ellis ALICANTE unit

Fairness in ML

A general introduction about Fairness in Algorithmic ML

Adrián Arnaiz-Rodríguez
1y PhD Student
ELLIS Alicante
Universidad de Alicante

17 January 2022

European Union
European Regional Development Fund
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
GENERALITAT VALENCIANA
Conselleria d'Innovació, Universitats, Ciència i Societat Digital
Sabadell Fundación

- Introduction to Algorithmic Fairness
- Bias
- Fairness definitions
- Limitations in definitions
- Imposing Fairness
- Current prominent approaches
- Datasets
- Libraries
- History and conceptual concerns
- General conclusions

Algorithmic bias problem and fairness at a glance

ML is used for critical decision making

How bias appears in society:

- Sources of bias
- Examples of bias

Challenges of AI

- Uncover bias/unfairness
- Measure bias (definitions Fairness)
- Mitigate bias
- Real world applications

How do we formulate the bias-fairness problem in every problem set up?

How do we detect the bias and how to solve it?

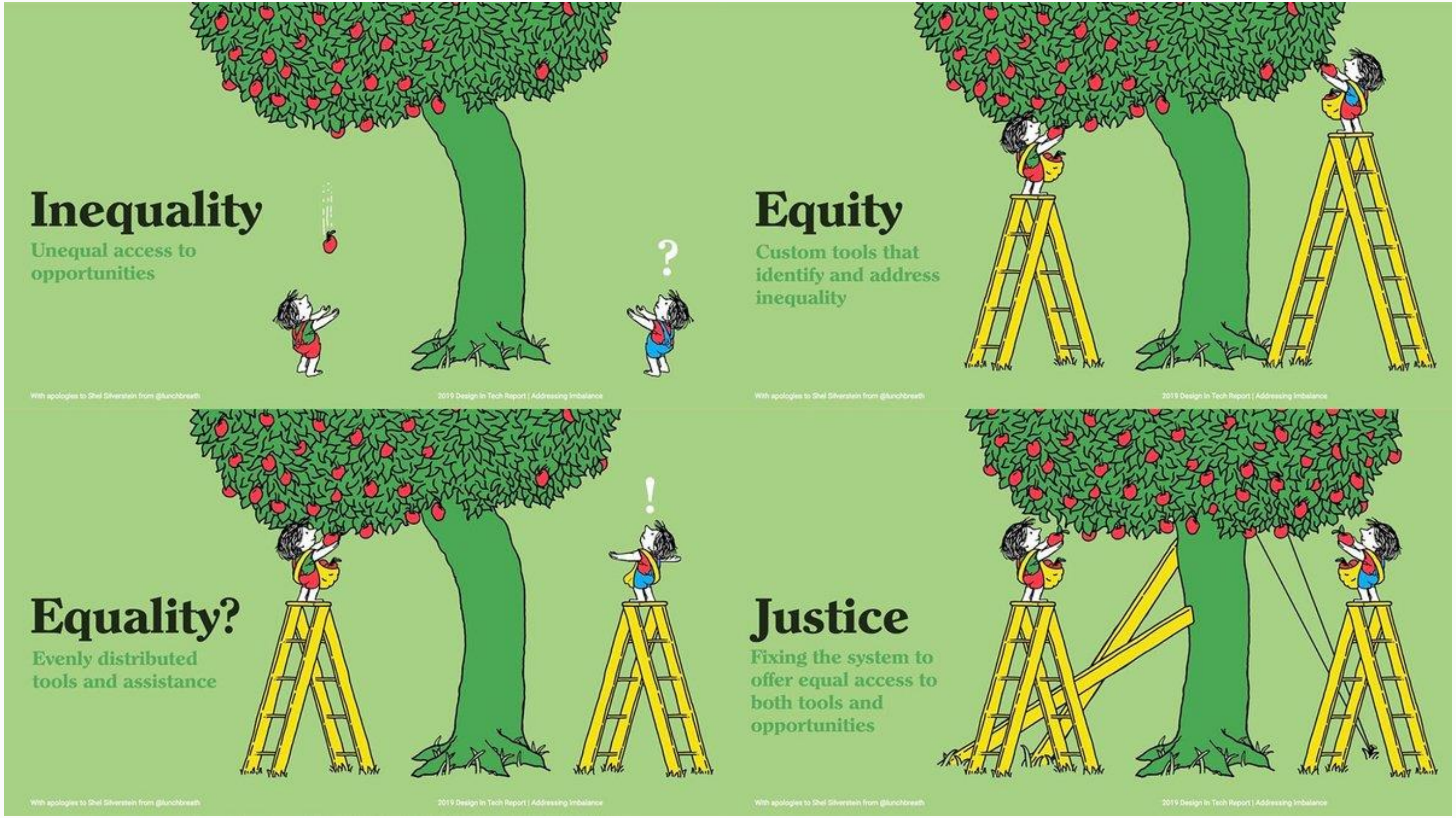
How could we define and measure bias or fairness?

Which are the ethical principles that follows each definition of bias and fairness?

Which are the implications in the real-world problems and, specifically in our own value system?

What is fairness for you?

Justice, equality and equity



Introduction to Algorithmic Fairness

ML for critical decision making

- ML models are becoming the main tools for addressing complex societal problems in many consequential areas of our lives
  - Education
  - Justice: pretrial and detention
  - Security
  - Health
  - Child Maltreatment screening
  - Social Services
  - Hiring
  - Finance
  - Advertising

- Each one with its own objectives
  - Reduce cost
  - Maximize social benefit
  - …

✓ Privacy ✓ Transparency ✓ Accountability	✓ Reliability ✓ Autonomy ✓ **Fairness**

Ethical implications

Many of these concepts do not have universally accepted definitions

Harms from Algorithmic Decision-Making

INDIVIDUAL HARMS ILLEGAL DISCRIMINATION / UNFAIR PRACTICES	COLLECTIVE SOCIAL HARMS
HIRING	LOSS OF OPPORTUNITY
EMPLOYMENT	
INSURANCE & SOCIAL BENEFITS	
HOUSING	
EDUCATION	
CREDIT	ECONOMIC LOSS
DIFFERENTIAL PRICES OF GOODS	
LOSS OF LIBERTY	SOCIAL STIGMATIZATION
INCREASED SURVEILLANCE	
STEREOTYPE REINFORCEMENT	
DIGNATORY HARMS	

