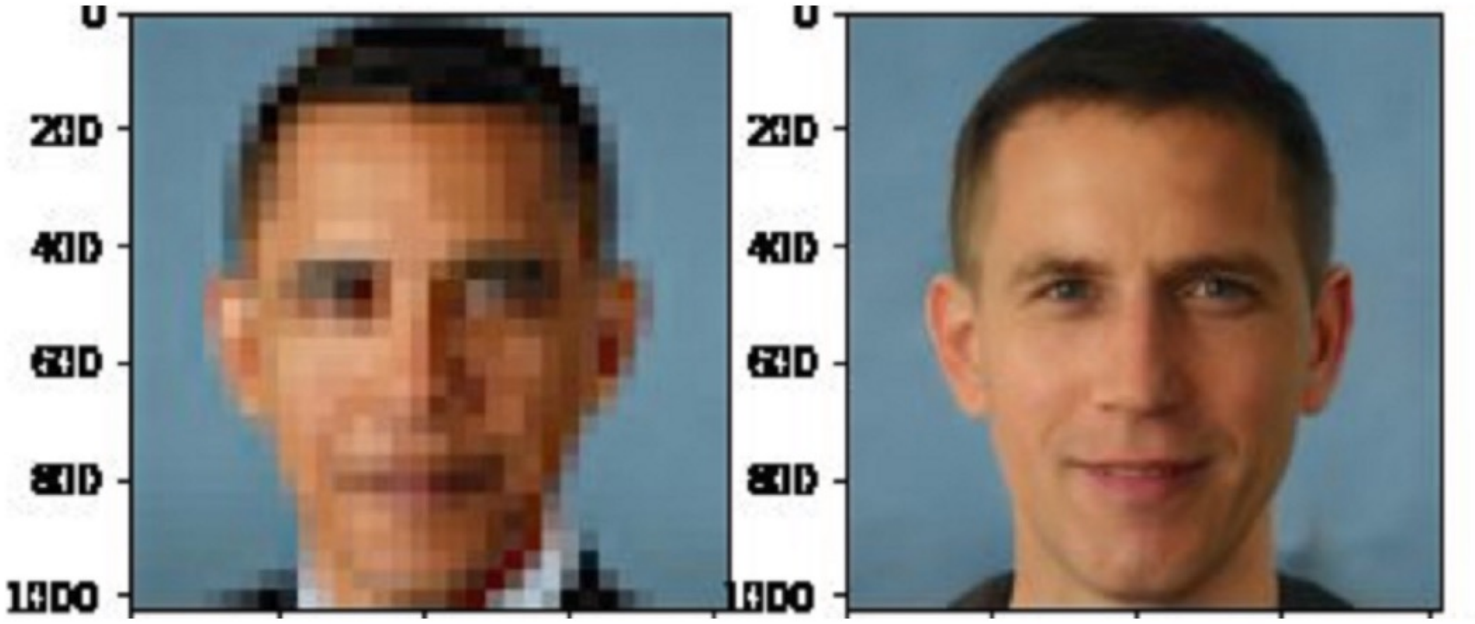
Table 3: Mean class-wise accuracy of the models, broken down by attribute labels on RAF-DB (Cau: Caucasian, AA: African-American, M: Male, F: Female).

	Without Augmentation			With Augmentation		
	Baseline	Attri-aware	Disentangle	Baseline	Attri-aware	Disentangle
Male	65.3%	67.4%	62.5%	72.3%	73.7%	**74.2%**
Female	63.5%	64.9%	61.0%	74.1%	74.1%	**74.4%**
Cau	65.9%	68.3%	63.4%	74.7%	74.9%	**75.6%**
AA	68.1%	62.8%	58.4%	76.3%	76.3%	**76.6%**
Asian	60.0%	59.8%	54.4%	67.8%	69.9%	**70.4%**
0-3	63.6%	59.9%	56.7%	**80.2%**	71.9%	65.0%
4-19	59.5%	58.8%	57.0%	61.1%	63.7%	**69.9%**
20-39	65.9%	68.2%	62.9%	74.9%	75.8%	**76.4%**
40-69	65.0%	63.4%	60.1%	73.8%	**74.4%**	72.1%
70+	51.3%	53.6%	51.6%	60.8%	54.3%	**62.2%**
M-Cau	65.3%	69.3%	63.6%	73.3%	73.9%	**74.5%**
M-AA	77.0%	70.4%	63.2%	66.4%	**80.2%**	78.7%
M-Asian	61.2%	58.6%	56.2%	67.8%	68.4%	**70.2%**
F-Cau	64.1%	66.2%	62.2%	74.7%	74.9%	**75.5%**
F-AA	61.6%	57.9%	62.8%	**87.6%**	75.8%	74.6%
F-Asian	59.1%	59.5%	52.4%	65.6%	68.4%	**69.0%**

T. Xu, J. White, S. Kalkan, H. Gunes, "Investigating Bias and Fairness in Facial Expression Recognition", ECCV2020 Workshop: ChaLearn Looking at People workshop ECCV: Fair Face Recognition and Analysis, 2020.

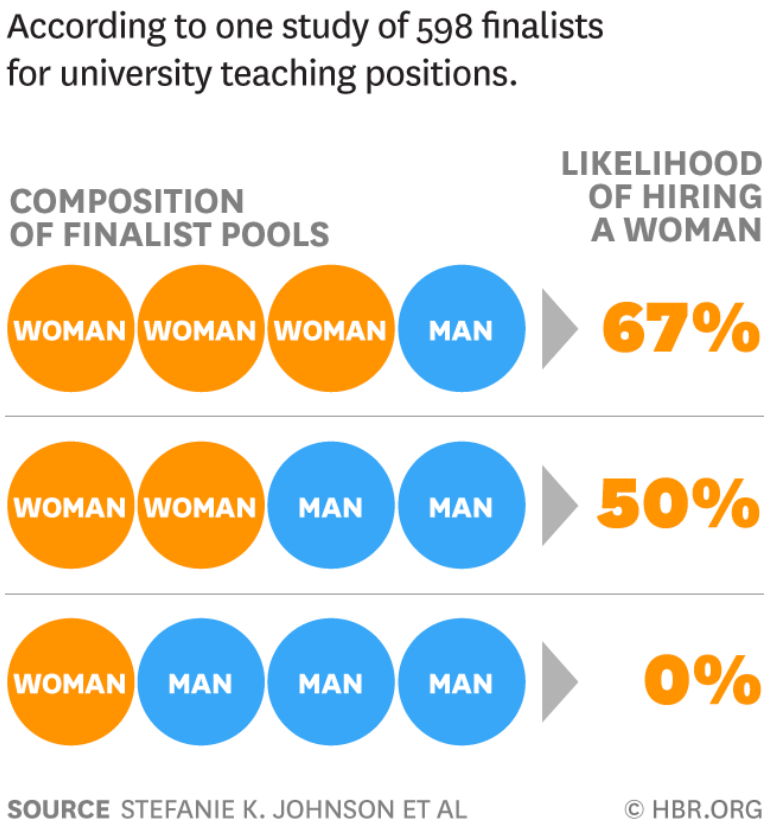
Bias in Machine Learning

- ML models may be biased against minorities



Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Bias and Fairness in ML:
Examples: Hiring Personnel

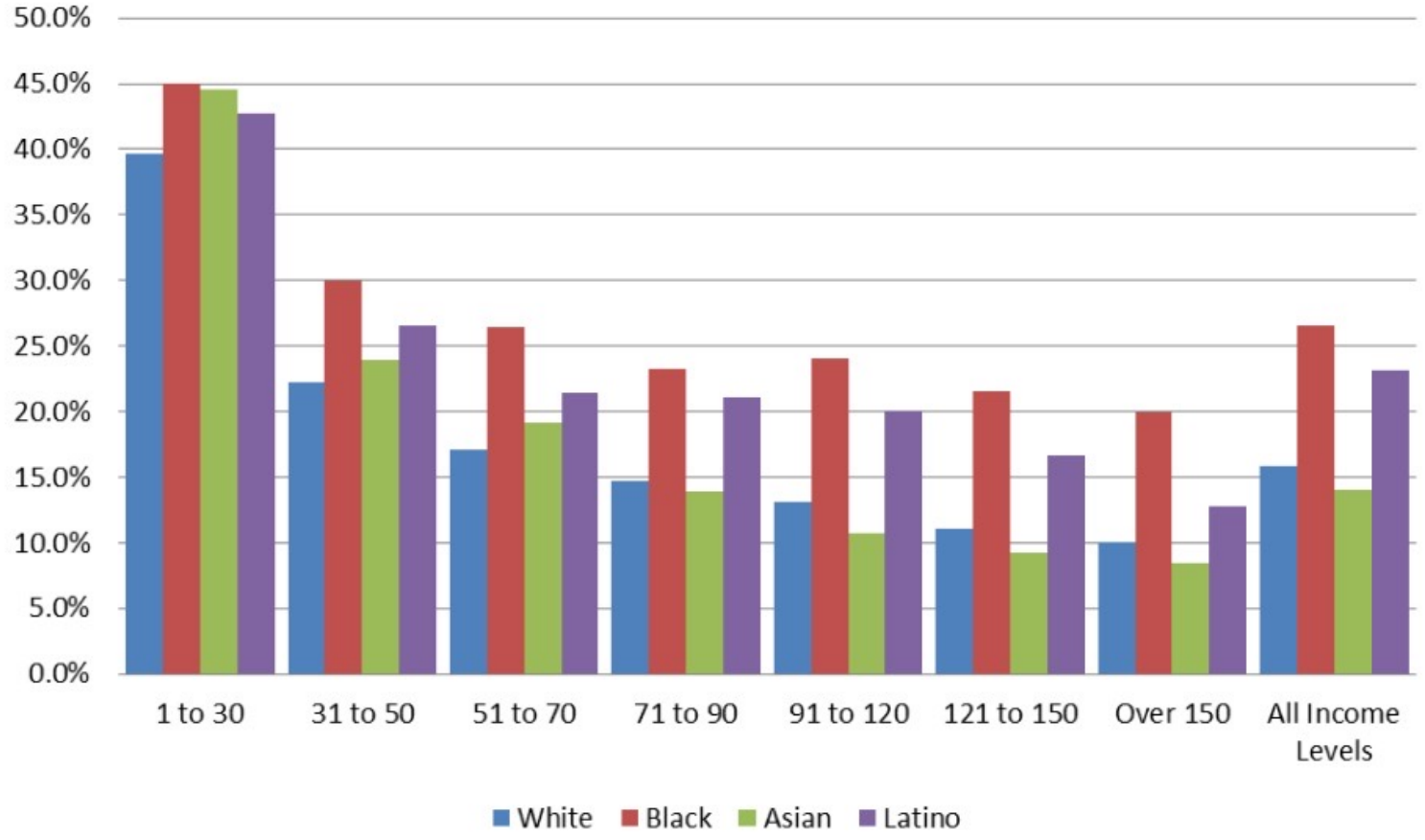


For other examples:
https://harver.com/blog/hiring-biases/

Source: https://hbr.org/2016/04/if-theres-only-one-woman-in-your-candidate-pool-theres-statistically-no-chance-shell-be-hired

Bias and Fairness in ML:
Examples: Bank Credit Applications

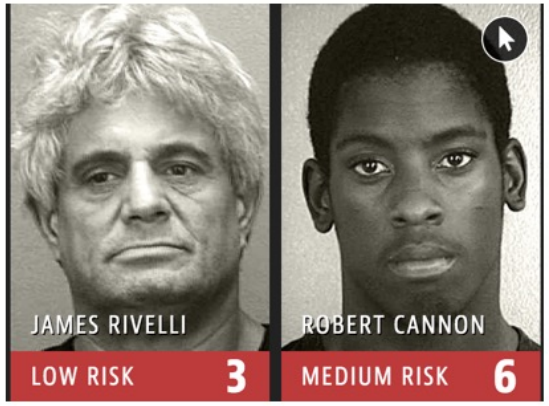
Figure 6. Denial Rates by Income Level and Race: All Loans, 2013 and 2014



Source: 2013-2014 HMDA. Data compiled by the Federal Reserve Bank of Boston

Source: http://www.bostonfed.org/commdev/issue-briefs/2016/cdbrief22016.pdf

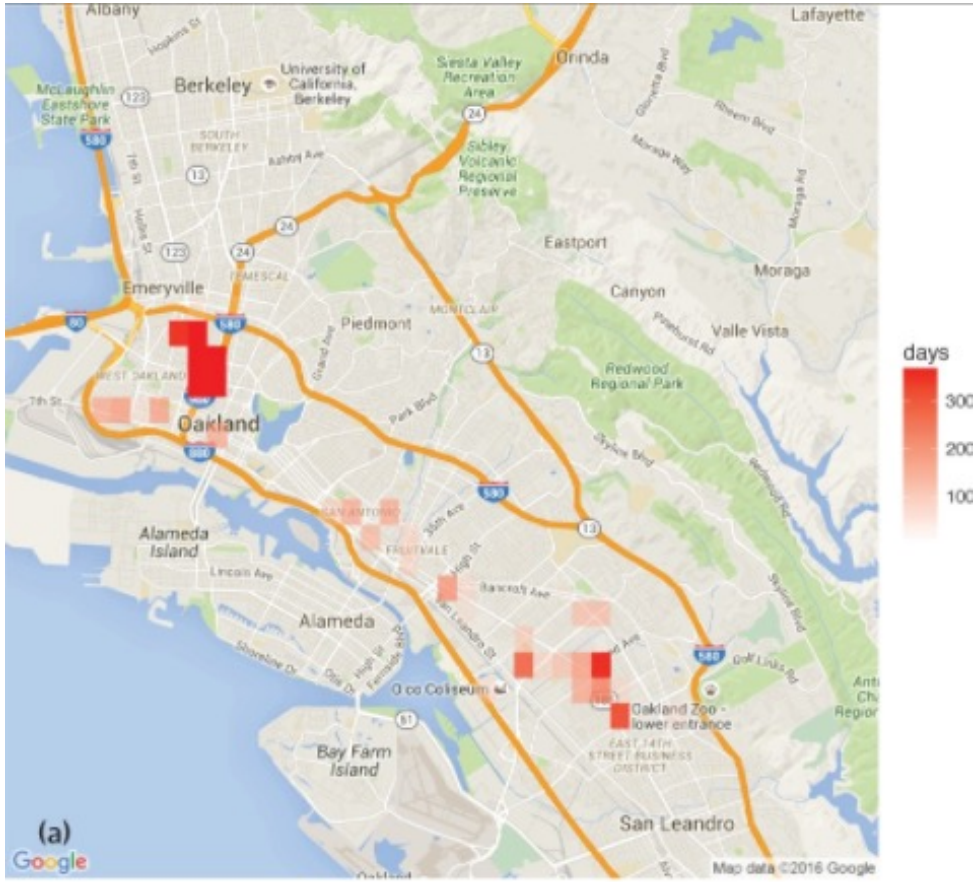
Bias and Fairness in ML:
Examples: Criminal Assessment