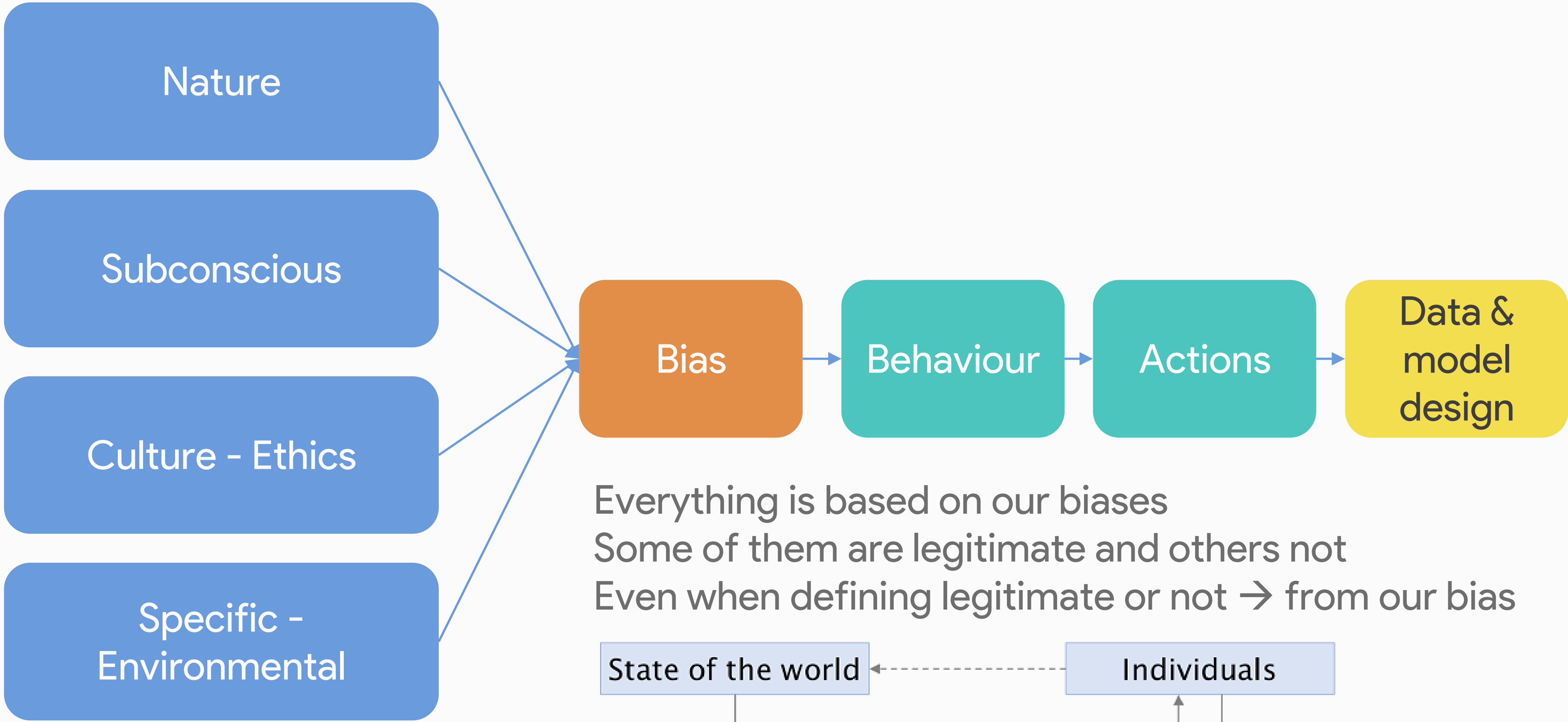
Chart Contents Courtesy of Megan Smith, Former CTO of the United States

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In FAccT. PMLR. http://gendershades.org/overview.html

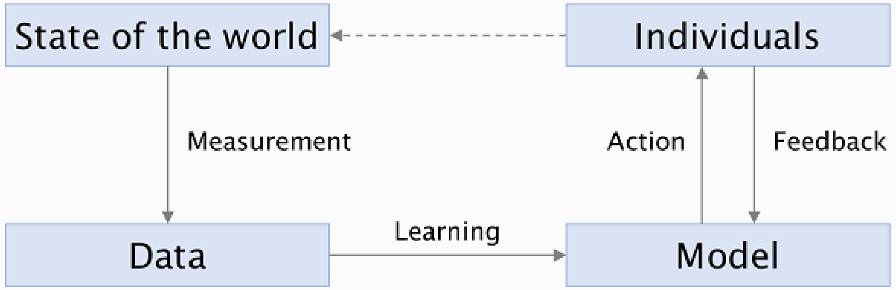
ML for critical decision making - examples

- Finance
  - *A. Byanjankar, M. Heikkilä, and J. Mezei. Predicting credit risk in peer-to-peer lending: A neural network approach. In IEEE Symposium Series on Computational Intelligence, 2015*
- Hiring
  - *M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. Technical report, Upturn, 2018*
- Pretrial and detention
  - *J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks., 2016.*
- Child maltreatment screening
  - *A. Chouldechova, E. Putnam-Hornstein, D. Benavides-Prado, O. Fialko, and R. Vaithianathan. A case study of algorithmassisted decision making in child maltreatment hotline screening decisions. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pages 134–148, 2018.*
- Education
  - *L. Oneto, A. Siri, G. Luria, and D. Anguita. Dropout prediction at university of genoa: a privacy preserving data driven approach. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2017.*
- Social Services
  - *V. Eubanks. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press, 2018*

Bias is implicit in every decision we make



Everything is based on our biases
Some of them are legitimate and others not
Even when defining legitimate or not → from our bias



Human centric ML approaches

AI systems learning moral notions

*AI-based systems can **learn moral notions** or ethical behaviors and then **autonomously behave ethically***

- Comparative Moral Turing Test
- Ethical Turing Test

➢ Evaluate the morality of the choices of automated systems

➢ **Branch quite unexplored:** difficult connection between philosophy, ethic and technical problems

➢ AGI related

How humans should design AI systems to minimize harms

*Designing for **minimizing** harms derived from **poor design**, **bad applications** and **misuse** of the systems*

- **Algorithmic Fairness**
- Privacy Preserving Data Mining – Federated Learning
- Explainable AI [2] & Interpretable AI
- Adversarial Learning

➢ Many more examples due to many different ML methods and problems addressed

HCML Perspective: building responsible AI including human relevant requirements, but also considering broad societal issues [1]

- Safety, **Fairness**, privacy, accountability & interpretability - Ethics and legislation

Franco, D., Navarin, N., Donini, M., Anguita, D., & Oneto, L. (2022). **Deep fair models for complex data: Graphs labeling and explainable face recognition.** Neurocomputing, 470
1. A.F. Winfield, K. Michael, J. Pitt, V. Evers, **Machine ethics: the design and governance of ethical ai and autonomous systems**, Proceedings of the IEEE 107 (2019) 509–517
2. D. Gunning, **Explainable artificial intelligence (xai)**, Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2).

Human centric ML approaches

AI systems learning moral notions

*AI-based systems can **learn moral notions** or ethical behaviors and then **autonomously behave ethically***

- Comparative Moral Turing Test
- Ethical Turing Test
- ➢ Evaluate the morality of the choices of automated systems
- ➢ **Branch quite unexplored:** difficult connection between philosophy, ethic and technical problems

How humans should design AI systems to minimize harms

*Designing for **minimizing** harms derived from **poor design, bad applications** and **misuse** of the systems*

- **Algorithmic Fairness**
- Privacy Preserving Data Mining – Federated Learning
- Explainable AI [2] & Interpretable AI
- Adversarial Learning
- ➢ Many more examples due to many different ML methods and problems addressed

HCML Perspective: building responsible AI including human relevant requirements, but also considering broad societal issues [1]

- Safety, **Fairness**, privacy, accountability & interpretability - Ethics and legislation

Franco, D., Navarin, N., Donini, M., Anguita, D., & Oneto, L. (2022). **Deep fair models for complex data: Graphs labeling and explainable face recognition.** Neurocomputing, 470
1. A.F. Winfield, K. Michael, J. Pitt, V. Evers, **Machine ethics: the design and governance of ethical ai and autonomous systems**, Proceedings of the IEEE 107 (2019) 509–517
2. D. Gunning, **Explainable artificial intelligence (xai)**, Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2).

What should we consider to formally defining fairness?

- How we **define different discriminations**?
- What are the **main sources of bias**?
- How we **define fairness** and **measure it**?
- How do we **find bias** in our models?
- How we **mitigate bias** / **impose fairness** in our models?
  - What kind of different approaches are there?
- What are some examples of **real applications**?

Hints on the complexity of formally defining fairness

Different kind of **discriminations**

What is discrimination?

Many **sources of bias**

How is it caused?

Different **fairness definitions** based on different fundamentals

How can we define unfairness and how I measure it?

Countless **types of models** in which bias is analyzed and fairness is imposed

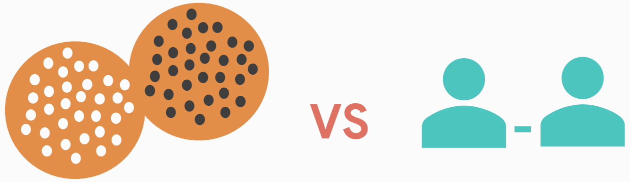
How can we find unfair models?
How can we implement fair models?

Numerous real problems

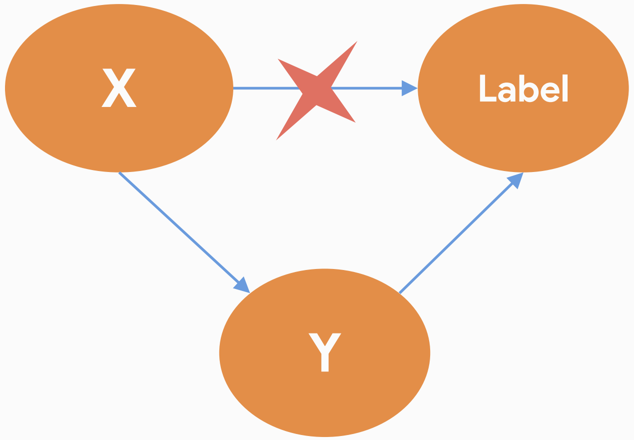
How do we eventually apply this?

Algorithmic Fairness

- Algorithmic Fairness deals with the problem of developing AI-based systems able to treat:
  - **Subgroups in the population <u>equally</u>** → **Group fairness**
  - **Similar individuals in a <u>similar</u> way** → **Individual Fairness**



- Subgroups → determined by means of sensitive attributes, considered for decisions
  - *Gender*, *incomes*, *ethnicity*, and *sexual* or *political orientation* and so on



How do we define equally?
How we define similar?