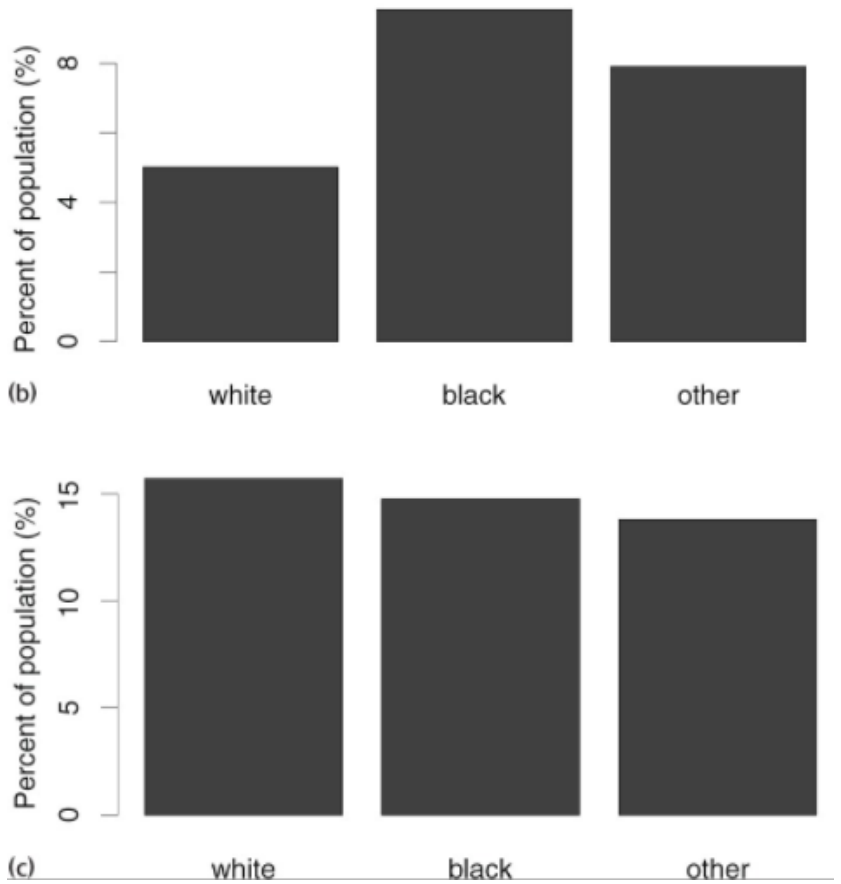


Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016) Machine Bias.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bias and Fairness in ML:
Examples: Criminal Assessment





(a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race

https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x

Bias in Machine Learning

- ML models may be biased against minorities

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Bolukbasi et al. 2016 : https://arxiv.org/abs/1607.06520
Image from: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

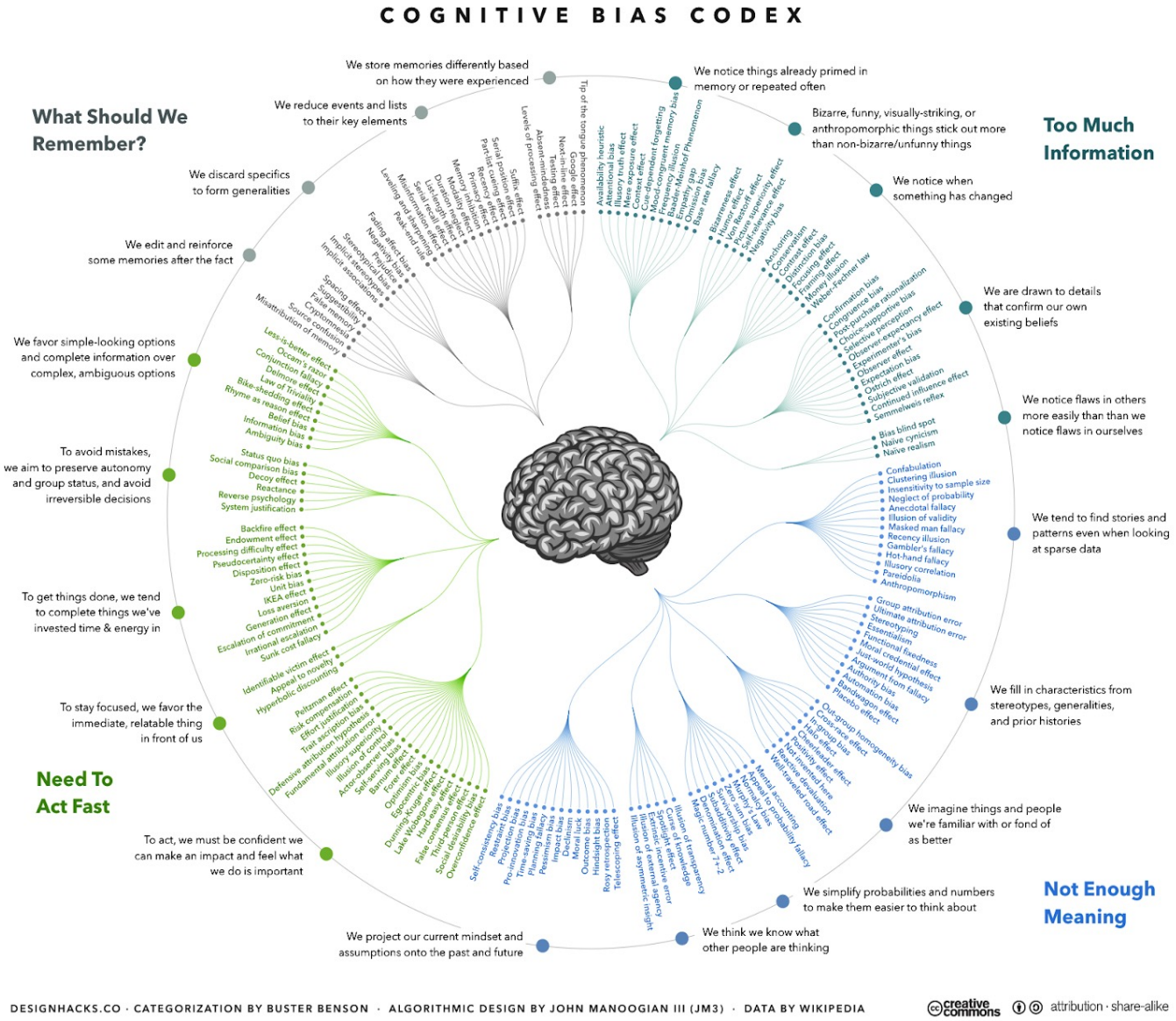
Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Bias and Fairness in ML:
A Fundamental Problem

- Current AI systems heavily use machine learning based solutions
- Any widespread use of such systems necessitates these systems to be fair across different demographic groups
- Who are responsible?
  - Decision makers
  - Researchers
  - Engineers
  - Everyone

Bias and Fairness in ML: Sources of Bias

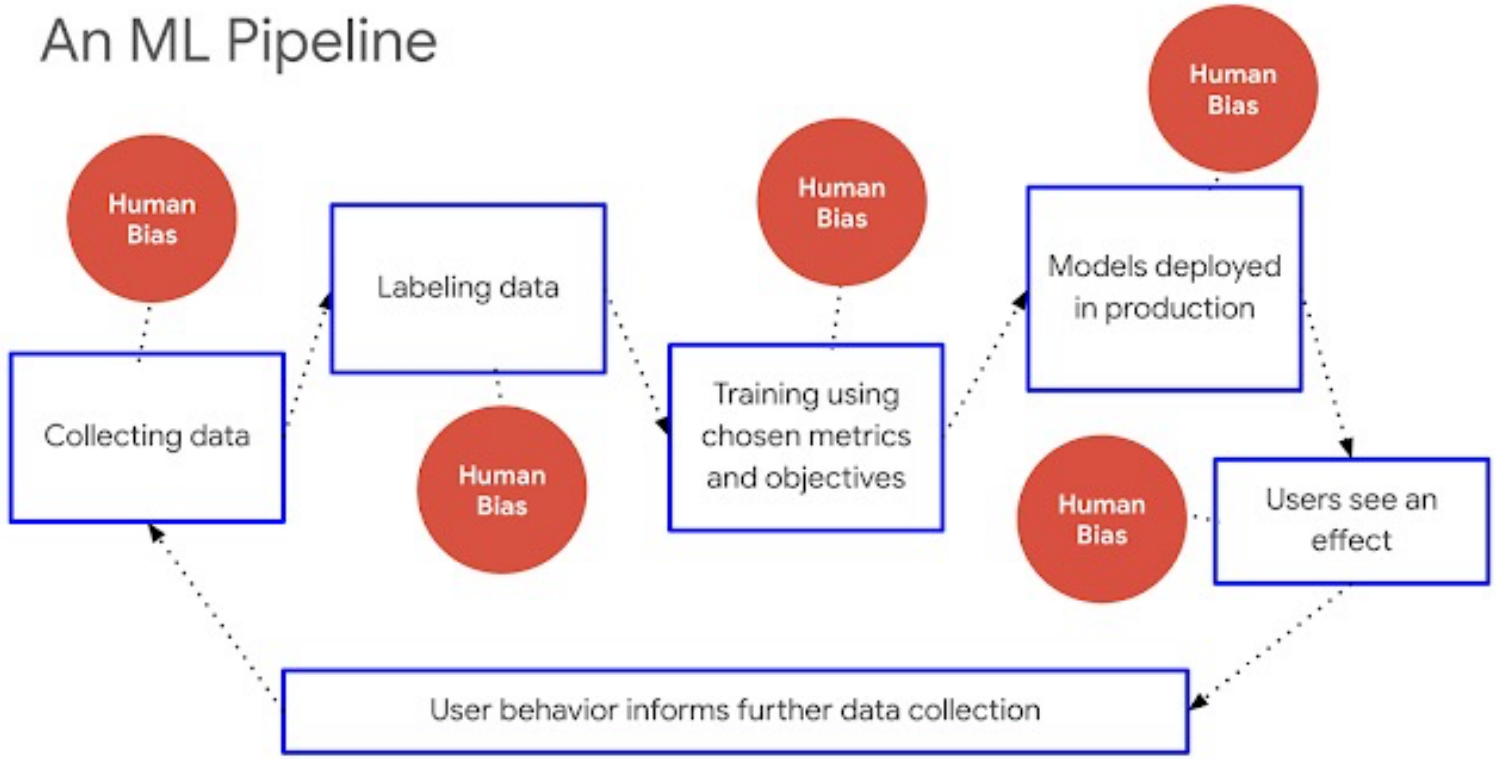
Main Source of Bias



https://miro.medium.com/max/3200/0*R1AFqFSih42NRTMl

Bias and Fairness in ML: Sources of Bias

Main Source of Bias



https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html

Sources of Bias

- **Data representation:** Distribution of inputs $p(x)$
- **Tainted labels:** Distribution of label assignments $p(y \mid x)$
- **Sensitive features:** Selecting what features to include for each sample (e.g., whether to include sensitive attributes such as race and gender)

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Data Representation

- Less data from minority groups → Higher error on minority groups
- **Example:** Many clinical trials historically recruited largely white males, leading to biases in understanding outcomes and side effects
- **Example:** Focus on easily accessible data (e.g. recent tweets, or easily measured features of people) can lead to biased datasets
- Need to be careful to gather representative datasets

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Does balanced data guarantee fairness?

- No!
- On the D-Vlog dataset (for depression detection), females have more samples have more samples but suffer from biased predictions!

Table 3: Dataset distribution and target attribute breakdown across datasets. Abbreviations: F: Female. M: Male. T: Total. Y_0: Control group. Y_1: MHD group. NA: Not available.

	Depresjon			Psykose			D-Vlog		
	Y_0	Y_1	T	Y_0	Y_1	T	Y_0	Y_1	T
M	150	160	310	NA	246	NA	140	182	322
F	252	131	383	NA	39	NA	266	373	639
T	402	291	693	402	285	687	405	555	961

	D-Vlog						
Metrics	B	Pre (M)	Pre (F)	In (M)	In (F)	Post (M)	Post (F)
$\mathcal{M}_{Acc}$	0.64	0.66	0.65	0.65	0.65	0.64	0.64
$\mathcal{M}_{P}$	0.70	0.71	0.69	0.72	0.69	0.68	0.67
$\mathcal{M}_{R}$	0.69	0.71	0.73	0.65	0.73	0.71	0.73
$\mathcal{M}_{F1}$	0.69	0.70	0.71	0.68	0.71	0.69	0.70
$\mathcal{M}_{SP}$	0.92	1.02	1.01	0.95	1.00	**1.24**	0.92
$\mathcal{M}_{Opp}$	1.09	**1.25**	**1.24**	1.18	1.19	**1.38**	1.16
$\mathcal{M}_{Odd}$	**1.84**	**2.13**	**2.09**	**2.45**	**1.87**	**1.42**	**2.27**
$\mathcal{M}_{EAcc}$	1.09	1.19	**1.21**	1.10	1.15	1.14	**1.21**
Consistency	3/4	2/4	1/4	3/4	2/4	1/4	2/4

Larger-than-1 indicates bias for females

Cheong, J., Kuzucu, S., Kalkan, S., & Gunes, H. (2023). Towards gender fairness for mental health prediction. International Joint Conferences on Artificial Intelligence Organization.

Does balanced data guarantee fairness?

- No!
- On the D-Vlog dataset (for depression detection), females have more samples have more samples but suffer from biased predictions!
- Why?
  - Videos/samples are truncated, which affects female samples more.

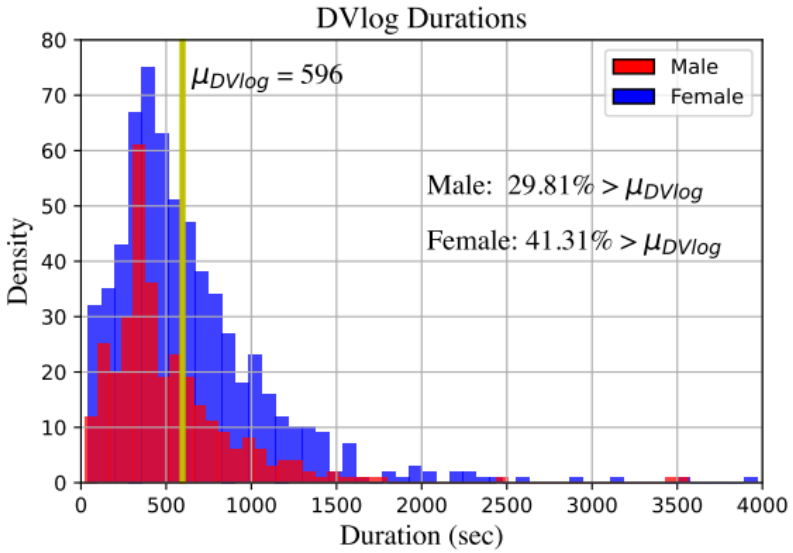


Figure 1: Male (red) and female (blue) vlog duration distribution in DVlog dataset. Yellow vertical line indicates the mean duration point. For females and males, 41.31% and 29.81% of the vlogs longer than the mean were truncated respectively.

Cheong, J., Kuzucu, S., Kalkan, S., & Gunes, H. (2023). Towards gender fairness for mental health prediction. International Joint Conferences on Artificial Intelligence Organization.

Does balanced data guarantee fairness?

- No!
- On the D-Vlog dataset (for depression detection), females have more samples have more samples but suffer from biased predictions!
- Why?
  - Videos/samples are truncated, which affects female samples more.
  - Gender differences in depression manifestation

Factor 2: Gender Differences in Depression Manifestation and Diagnosis. Females and males tend to show different symptom profiles when depressed [Floyd, 1997; Barsky *et al.*, 2001; Ogrodniczuk and Oliffe, 2011]. Though existing literature does not provide a conclusive indication of whether males or females are harder to diagnose, existing literature suggests that there are factors (e.g. physician bias, hormonal effects) which may make it more difficult to diagnose depression in females compared to males [Floyd, 1997; Barsky *et al.*, 2001] (F3.2).

Cheong, J., Kuzucu, S., Kalkan, S., & Gunes, H. (2023). Towards gender fairness for mental health prediction. International Joint Conferences on Artificial Intelligence Organization.

Sensitive Attributes as Features

- When should sensitive attributes be used as features?

- **Example:** Predicting diabetes risk
  - Race is a sensitive attribute that may not cause diabetes, but may be correlated with unrecorded features that cause diabetes
  - What if an insurance company decides that people of some races are at higher risk and should pay higher premium?

- Omitting sensitive attributes is not enough!
  - Other features such as current income may be correlated with race/gender

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Sources of Bias

- Need to gather representative sample
- Need to ensure labels are unbiased
- Need to think carefully about whether to include sensitive attributes

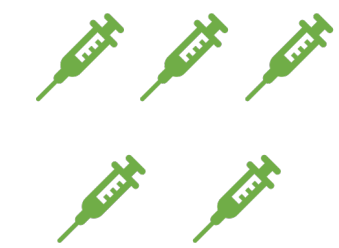
Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Fairness Notions

What does it mean to be fair?

In Philosophy:

- **Egalitarianism** (social equality, equality of opportunity) and **distributive justice** focus on the process of fair distribution of resources.
- **Utilitarianism, consequentialism** focus on the outcomes of the the processes on overall social welfare.



5 doses of medicine

Egalitarianism: Every individual deserves equal change at the resource.

Utilitarianism: Prioritize outcome that maximizes overall well-being.

What does it mean to be fair?

- Two recurring main principles:
  - **Equality/liberty**: Equal access to basic rights
  - **Difference**: Inequalities do not harm disadvantaged groups
- These principles are rather abstract and do not answer how differences among individuals can be considered (equity) on what values (power, need, responsibility):
  - Rawls, J., 2017. A theory of justice. In Applied ethics (pp. 21-29). Routledge.
  - Forsyth, Donelson R. 2006. "Conflict." pp. 388–389 in Group Dynamics (5th ed.), by D. R. Forsyth. Belmont, CA: Wadsworth Cengage Learning.
  - Deutsch, M. 1975. "Equity, equality, and need: What determines which value will be used as the basis of distributive justice?." Journal of Social Issues 31:137–149.
- This has led to many definitions of fairness in sociology, law, economics and politics.

Challenges with defining fairness

Optimizing a statistical fairness measure might violate fairness definitions in Philosophy

Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There?

Matthias Kuppler[1], Christoph Kern[1], Ruben L. Bach[1], and Frauke Kreuter[2,3]

[1] *School of Social Sciences, University of Mannheim, Germany*
[2] *Department of Statistics, LMU Munich, Germany*
[3] *Joint Program in Survey Methodology, University of Maryland, USA*

The advent of powerful prediction algorithms led to increased automation of high-stake decisions regarding the allocation of scarce resources such as government spending and welfare support. This automation bears the risk of perpetuating unwanted discrimination against vulnerable and historically disadvantaged groups. Research on algorithmic discrimination in computer science and other disciplines developed a plethora of fairness metrics to detect and correct discriminatory algorithms. Drawing on robust sociological and philosophical discourse on distributive justice, we identify the limitations and problematic implications of prominent fairness metrics. We show that metrics implementing equality of opportunity only apply when resource allocations are based on deservingness, but fail when allocations should reflect concerns about egalitarianism, sufficiency, and priority. We argue that by cleanly distinguishing between prediction tasks and decision tasks, research on fair machine learning could take better advantage of the rich literature on distributive justice.

Challenges with defining fairness

Optimizing a statistical fairness measure might lead to overall lower social welfare

- See also: Hu, L., & Chen, Y. (2020, January). Fair classification and social welfare. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 535-545).

Learning to Be Fair: A Consequentialist Approach to Equitable Decision Making

Alex Chohlas-Wood, Madison Coots, Henry Zhu, Emma Brunskill, Sharad Goel

Published Online: 18 Dec 2024 | https://doi.org/10.1287/mnsc.2022.00345

Abstract

In an attempt to make algorithms *fair*, the machine learning literature has largely focused on equalizing decisions, outcomes, or error rates across race or gender groups. To illustrate, consider a hypothetical government rideshare program that provides transportation assistance to low-income people with upcoming court dates. Following this literature, one might allocate rides to those with the highest estimated treatment effect per dollar while constraining spending to be equal across race groups. That approach, however, ignores the downstream consequences of such constraints and, as a result, can induce unexpected harm. For instance, if one demographic group lives farther from court, enforcing equal spending would necessarily mean fewer total rides provided and potentially more people penalized for missing court. Here we present an alternative framework for designing equitable algorithms that foregrounds the consequences of decisions. In our approach, one first elicits stakeholder preferences over the space of possible decisions and the resulting outcomes—such as preferences for balancing spending parity against court appearance rates. We then optimize over the space of decision policies, making trade-offs in a way that maximizes the elicited utility. To do so, we develop an algorithm for efficiently learning these optimal policies from data for a large family of expressive utility functions. In particular, we use a contextual bandit algorithm to explore the space of policies while solving a convex optimization problem at each step to estimate the best policy based on the available information. This consequentialist paradigm facilitates a more holistic approach to equitable decision making.

Challenges with defining fairness

- Impossibility theorem [1, 2]:

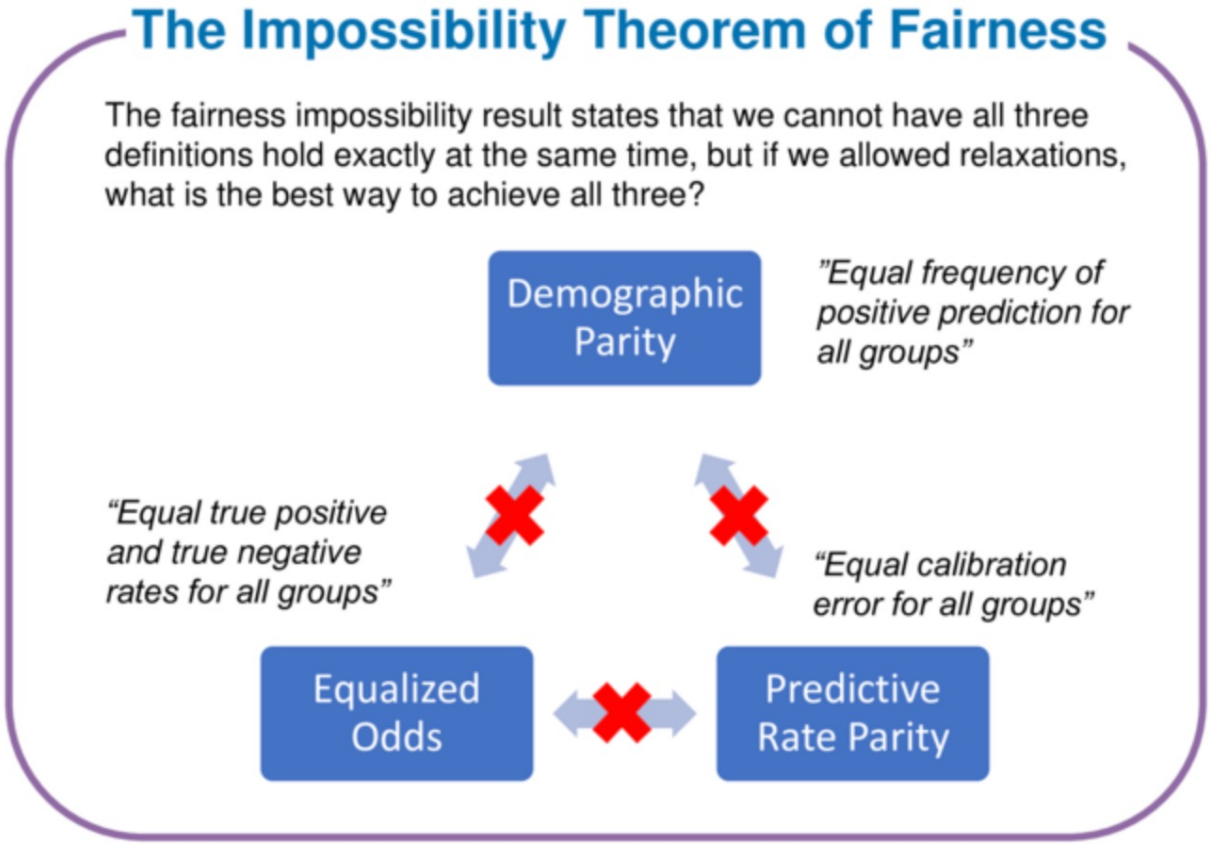


Fig from: https://neurips.cc/virtual/2022/poster/52996

[1] Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807, 2016.
[2] Miconi. The impossibility of "fairness": a generalized impossibility result for decisions. arXiv:1707.01195, 2017.

Challenges with defining fairness

- Too many statistical fairness definitions:
  - Individual fairness, group fairness, causal fairness, intersectional fairness, demographic parity, statistical parity, equal opportunity, equalized odds, uncertainty fairness, …

- Accuracy-fairness tradeoff

Fairness Definitions

Blind Fairness (Also known as fairness through unawareness)

- Predictive model should ignore sensitive attributes
- **Problem:** Other attributes may be correlated with sensitive attributes
  - Race is correlated with poverty
- **Problem:** It is "fair" to randomly predict for one subgroup as long as sensitive attributes are omitted
  (If the majority group has e.g. 95% of the data, the ML model will make random predictions for the minority (5%) group. Removing the sensitive attribute might give a false impression of fairness while a clear presence of unfairness)
- **Problem**: How do you remove sensitive attribute from an image?