Algorithmic Fairness

- How to enhance ML models with fairness requirements, not unethically biasing decisions

Training data Bias → Unfair decisions due to sensitive attributes
Model inaccuracies → Unfair decisions due to sensitive attributes

- Ensure that the outputs of a model DO NOT depend on sensitive attributes
  - *In a way that is considered unfair - differences due to such traits cannot be reasonably justified*

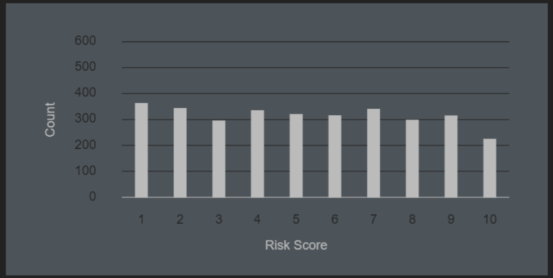
$$F(X) = R,\ A \in X \rightarrow \mathrm{R} \perp \mathrm{A}$$

- Many approaches: properties of the model outputs with respect to the sensitive attributes
- **Relationships among all relevant variables in the data → unfairness underlying**
  - If not → COMPAS: biased recidivism application even not using sensitive data

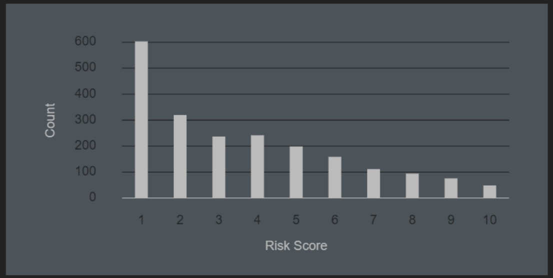
L. Oneto, S. Chiappa, Fairness in machine learning, Recent Trends in Learning From Data (2020)

Two Petty Theft Arrests

Black Defendants' Risk Scores



White Defendants' Risk Scores



Two Drug Possession Arrests

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%

Machine Bias

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

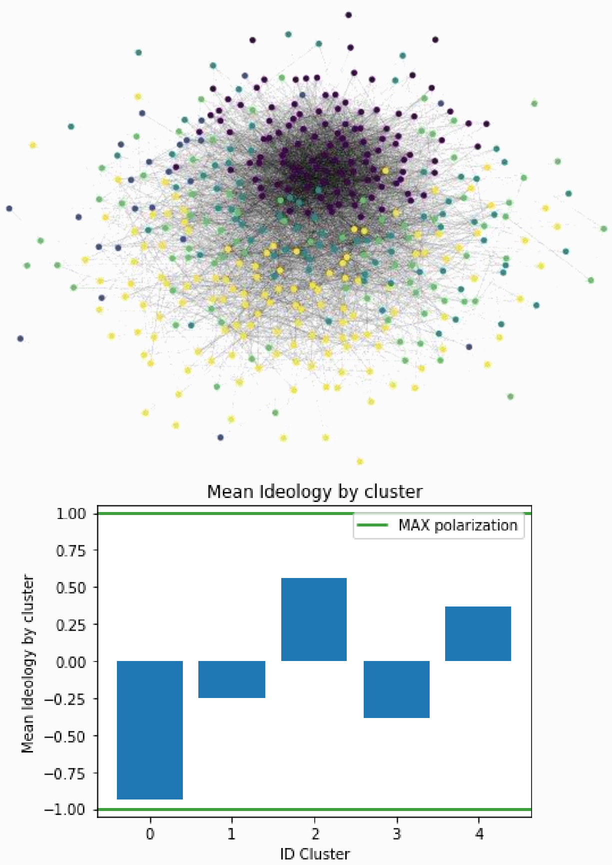
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

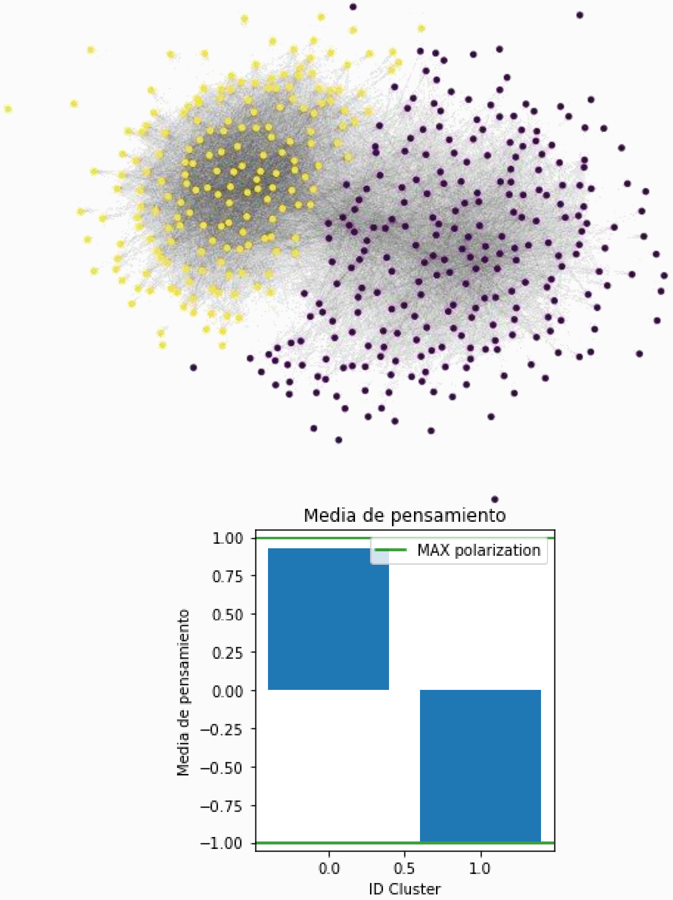
Correctional Offender Management Profiling for Alternative Sanctions - COMPAS

Not only fair decisions: echo chambers

- US House of Representatives 1973 VS 2016
- Two politicians are linked if they have supported 3+ initiatives together



1973



2016

Personal Project: https://twitter.com/Arnaiztech/status/1331996276045582339

Before kicking off: spoiler

- There are quite a lot different approaches to mitigating unfairness.
- No single approach is universally best → No free lunch ☹
- Choosing the most appropriate one will require:

Expert judgement

Knowledge of relevant legal and compliance requirements

Context in which we are working

Takeaway: Choosing Fairness metric and method <u>highly depends on the context</u>

No universal fairness definition or bias mitigation / imposing fairness approach

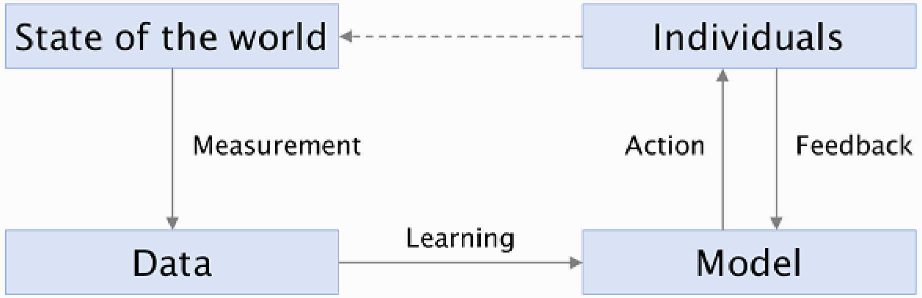
Bias

Different types

Bias & Sources

1. How law define bias?
   - Disparate treatment
   - Disparate impact

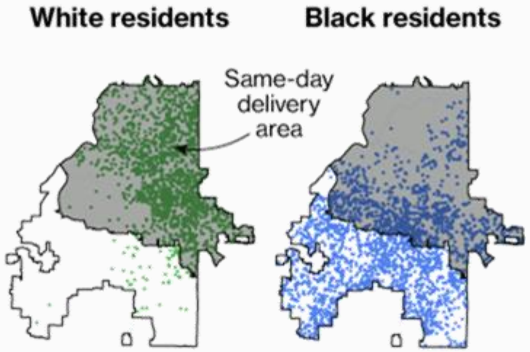
2. Bias in in ML
   - By source
   - By interaction

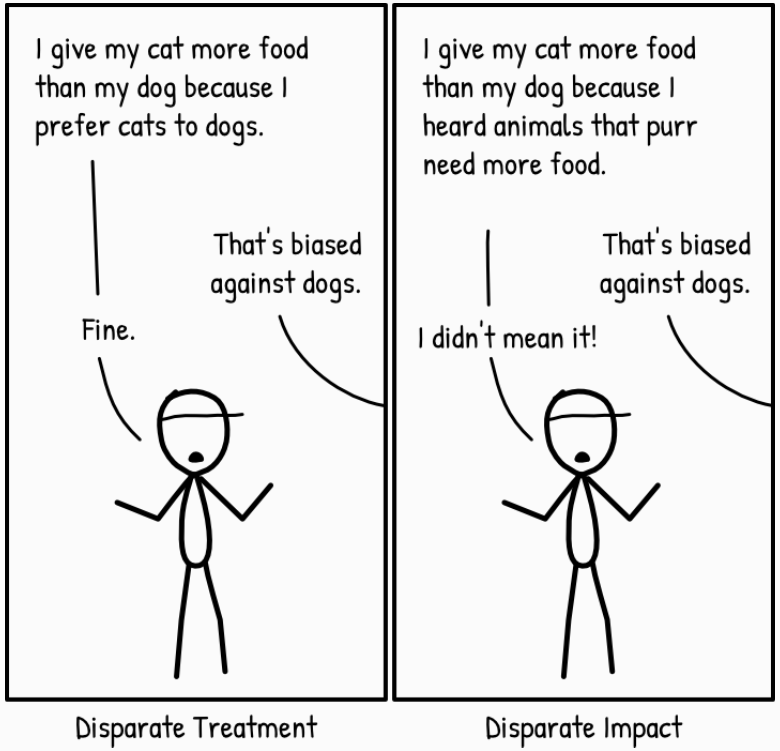


Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671

Disparate Treatment and Impact

- Anti-discrimination laws in various countries prohibit unfair treatment of individuals
- Legal or ethical support and formalize it quantitively
  - **Disparate treatment:**
    - Decisions are (partly) based on the subject's sensitive attribute
    - Explicit or intentional
  - **Disparate impact:**
    - Outcomes or implemented policy disproportionately hurt people with certain sensitive attribute
    - Implicit or unintentional





Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671
Lim Swee Kiat. Retrieved December 2021. Machines go Wrong. https://machinesgonewrong.com/fairness/
Ingold, D. and Soper, S., 2016. Amazon doesn't consider the race of its customers. Should It?. Bloomberg News.

Sources of Bias – Data

Bias in historical data
- **Skewed** towards groups or **imbalanced** limited information
  - Amazon, COMPAS or 2018-CEO-image-search
- Easy to ignore biases and surrogate variables for protected attributes
- Label imperfectly observed: Label bias
- Record of crimes comes from crimes observed by police

Bias in data collection mechanisms
- Inherent biases in the data collection mechanisms
- Lack of representativeness
- Crowdsourcing from a technology that only uses a type of people → Autonomous car related with wealthier

Bias in alternate sources of data
- "New" sources of data: worldwide web, social media, blogs
- Digital footprint variables: computer brand or type of device
  - Proxies of protected attributes
  - Socio-economic variables → surrogates for protected groups

Selective labels - Unobservable Outcomes
- Observed outcomes are consequence of the existing choices of the human decision-makers
  - → Label distribution based on previous policy
- Was former policy accurate or biased?
- Would they have defaulted if had they been approved for a mortgage? → Counterfactual
- Tainted samples → Decision-maker bias
- *We observe loan defaults only for those who received a mortgage → we do not have any information for those who were denied*
- *We observe whether a defendant fails to return for their court appearance only if the human judge decides to release the defendant on bail*

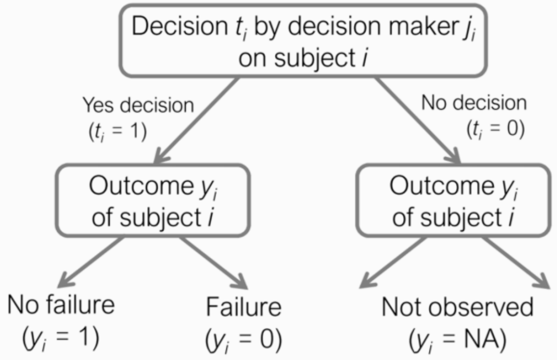


Figure 1: Selective labels problem.