Adapted from: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Blind Fairness

- **Legally Protected Attributes**
  - Race, sex, color, religion, national origin (Civil Rights Act of 1964, Equal Pay Act of 1963)
  - Age (Discrimination in Employment Act of 1967)
  - Citizenship (Immigration Reform and Control Act)
  - Pregnancy (Pregnancy Discrimination Act)
  - Familial status (Civil Rights Act of 1968)
  - Disability (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
  - Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act)
  - Genetic information (Genetic Information Nondiscrimination Act)

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Case Study: Criminal Justice

- Software by Northpointe to predict **recidivism** for defendants
  - I.e., risk of committing future crimes

- Used to help make bail, sentencing, and parole decisions

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Case Study: Criminal Justice

- **Features:** 137 questions answered by defendants or criminal records:
  - "Was one of your parents ever sent to jail or prison?"
  - "How many of your friends/acquaintances are taking drugs illegally?"
  - "How often did you get in fights while at school?"
  - Agree or disagree? "A hungry person has a right to steal"
  - Agree or disagree? "If people make me angry or lose my temper, I can be dangerous."

- Exact algorithm and model is a trade secret

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Case Study: Criminal Justice

- Race is **not** a feature
- **Problem:** Correlated features
  - E.g., poverty, joblessness and social marginalization
  - One of the developers of the system said it is difficult to construct a score that doesn't include items that can be correlated with race
  - "If those are omitted from your risk assessment, accuracy goes down"
- Similar to Amazon hiring bias example

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Individual Fairness

- "Similar" individuals (differing only on sensitive attributes) should receive "similar" outcomes
  - The prediction function $f: X \to Y$ should be Lipschitz continuous:

$$\|x - x'\| \leq \epsilon \Rightarrow |f(x) - f(x')| \leq \epsilon'$$

- **Problem:** How to define "similar"?
  - What if we include someone's accent or attire as a feature?
  - Accent may be correlated with race, in which case $\|x - x'\|$ is always large for two individuals of different race, even if race is not included as a feature

- **Problem:** Scales poorly to high-dimensional spaces

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Group Fairness

- Equalize “fairness metrics” across “subgroups”

- **Remaining challenges**
  - Need to define “subgroups” (e.g., ethnicity, gender, etc.)
  - Need to define “fairness metrics” (e.g., rate of positive outcomes, false positive/negative rates, etc.)

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Group Fairness

- **Problem setup**
  - Sensitive attribute A
  - ML model R mapping input features X to prediction $\hat{Y} = R(X)$
  - True outcome Y (typically binary, and $Y = 1$ is the "good" outcome)

- **Group fairness:** Account for performance on subgroups

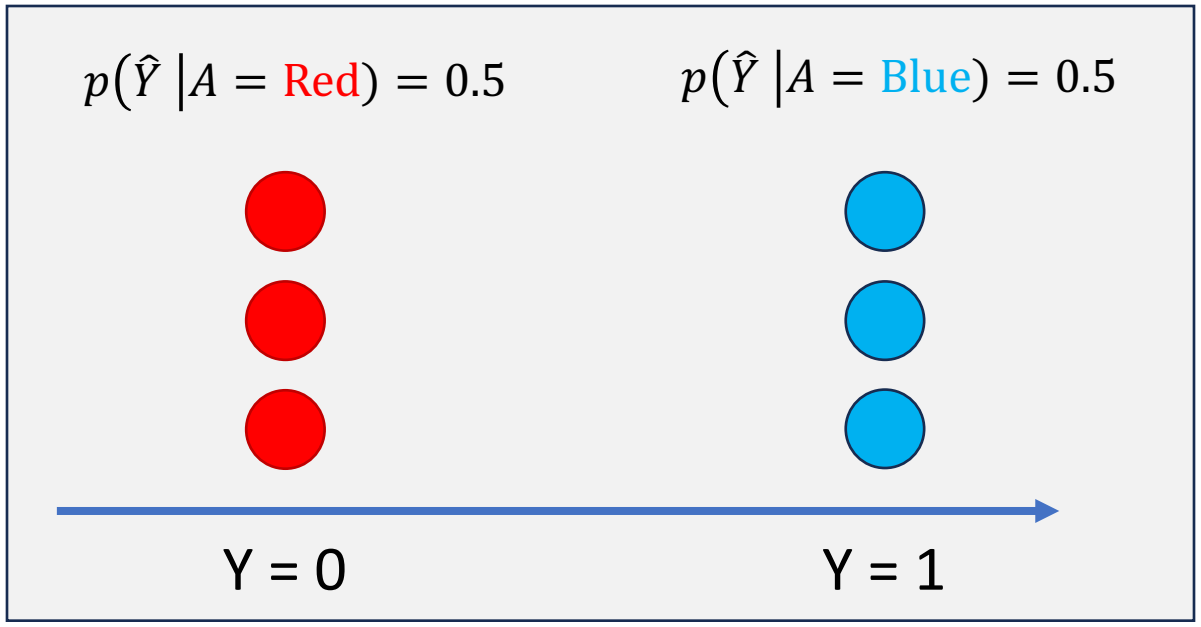
$$\text{Fairness metric} = F\big(L(f; X_1), \ldots, L(f; X_k)\big)$$

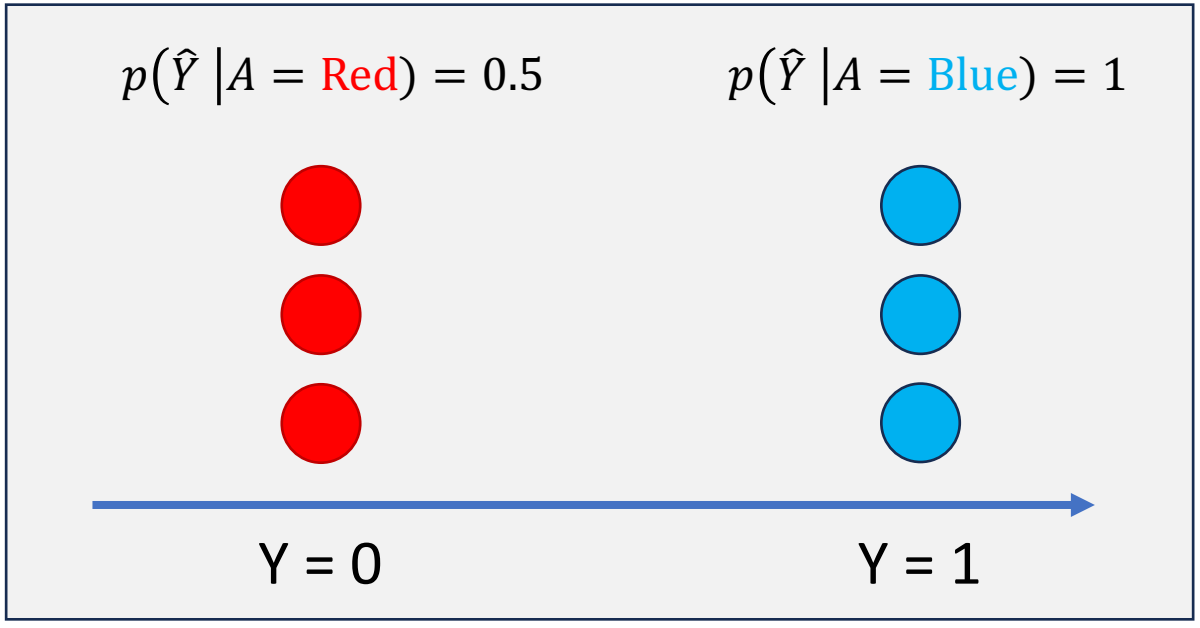
- **Example:** Insurance risk prediction
  - $A = \text{age}$, $R = \text{predicted cost}$, $Y = \text{true cost}$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Group Fairness Principles (1/3): Independence

- Prediction ($\hat{Y}$) is independent of sensitive attribute (A).
- Formally: $\hat{Y} \perp A$
- Groups should receive positive outcomes at same rate.
- Does not care about correct labels / ground truth!





Group Fairness Principles: Independence

- Prediction ($\hat{Y}$) is independent of sensitive attribute (A).
- Formally: $\hat{Y} \perp A$
- Groups should receive positive outcomes at same rate.
- Measures:
  - Demographic parity: $p\left(\hat{Y} = 1 \,\middle|\, \mathrm{A} = \mathrm{red}\right) = p\left(\hat{Y} = 1 \,\middle|\, \mathrm{A} = \mathrm{blue}\right)$
  - Disparate Impact: $\frac{p(\hat{Y}=1 \mid \mathrm{A}=\mathrm{minority})}{p(\hat{Y}=1 \mid \mathrm{A}=\mathrm{majority})}$ (Prefer this ratio to be $\geq 1 - \epsilon$)
  - Statistical Parity Difference:

$$\left|p\left(\hat{Y} = 1 \,\middle|\, \mathrm{A} = \mathrm{red}\right) - p\left(\hat{Y} = 1 \,\middle|\, \mathrm{A} = \mathrm{blue}\right)\right|$$

(Prefer this disparity to be $\leq \epsilon$)

Group Fairness Principles: Independence

- In regression problems:
  - $p(R \geq r \mid \mathrm{A} = \mathrm{red})$ vs $p(R \geq r \mid \mathrm{A} = \mathrm{blue})$
  - $mean(\hat{Y} \mid \mathrm{A} = \mathrm{red})$ vs $mean(\hat{Y} \mid \mathrm{A} = \mathrm{red})$
  - …

Group Fairness Principles: Independence

Example: Four-fifths rule [1] (Applies disparate impact ratio)

- Employed by the US Equal Employment Opportunity Commission:
  - "According to the EEOC, a selection rate for any group that is less than four-fifths (or 80%) of the rate for the group with the highest selection rate may indicate adverse impact"

- Example:
  - "Group A has 500 applicants and 100 were selected; a 20% selection rate
  - Group B has 120 applicants and 17 were selected; a 14.17% selection rate
  - The ratio is 0.1417/0.20 = 0.7083. This is below 0.80, so the procedure is biased against Group B."

[1] https://assess.com/four-fifths-rule/

Group Fairness: Independence

- **Problem:** Can assign randomly for one subgroup, no guarantee on quality of predictions across subgroups
  - Independence only cares about quantity, not quality.
  - It mandates that you approve the same percentage of people from Group A and Group B,
  - It doesn't check if you are approving the right people.

Example: Assume a personnel-hiring scenario. If 50% of the majority group is accepted, the decision maker needs to select 50% from the minority group.

If the minority group has few individuals, “qualified” individuals from the majority group will be rejected while “unqualified” individuals from the minority group might be accepted.

Group Fairness: Independence

- **Problem:** What if the base rates are not equal?

Example: Assume that 18-year-old drivers are more likely to have accidents (high risk) compared to 40-year-old drivers (low risk).

- Base Rate of Safe Driving (Young): 30%
- Base Rate of Safe Driving (Older): 80%

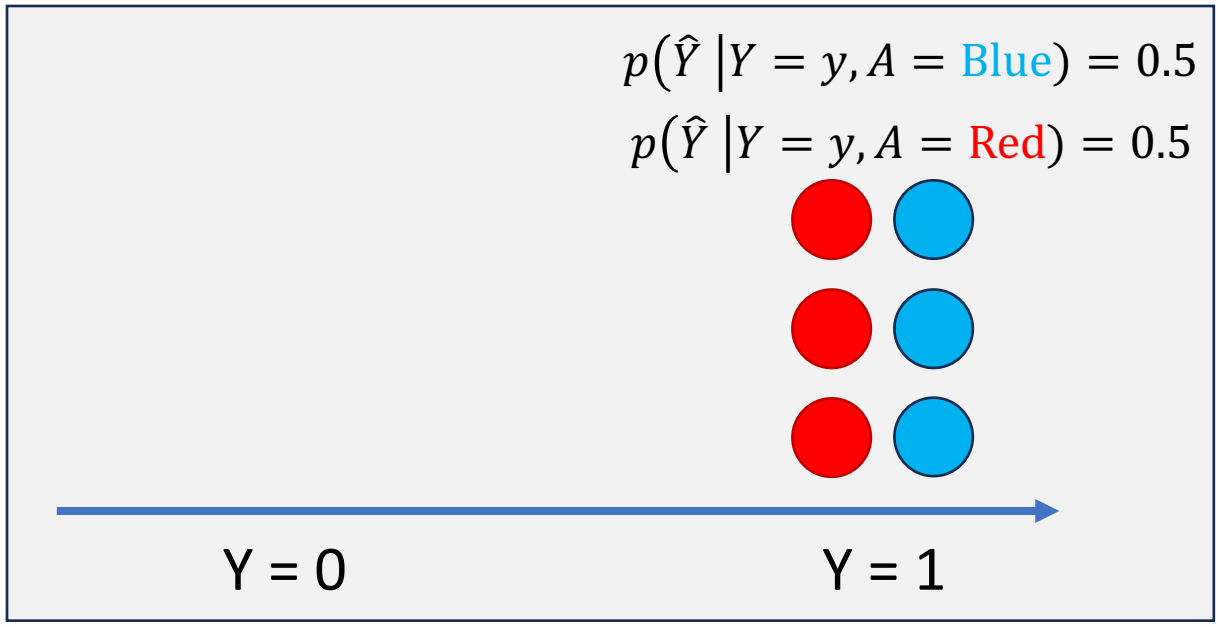
"Independence" requires the model to predict "Safe Driver" at the exact same rate for both groups:

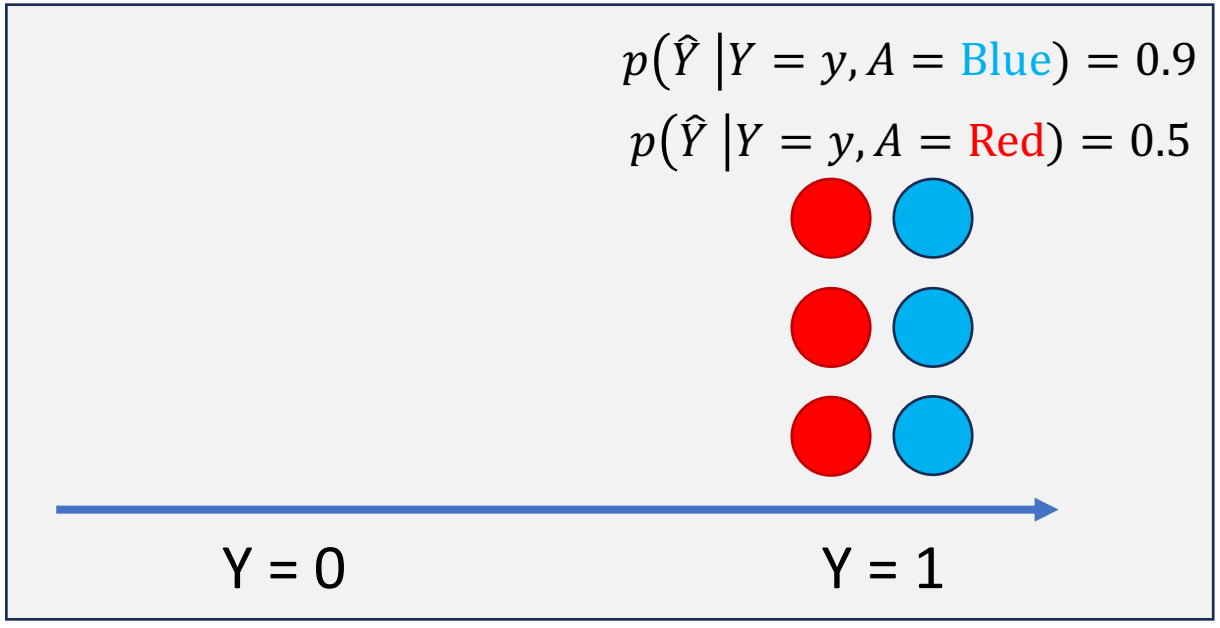
- It must either **punish** older drivers (rejecting safe drivers to lower their rate to the young group's level).
- Or it must **subsidize** young drivers (accepting dangerous drivers to raise their rate to the older group's level).

(borrowed from the directed separation notion in probabilistic graphical models)

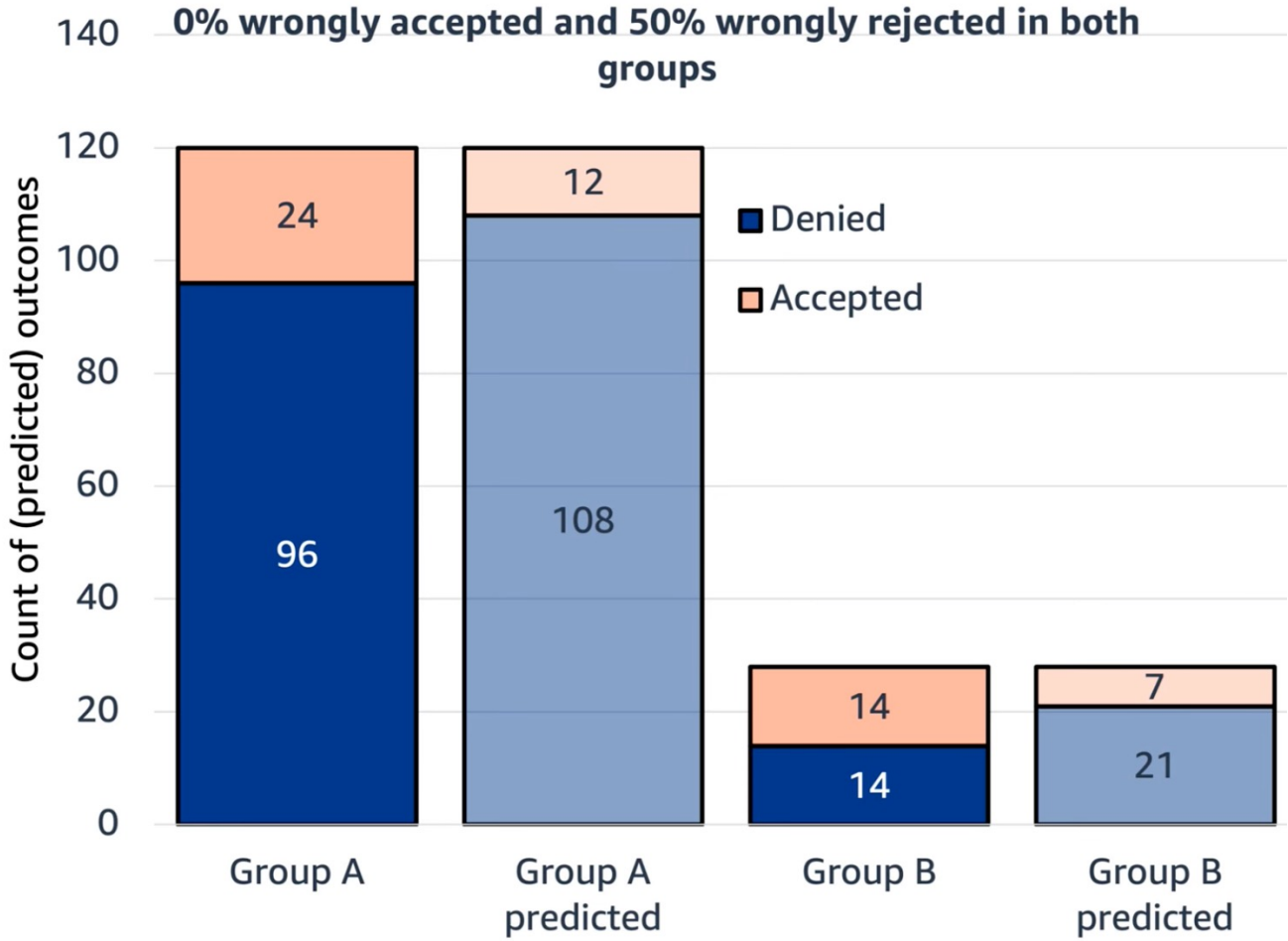
Group Fairness Principles (2/3): Separation

- Prediction is independent of sensitive attribute for positive (negative) cases
- Formally: $\hat{Y} \perp A \mid Y$
- Y **separates** $\hat{Y}$ and A. If you know Y, the outcome should not depend on A.





Group Fairness Principles: Separation



Plot from: https://www.youtube.com/watch?v=YQ5jkl_gwBA

Group Fairness Principles: Separation

- Prediction is independent of sensitive attribute for positive (negative) cases
- Formally: $\hat{Y} \perp A \mid Y$
- Y separates $\hat{Y}$ and A. If you know Y, the outcome should not depend on A.
- Measures:
  - Equal opportunity (equal True Positive Rates):
    $$p(\hat{Y} = 1 \mid Y = 1, A = \text{red}) = p(\hat{Y} = 1 \mid Y = 1, A = \text{blue})$$
  - Equalized odds (equal True Positive Rates and False Positive Rates):
    $$p(\hat{Y} = 1 \mid Y = y, A = \text{red}) = p(\hat{Y} = 1 \mid Y = y, A = \text{blue})$$
  - Predictive equality (equal False Positive Rates)
    $$p(\hat{Y} = 1 \mid Y = 0, A = \text{red}) = p(\hat{Y} = 1 \mid Y = 0, A = \text{blue})$$

Case Study: Criminal Justice

MACHINE BIAS

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson, Dec. 30, 2016, 4:44 p.m. EST

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

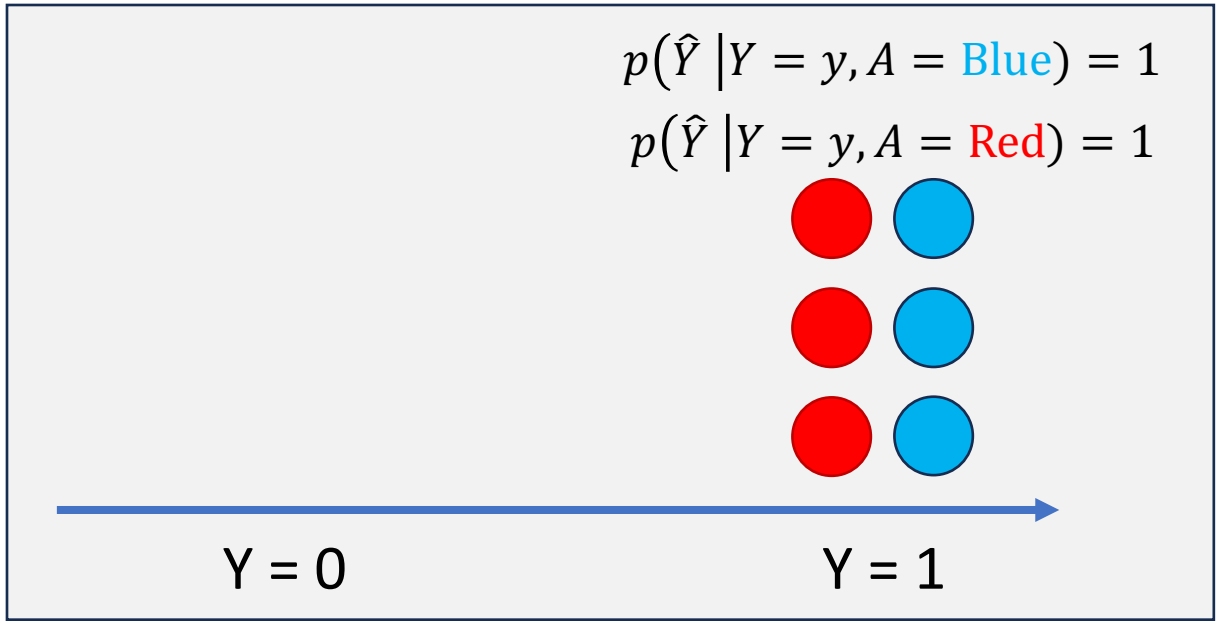
False Positive Rate

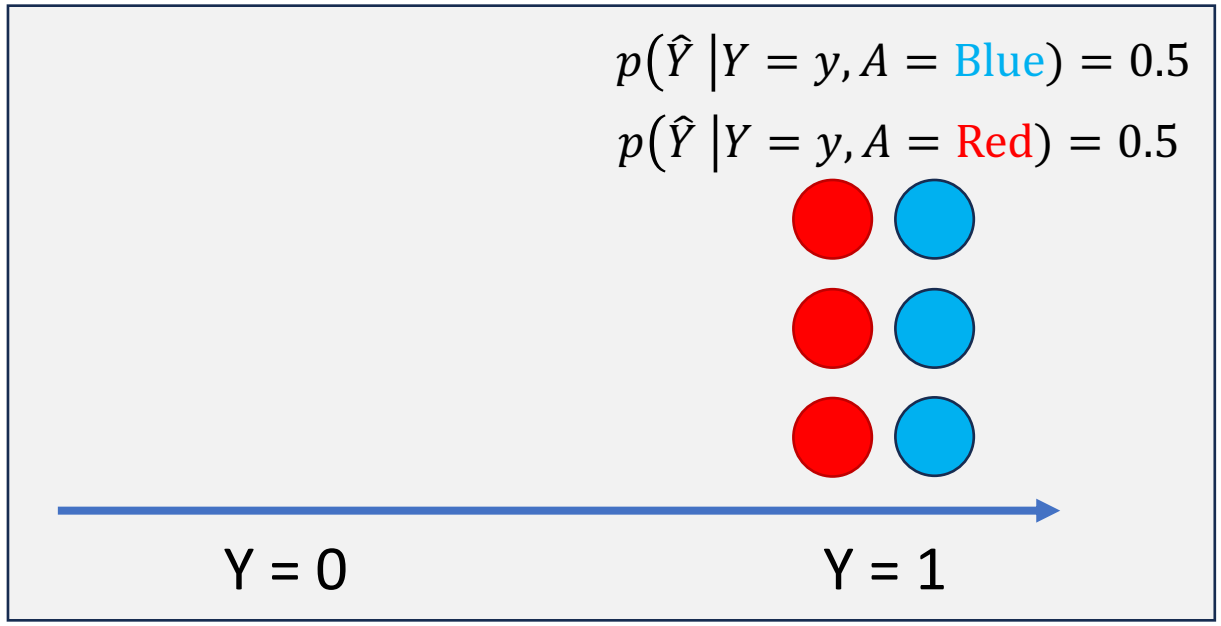
False Negative Rate

Violates the separation principle

Group Fairness Principles (3/3): Sufficiency

- Model is calibrated; i.e., true probabilities match prediction probabilities across sensitive attributes
- Sufficiency: $Y \perp A \mid \hat{Y}$
- Knowing the prediction ($\hat{Y}$) is **sufficient** to know Y; A is redundant.





Group Fairness Principles: Sufficiency

- Model is calibrated; i.e., true probabilities match prediction probabilities across sensitive attributes
- Sufficiency: $Y \perp A \mid \hat{Y}$
- Measures:
  - Calibration

$$p(Y = 1 \mid \hat{Y} = y, A = \text{red}) = p(Y = 1 \mid \hat{Y} = y, A = \text{blue})$$

("If a Black defendant and a White defendant both get a "Risk Score of 7," they should both have the exact same 70% chance of re-offending")

  - Positive Predictive Value Disparity

$$p(Y = 1 \mid \hat{Y} = 1, A = \text{red}) = p(Y = 1 \mid \hat{Y} = 1, A = \text{blue})$$

("When the model says 'Yes,' is it equally trustworthy for both groups?")