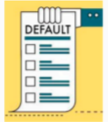
Barocas, S., & Selbst, A. D. (2016). **Big data's disparate impact.** Calif. L. Rev., 104, 671
Manuel Gomez Rodriguez et al. (2020). **Human-Centric Machine Learning Feedback loops, Human-AI Collaboration and Strategic Behavior** [Link]. Web
Corbett-Davies & Goel. (2018). **The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning**
Lakkaraju, H. et al. (2017). **The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables.** 23rd SIGKDD

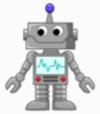
Examples of selective label

Decisions $d(\boldsymbol{x}) \in \{0,1\}$ $\sim$ **Decision policy** $\pi(d \,|\, \boldsymbol{x})$ **Informed by** $\rightarrow$ **Label predictions** $y \in \mathcal{Y}$ $\sim$ **Predictive model** $P_\theta(y \,|\, \boldsymbol{x})$

Features $\boldsymbol{x} \sim P(\boldsymbol{x})$

Aim to predict a ground truth label $y \sim P(y \,|\, \boldsymbol{x})$



Decisions $d(\boldsymbol{x}) \in \{0,1\}$ $\sim$ **Decision policy** $\pi(d \,|\, \boldsymbol{x})$ **Informed by** $\rightarrow$ **Label predictions** $y \in \mathcal{Y}$ $\sim$ **Predictive model** $P_\theta(y \,|\, \boldsymbol{x})$

Individual is rejected

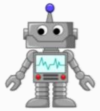
Individual receives loan

$\mathcal{Y} = \{0,1\}$

Individual defaults

Individual pays back



Decisions $d(\boldsymbol{x}) \in \{0,1\}$ $\sim$ **Decision policy** $\pi(d \,|\, \boldsymbol{x})$ **Informed by** $\rightarrow$ **Label predictions** $y \in \mathcal{Y}$ $\sim$ **Predictive model** $P_\theta(y \,|\, \boldsymbol{x})$

Individual remains jailed

Individual is released

$\mathcal{Y} = \{0,1\}$

Individual reoffends

Individual does not reoffend

Manuel Gomez Rodriguez et al. (2020). **Human-Centric Machine Learning Feedback loops, Human-AI Collaboration and Strategic Behavior** [Link]. Web

Sources of Bias – Algorithm

- The **automated** nature of modern ML
  - Millions of automated data-transformations to get a tiny improvement in predictive performance
  - Don’t carefully review the selected variables → surrogate variables and proxy discrimination
- **Overfitting** and **hyperparameter tunning** can amplify biases
- **Opaqueness** and **lack of interpretability** of complex ML algorithms
  - If one can identify the input-output relationships → easier to isolate potential algorithmic bias

- **Inherent biases in programmers** conveyed to the algorithm
- **Unexpected decisions** in traditional programming
  - Deliveroo riders affected by the ranking algorithm → Reliability index

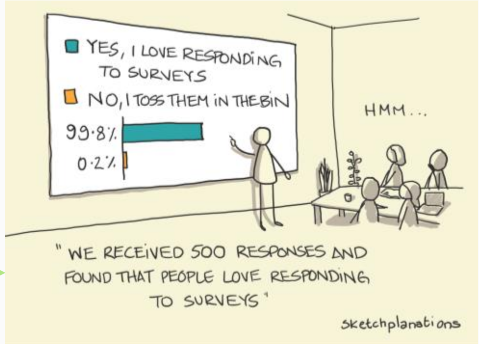
Personal and protected reasons → Shift Cancelation/ Acceptation → Reliability index → Ranking of good riders → Offered Shifts

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671
Mehrabi, N., et al. (2021). **A survey on bias and fairness in machine learning**. ACM Computing Surveys (CSUR), 54(6), 1-35
Jonathan Keane (2021). Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules. Web: Forbes

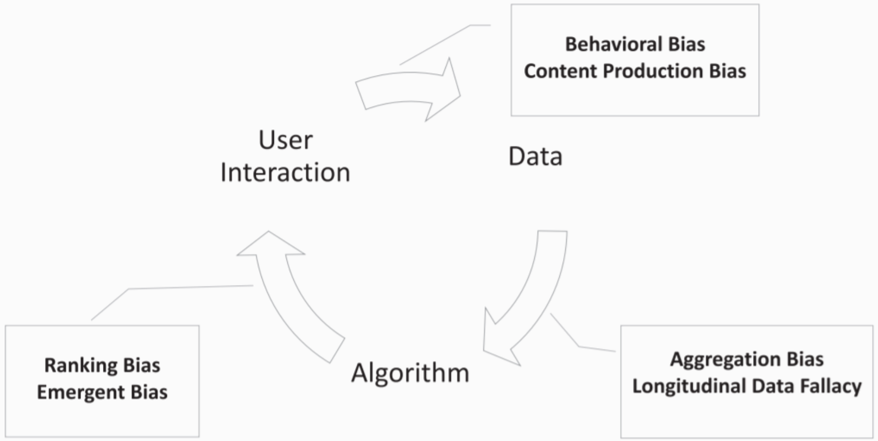
Sources of Bias – By interaction

- **Data to Algorithm**
  - Measurement Bias
  - Omitted Variable Bias
  - Representation Bias
  - Aggregation Bias
    - E.g., Sympson paradox
  - Sampling Bias
  - Longitudinal Data Fallacy
  - Linking Bias
  - Proxie

- **User to Data**
  - Historical Bias
  - Population Bias
  - Self-selection Bias
  - Social Bias
  - Behavioral Bias
  - Survivorship bias
  - Temporal Bias
  - Content production bias



- **Algorithm to User**
  - Algorithmic Bias
  - User Interaction Bias - Ranking
  - Popularity Bias
  - Emergent Bias
  - Evaluation Bias



Mehrabi, N., et al. (2021). **A survey on bias and fairness in machine learning**. ACM Computing Surveys (CSUR), 54(6), 1-35
Ricardo Baeza-Yates. 2018. **Bias on the web**. Commun. ACM 61, 6