  - Negative Predictive Value Disparity
  - False Positive Value Disparity

Group Fairness: Sufficiency

- Outcome should be independent of risk score given age:

$$P(\text{true outcome}, \text{age} \mid \text{risk score}) = P(\text{true outcome} \mid \text{risk score})$$

- Equivalently, calibrated conditional on each subgroup

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Group Fairness

Independence: $\hat{Y} \perp A$
Separation: $\hat{Y} \perp A \mid Y$
Sufficiency: $Y \perp A \mid \hat{Y}$

- Three notions are incompatible!

  Proposition 2. *Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.*

  Proposition 3. *Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then, independence and separation cannot both hold.*

  Proposition 5. *Assume Y is not independent of A and assume $\hat{Y}$ is a binary classifier with nonzero false positive rate. Then, separation and sufficiency cannot both hold.*

- Thus, need carefully choose what kinds of fairness we ask for

Solon Barocas, Moritz Hardt, Arvind Narayanan, "Fairness and Machine Learning"

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Group Fairness

- Impossibility theorem [1, 2].

Counter-argument [3]**:** Strict equality of metrics is not necessary, approximate equality is sufficient. This then eliminates the conflict between the three criteria.

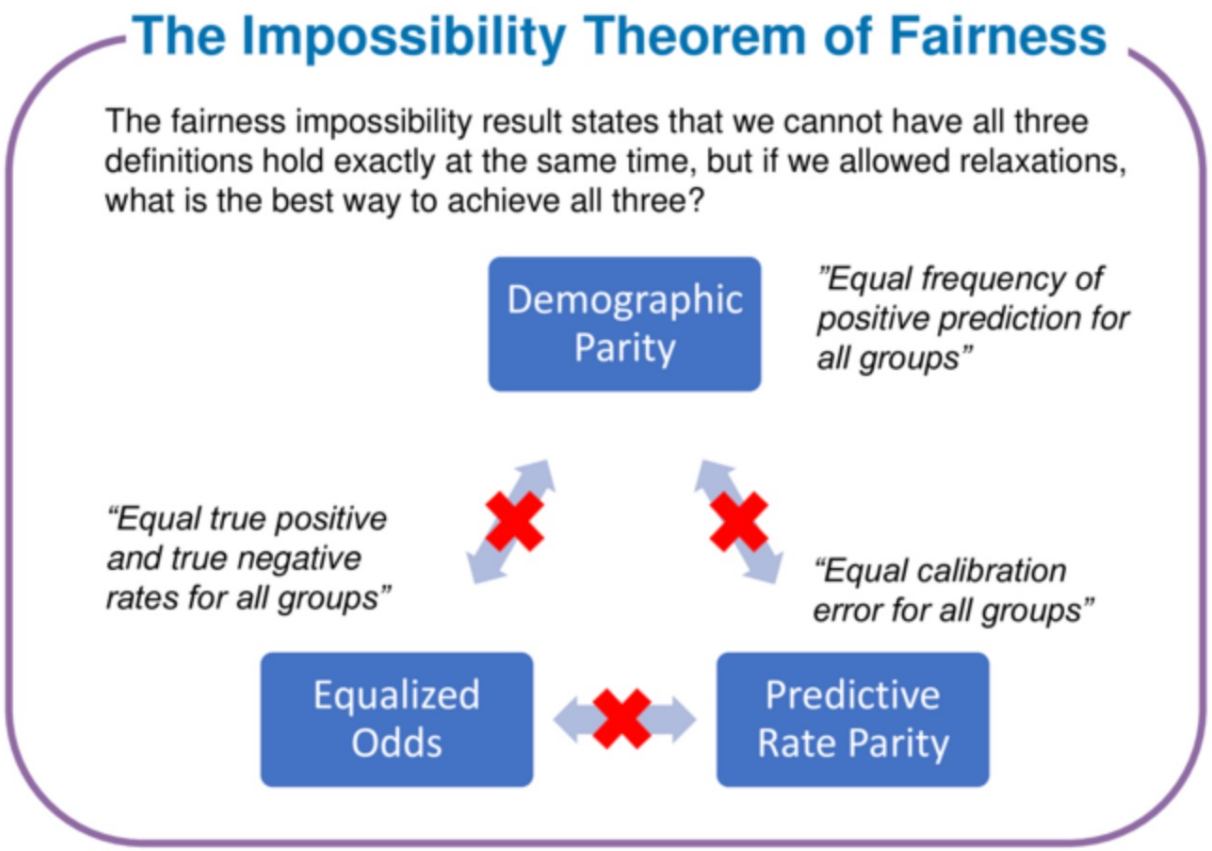
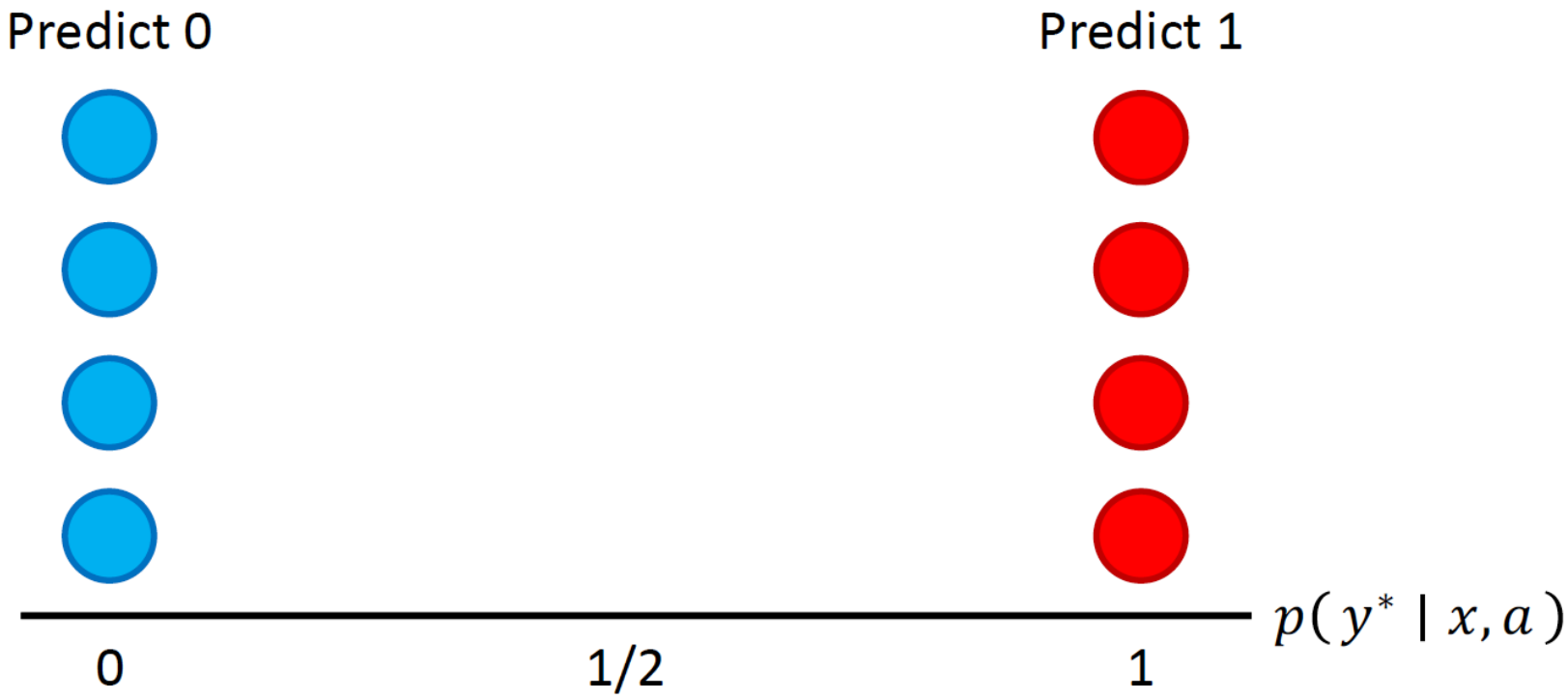


Fig from: https://neurips.cc/virtual/2022/poster/52996

[1] Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807, 2016.
[2] Miconi. The impossibility of "fairness": a generalized impossibility result for decisions. arXiv:1707.01195, 2017.
[3] Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., & Stoyanovich, J. (2023). The possibility of fairness: Revisiting the impossibility theorem in practice. ACM Conference on Fairness, Accountability, and Transparency.

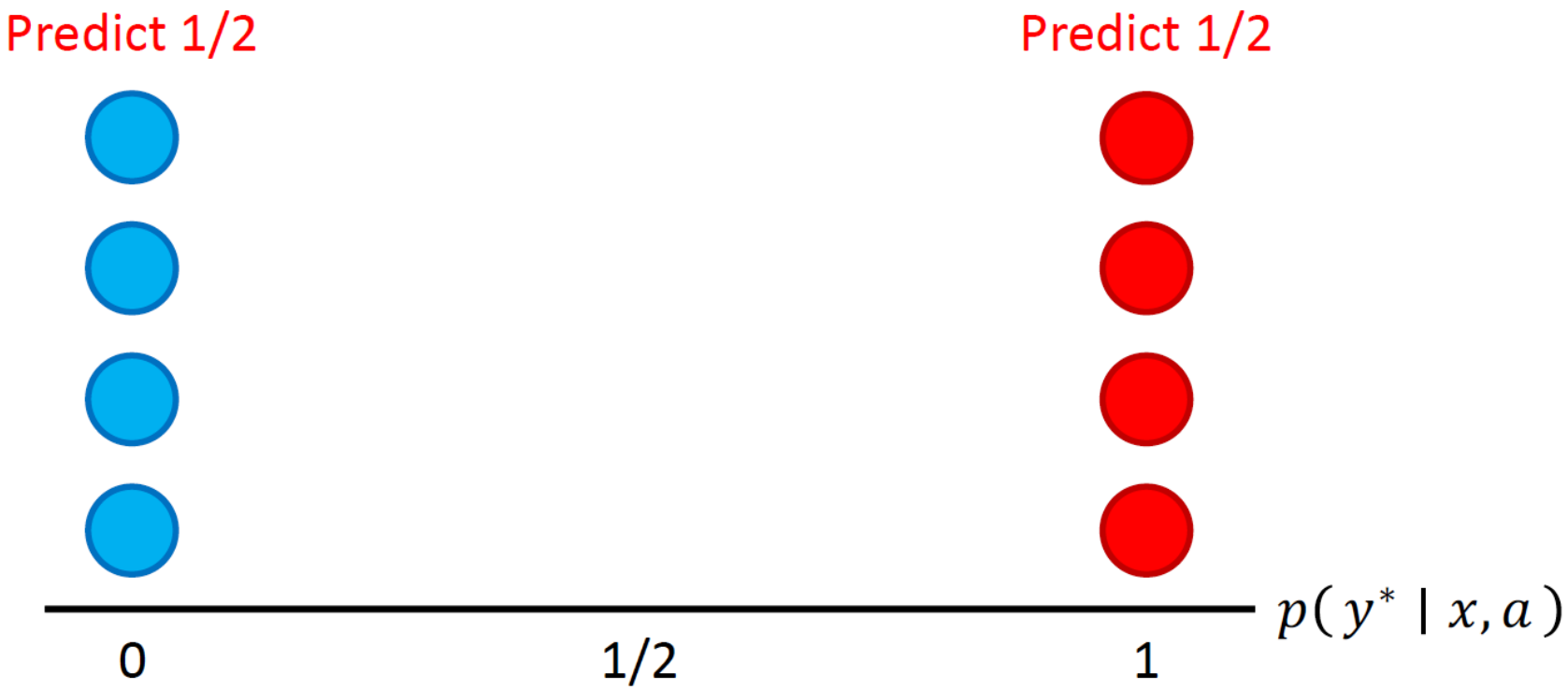
Violates independence / demographic parity:	$p(\hat{Y} = 1 \mid \mathrm{A} = \mathrm{blue}) \neq p(\hat{Y} = 1 \mid \mathrm{A} = \mathrm{red})$
Satisfies separation / equal odds	$p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \mathrm{blue}) = p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \mathrm{red})$
Satisfies sufficiency / calibration	$p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \mathrm{blue}) = p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \mathrm{red})$

A "perfect" predictor that does not make any errors!



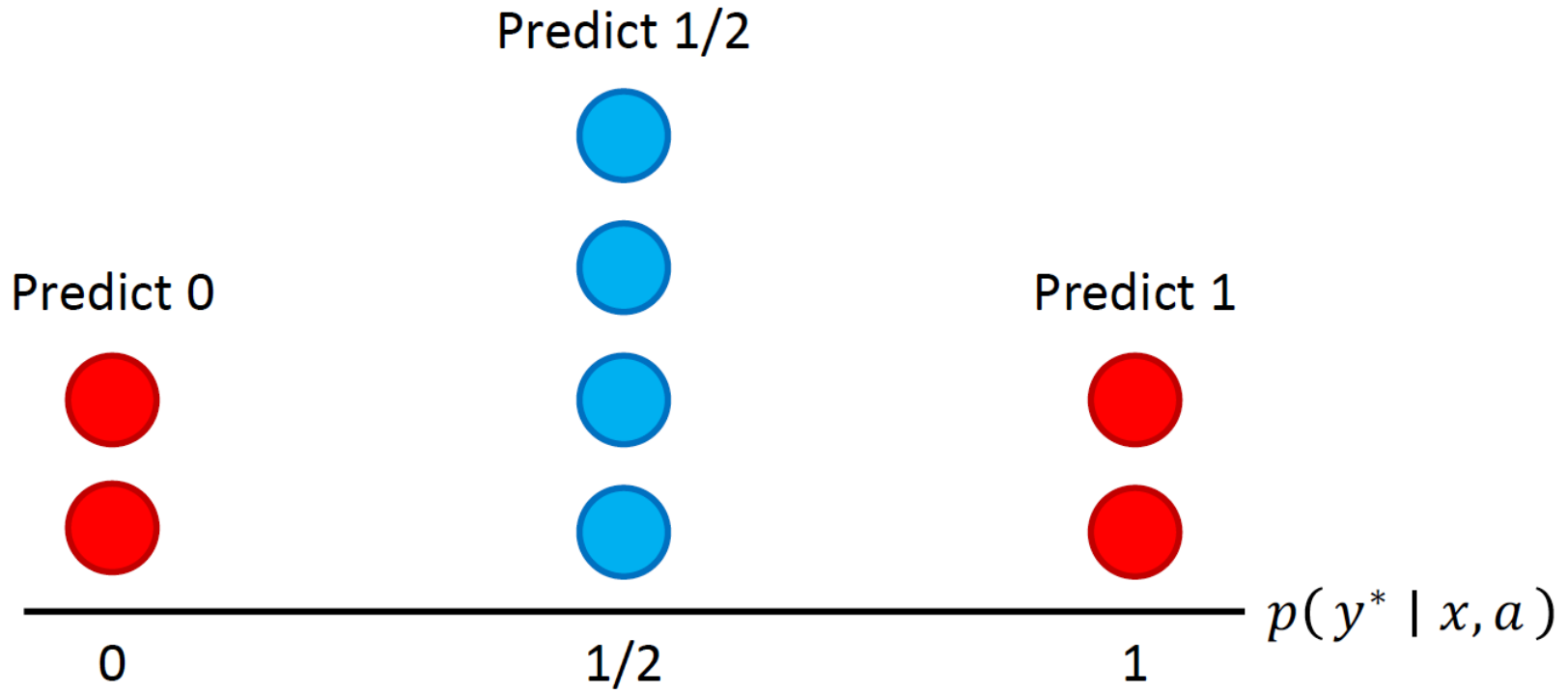
Adapted from: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Satisfies independence / demographic parity:	$p(\hat{Y} = 1 \mid \mathrm{A} = \text{blue}) = p(\hat{Y} = 1 \mid \mathrm{A} = \text{red})$
Violates separation / equal odds	$p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \text{blue}) \neq p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \text{red})$
Violates sufficiency / calibration	$p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \text{blue}) \neq p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \text{red})$



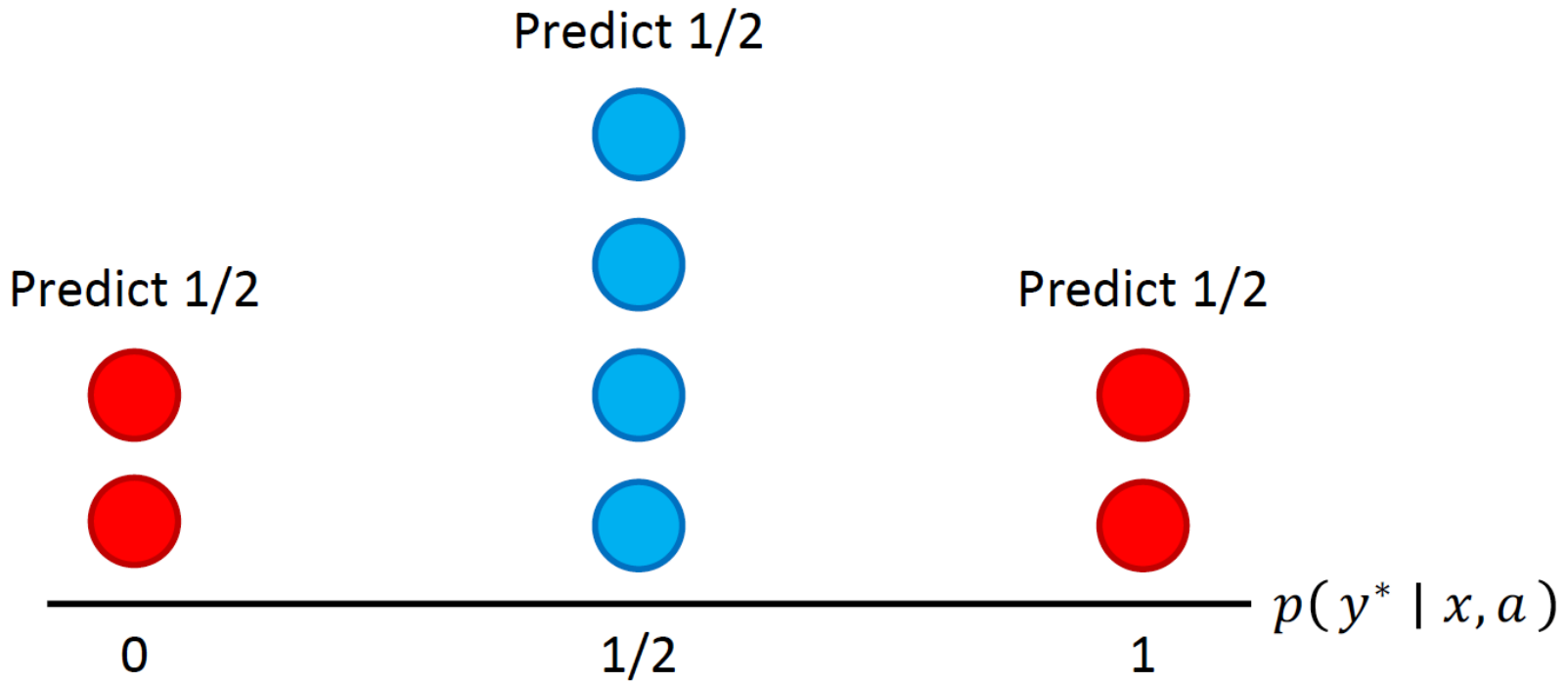
Adapted from: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Satisfies independence / demographic parity:	$p(\hat{Y} = 1 \mid \mathrm{A} = \mathrm{blue}) = p(\hat{Y} = 1 \mid \mathrm{A} = \mathrm{red})$
Violates separation / equal odds	$p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \mathrm{blue}) \neq p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \mathrm{red})$
Satisfies sufficiency / calibration	$p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \mathrm{blue}) = p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \mathrm{red})$



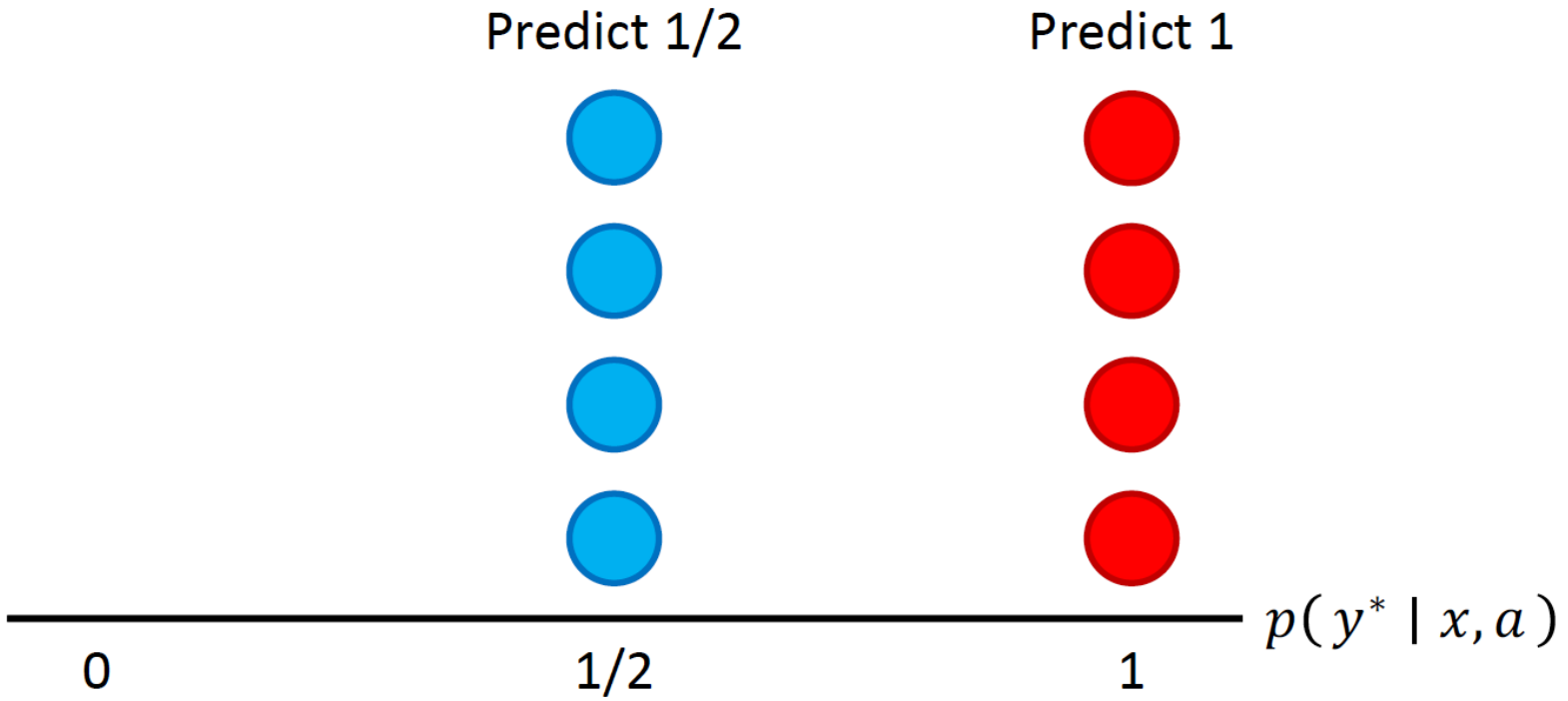
Adapted from: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Satisfies independence / demographic parity:	$p(\hat{Y} = 1 \mid \mathrm{A} = \mathrm{blue}) = p(\hat{Y} = 1 \mid \mathrm{A} = \mathrm{red})$
Satisfies separation / equal odds	$p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \mathrm{blue}) = p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \mathrm{red})$
Satisfies sufficiency / calibration	$p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \mathrm{blue}) = p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \mathrm{red})$



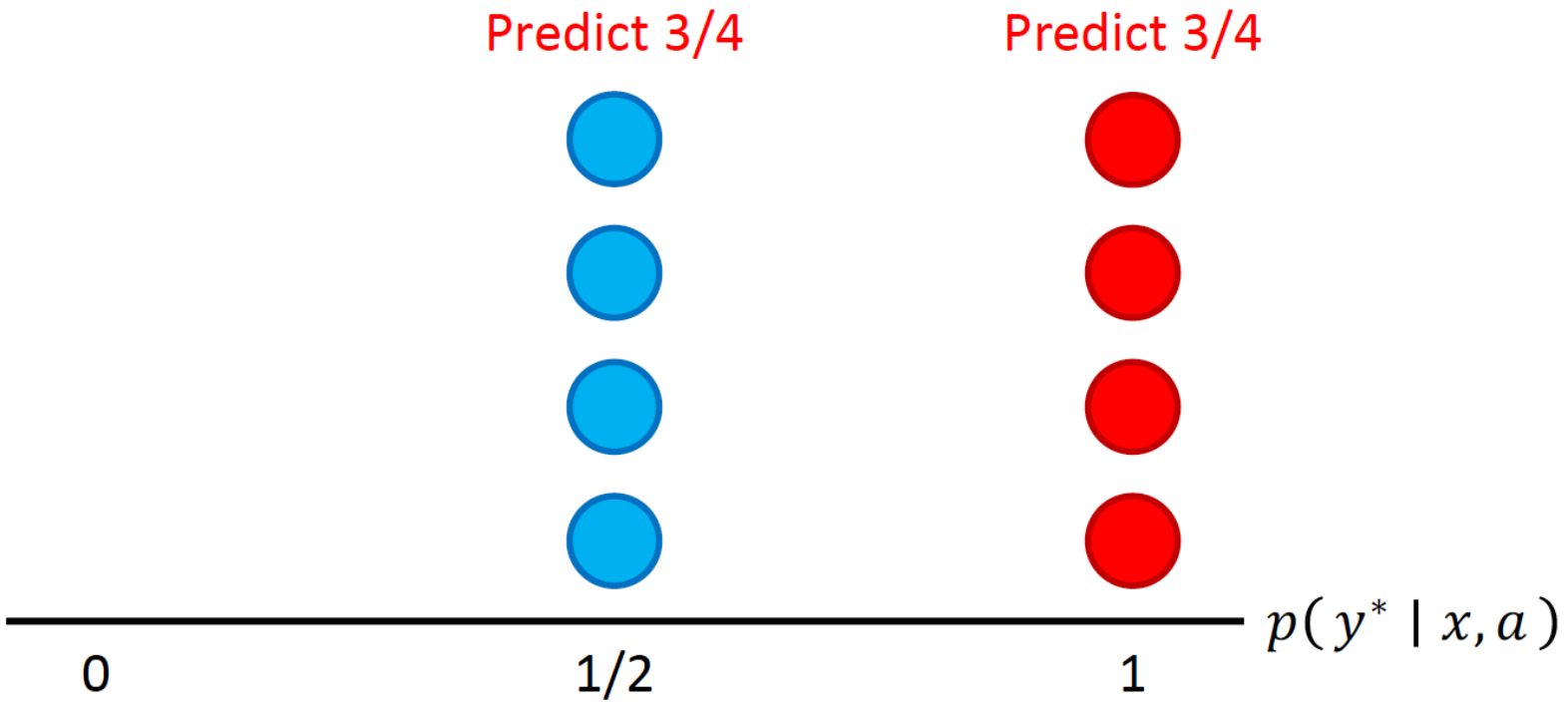
Adapted from: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Violates independence / demographic parity:	$p(\hat{Y} = 1 \mid \mathrm{A} = \text{blue}) \neq p\,(\hat{Y} = 1 \mid \mathrm{A} = \text{red})$
Violates separation / equal odds	$p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \text{blue}) \neq p\,(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \text{red})$
Satisfies sufficiency / calibration	$p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \text{blue}) = p\,(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \text{red})$