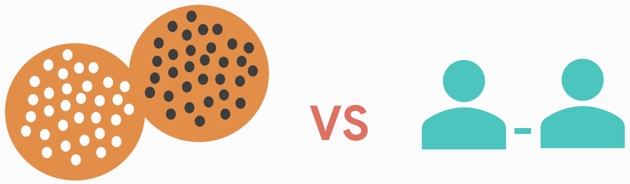
Fairness definitios and metrics

$P(S{=}s \mid A{=}a) = P(S{=}s \mid A{=}b)$

Several notions of fairness already exist in the literature

Recap: Algorithmic Fairness

- Algorithmic Fairness deals with the problem of developing AI-based systems able to treat:
  - **Subgroups in the population equally** → **Group fairness**
  - **Similar individuals in a similar way** → **Individual Fairness**
  - Other newer approaches



- Subgroups → determined by means of sensitive attributes, considered for decisions
  - *Gender*, *incomes*, *ethnicity*, and *sexual* or *political orientation* and so on

- Ensure that the outputs of a model DO NOT depend on sensitive attributes
  - *In a way that is considered unfair - differences due to such traits cannot be reasonably justified*

$$F(X) = R,\ \ A \in X \rightarrow \mathbf{R} \perp \mathbf{A}$$

  - Many approaches: properties of the model outputs with respect to the sensitive attributes

How do we define equally?

How we define similar?

Decision Rules: Classification

- Each individual has a set of features:
  - $x_i \in \mathbb{R}^p$
- x can be partitioned into protected and unprotected features:
  - $x = (x_u, x_p)$
  - Set of protected features: $\boldsymbol{A} \in X$ → different A values leads to different protected groups
- Target of prediction
  - $y \in \{0, 1\}$
- Training samples
  - $D = \{(x_i, y_i)\}_i^N$
- Random variables X and Y that take on values $\boldsymbol{X} = x$ and $\boldsymbol{Y} = y$ for an individual drawn randomly from the population of interest
- Binary classification
  - $f: \mathbb{R}^p \rightarrow \{0, 1\}$, where $\hat{y} = f(x)$ or, in population level $\hat{Y} = f(X)$
- Risk score
  - True risk score: $\mathrm{r(x) = Pr(Y = 1|X = x)}$
  - Model approximation of risk score $\mathrm{s(x)}$ instead of binary decision and $\boldsymbol{d(x) = 1\ iff\ s(x) > t}$
  - $\mathrm{R{=}E[Y|X]}$

In binary classification → probability of decision S

Confusion matrix

Event	Condition	Notion $P(event \mid condition)$
$\hat{Y}=0$	$Y=0$	True Negative rate
$\hat{Y}=1$	$Y=0$	False Positive rate
$\hat{Y}=0$	$Y=1$	False Negative rate
$\hat{Y}=1$	$Y=1$	True Positive rate

Classical clf criteria

Event	Condition	Notion $P(event \mid condition)$
$Y=0$	$\hat{Y}=0$	Positive predicted value
$Y=1$	$\hat{Y}=1$	Negative predicted value

Additional clf criteria

		Predicted Label		
		$\hat{y}=1$	$\hat{y}=-1$	
True Label	$y=1$	True positive	False negative	$P(\hat{y} \neq y \mid y=1)$ False Negative Rate
	$y=-1$	False positive	True negative	$P(\hat{y} \neq y \mid y=-1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y}=1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y}=-1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

		Predicted Label	
		Positive	Negative
True Label	Positive	**True Positives** $PPV = \frac{TP}{TP+FP}$ $TPR = \frac{TP}{TP+FN}$	**False Negative** $FOR = \frac{FN}{FN+TN}$ $FNR = \frac{FN}{FN+TP}$
	Negative	**False Positive** $FDR = \frac{FP}{FP+TP}$ $FPR = \frac{FP}{FP+TN}$	**True Negatives** $NPV = \frac{TN}{TN+FN}$ $TNR = \frac{TN}{TN+FP}$

Confusion matrix allow us to go further accuracy in error explanations related with joint distributions of $(X, \hat{Y}, Y)$

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. Nips tutorial, 1, 2017
Zafar, M. et al. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. 26th WWW.
Verma, S., & Rubin, J. (2018). Fairness definitions explained. In 2018 ieee/acm fairware. IEEE.

More confusion matrix measures

$$Pr(\hat{Y} = y | Y = y)$$
$$Pr(Y = y | \hat{Y} = y)$$

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP), hit**	**False negative (FN), type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
Actual condition	**Negative (N)**	**False positive (FP), type I error, false alarm, overestimation**	**True negative (TN), correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

- Confusion matrix allow us to go further accuracy in error explanations related with joint distributions of $(X, \hat{Y}, Y)$
- However, it may seem quite unmanageable to try all possible combinations
- How do we leverage all this measures for fairness? → Add sensitive attribute to conditional probabilities

Wikipedia. Precision and recall. https://en.wikipedia.org/wiki/Precision_and_recall

Considering a binary protected attribute $A \in \{a, b\}$ and $\hat{Y} = d$

Group fairness: main definitions

Predicted Outcome ($\hat{Y}$) → A ⊥ S

- **Demographic parity** [1] → A ⊥ S (independence)

  P(d=1|A=a) = P(d=1|A=b)

Predicted ($\hat{Y}$) and Actual Outcomes (d)

- **Predictive parity** [2] - *Same PPV* → A ⊥ Y | S (sufficiency)

  P(Y=1 | d=1, A=a) = P(Y=1 | d=1 , A=b)

- Predictive equality - *Same FPR [TNR]*

  P(d=1 | Y=0, A=a) = P(d=1 | Y=0, A=b)

- **Equal opportunity** – *Same FNR [TPR]*

  P(d=0 | Y=1, A=a) = P(d=0| Y=1, A=b)

- **Equalized odds** [3]– *same TPR and FPR* → A ⊥ S | Y (separation)

  P(d=1 | Y=i , A=a) = P(d=1 | Y=i, A=b), ∀ i ∈ {0, 1}

- Conditional use accuracy equality – *same accuracy for G*

  P(Y=1 | d=1, A=a) = P(Y=1 | d=1, A=b) ∧
  P(Y=0 | d=0, A=a) = P(Y=0 | d= 0, A=b)

- Overall accuracy equality – general accuracy

  P(d=Y, A=a) = P(d=Y, A=b).