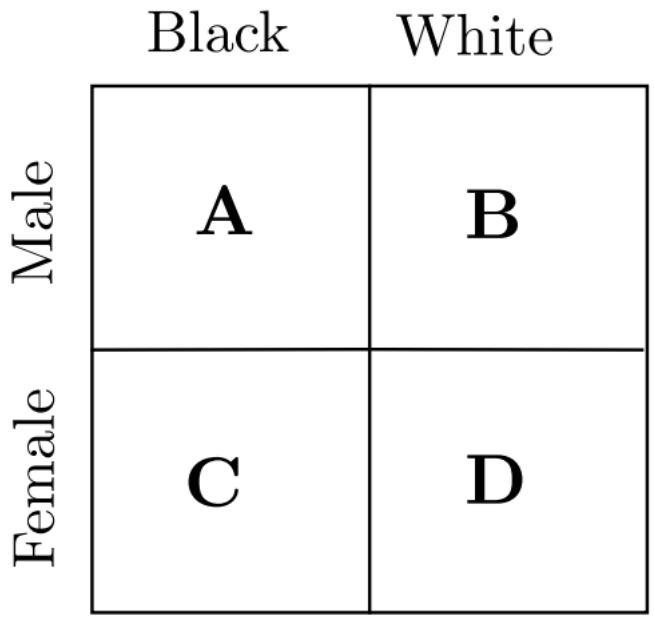
Adapted from: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Satisfies independence / demographic parity:	$p(\hat{Y} = 1 \mid \mathrm{A} = \text{blue}) = p(\hat{Y} = 1 \mid \mathrm{A} = \text{red})$
Violates separation / equal odds	$p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \text{blue}) \neq p(\hat{Y} = 1 \mid Y = \mathrm{y}, \mathrm{A} = \text{red})$
Violates sufficiency / calibration	$p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \text{blue}) \neq p(Y = 1 \mid \hat{Y} = \mathrm{y}, \mathrm{A} = \text{red})$



Adapted from: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Other Definitions: Intersectional Fairness



Gohar, U., & Cheng, L. (2023). A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. arXiv preprint arXiv:2305.06969.

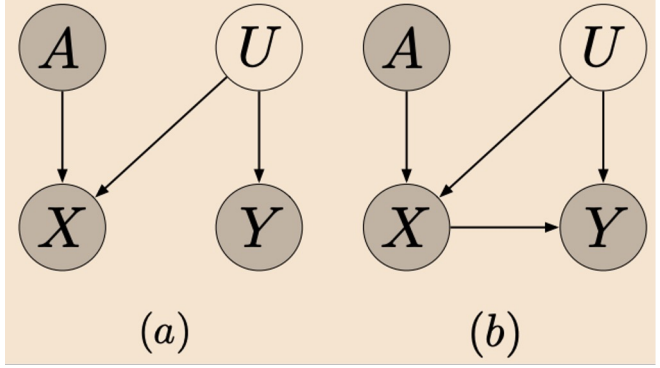
Other Definitions: Counterfactual Fairness

Given a predictive problem with fairness considerations, where A, X and Y represent the protected attributes, remaining attributes, and output of interest respectively, let us assume that we are given a causal model (U, V, F), where $V \equiv A \cup X$. We postulate the following criterion for predictors of Y.

Definition 5 (Counterfactual fairness). *Predictor* $\hat{Y}$ *is* **counterfactually fair** *if under any context* $X = x$ *and* $A = a$,

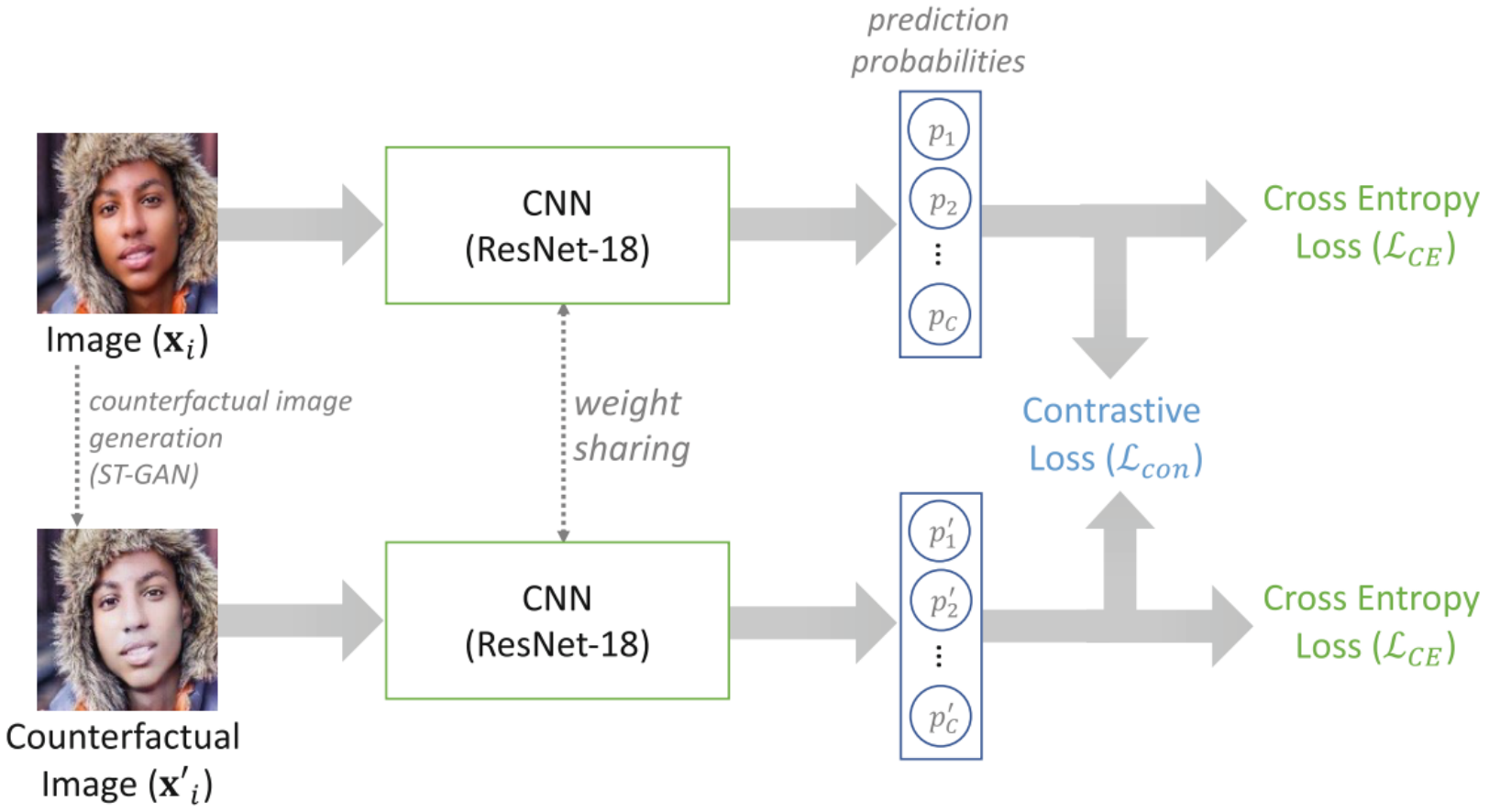
$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \quad (1)$$

for all y *and for any value* a' *attainable by* A.



Material: https://arxiv.org/pdf/1703.06856

Other Definitions: Counterfactual Fairness
Sample Study from Our Group



Cheong, J., Kalkan, S., & Gunes, H. (2022, October). Counterfactual fairness for facial expression recognition. In European Conference on Computer Vision (pp. 245-261). Cham: Springer Nature Switzerland.

Other Definitions: Uncertainty Fairness Sample Study from Our Group

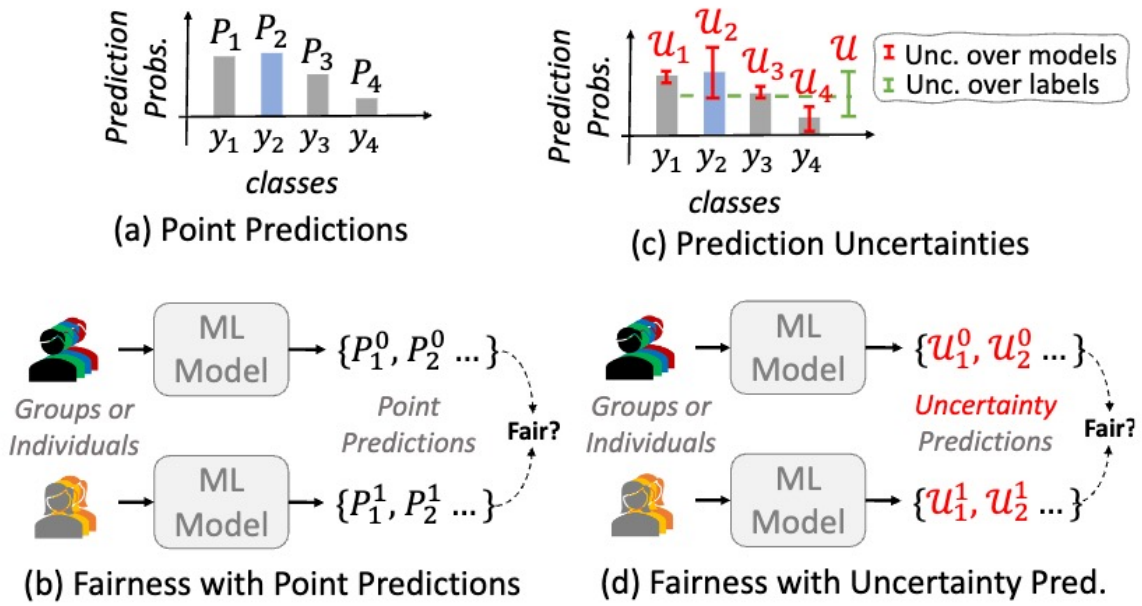


Figure 1: Existing fairness measures utilize point predictions for quantifying fairness, which ignores the uncertainty (variance) of the predictions **(a-b)**. We fill this gap by using uncertainty instead for measuring fairness **(c-d)**.

S. Kuzucu, J. Cheong, H. Gunes, S. Kalkan, "Uncertainty-based Fairness Measures", Journal of Artificial Intelligence Research, Special Issue on Fairness and Bias in AI, 81:307-335, 2024.

Other Definitions: Uncertainty Fairness
Sample Study from Our Group

Definition 4.1 (UNCERTAINTY-FAIRNESS MEASURE). *A model is fair if its uncertainties are the same across different groups. More formally, extending the definition in Section 3.2*

$$Fair(f;\mathcal{U},D) \equiv \mathcal{U}(D,f,G=0) = \mathcal{U}(D,f,G=1), \quad (8)$$

where $\mathcal{U}$ is an uncertainty measure, e.g., predictive uncertainty ($\mathcal{U}_p$), epistemic uncertainty ($\mathcal{U}_e$), or aleatoric uncertainty ($\mathcal{U}_a$) as introduced in Section 4.1.

S. Kuzucu, J. Cheong, H. Gunes, S. Kalkan, "Uncertainty-based Fairness Measures", Journal of Artificial Intelligence Research, Special Issue on Fairness and Bias in AI, 81:307-335, 2024.

Other Definitions: Uncertainty Fairness
Sample Study from Our Group

Proposition 4.1 (INDEPENDENCE OF UNCERTAINTY FAIRNESS). *Consider a predictor* $f(\cdot;\theta)$ *with point-predictions* $\{\hat{y}_i\}_i$ *(and associated probabilities* $\{P(\hat{y}_i|\mathbf{x}_i)\}_i$*) and uncertainties* $\{\mathcal{U}_i\}_i$ *(namely, predictive, epistemic and aleatoric). Then, uncertainty fairness* $Fair(f;\mathcal{U},D)$ *is independent to the conventional point-measure based fairness* $Fair(f;\mathcal{M},D)$. *More formally:*

- $Fair(f;\mathcal{M},D) \not\Rightarrow Fair(f;\mathcal{U},D)$.
- $Fair(f;\mathcal{U},D) \not\Rightarrow Fair(f;\mathcal{M},D)$.

$Fair(f;\mathcal{U},D)$ *does not imply* $Fair(f;\mathcal{M},D)$ *or vice versa.*

$Fair(f;M,D)$: Point-prediction-based fairness measure (e.g., equal opportunity)

S. Kuzucu, J. Cheong, H. Gunes, S. Kalkan, "Uncertainty-based Fairness Measures", Journal of Artificial Intelligence Research, Special Issue on Fairness and Bias in AI, 81:307-335, 2024.