- Treatment equality – same ratio of errors.

  (FN/FP)f=(FN/FP)m.

Predicted Probabilities (S) and Actual Outcome (d) → A ⊥ Y | S

- **Calibration** – *predictive parity but with probabilities* → A ⊥ Y | S

  P(Y=1 | S=s, A=a)= P(Y=1 | S=s, A=b), ∀ s ∈ [0, 1]

- Well calibration

  P(Y=1 | S=s, A=a)= P(Y=1 | S=s, A=b) = s, ∀ s ∈ [0, 1]

- Balance for positive class - *equal average predicted S for actual positives*

  E(S | Y=1, A=a) = E(S | Y=1, A=b)

- Balance for negative class - *same average predicted S for actual negatives*

  E(S | Y=0, A=a) = E(S | Y=0, A=b)

ML model should behave equally, or at least similarly, no matter whether it is applied to one subgroup in the population or to another one

Example of incompatibility
If different base rate P(Y=1|A=a) ≠ P(Y=1|A=b)
and satisfies predictive parity
→ Cannot satisfy Equalized odds

Barocas, S., Hardt, M., & Narayanan, A. (2017). **Fairness in machine learning**. Nips tutorial, 1, 2017
Verma, S., & Rubin, J. (2018). **Fairness definitions explained**. In 2018 ieee/acm fairware. IEEE.
[1] Cynthia Dwork,et al. 2012. **Fairness Through Awareness**. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference
[2] Alexandra Chouldechova. 2016. **Fair Prediction with Disparate Impact**: A Study of Bias in Recidivism Prediction Instruments. Big Data.
[3] Moritz Hardt, Eric Price, and Nati Srebro. 2016. **Equality of Opportunity in Supervised Learning**. In Advances in Neural Information Processing Systems

Definition clarification: Formal criteria

P(d=[0,1] | Y=[0,1]) AND P(Y=[0,1] | d=[0,1])

P(D=*d* | Y=*y*, A=a)=P(D=*d* | Y=*y*, A=b)

D \ Y	0	1
0	Predictive equality	Equal opportunity
1	Predictive equality Equal odds	Equal opportunity Equal odds

Group fairness and conditional statistical parity

P(Y=*y* | D=*d* , A=a)=P(Y=*y* | D=*d*, A=b)

Y \ D	0	1
0	Conditional use acc	Predictive parity
1		Predictive parity conditional use acc

Overrall accuracy

Definition clarification: Formal criteria

"Many fairness criteria have been proposed over the years, each aiming to formalize different desiderata. We'll start by jumping directly into the formal definitions of three representative fairness criteria that relate to many of the proposals that have been made." (Hardt et al., Fairness in Machine Learning book, 2019)

$P(S\|A)$	$P(S\|Y,A)$	$P(Y\|S,A)$
Independence	*Separation*	*Sufficiency*
$S \perp A$	$S \perp A \mid Y$	$A \perp Y \mid S$

Demographic parity

P(d=1|A=a) = P(d=1|A=b)

Positive Predicted Ratio
Equal acceptance rate

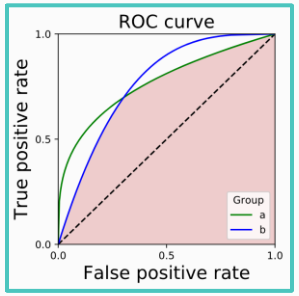
Equalized odds

P(d=1 | Y=i , A=a) = P(d=1 | Y=i, A=b), i ∈ 0, 1

Equal opportunity

P(d=0 | Y=1, A=a) = P(d=0| Y=1, A=b)

TPR - FPR
Equal error rates



Predictive Parity

P(Y=1 | d=1, A=a) = P(Y=1 | d=1 , A=b)

Calibration

P(Y=1 | S=s>t, A=a)= P(Y=1 | S=s>t, A=b)∀ t

PPV - NPV
Calibration by group

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. Nips tutorial, 1, 2017

Definition clarification: Formal criteria

List of demographic fairness criteria

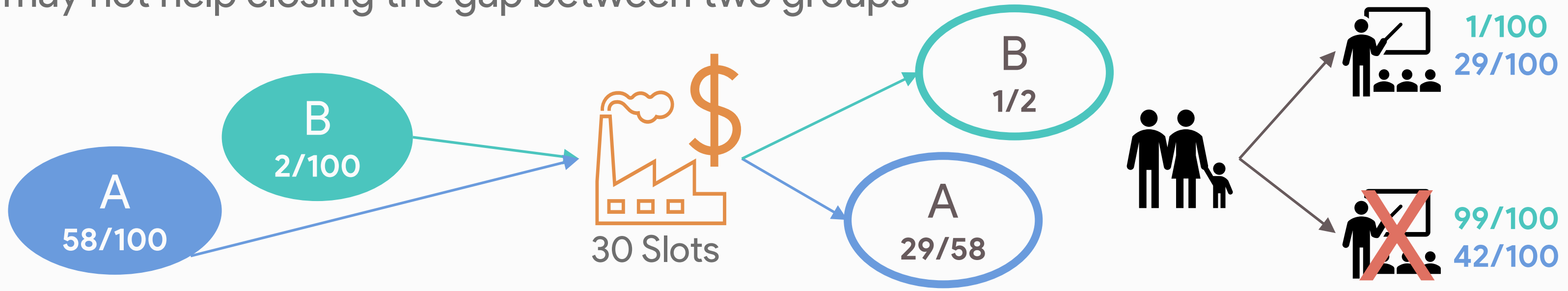
Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. Nips tutorial, 1, 2017

Group fairness gaps

- Proved that statistical definitions are insufficient [1, 2, 3, 4]
- Moreover, most valuable statistical metrics assume availability of actual, verified outcomes.
  - Problems with Selective label bias
- Similar individuals may not be treated equally for achieving measures of group fairness

- Demographic Parity [*Independence*]
  - Ignores any possible correlation between Y and A
  - E.g., Perfect predictor (S=Y) is not considered fair when base rates differ (i.e., $P[Y=1|A=a] \neq P[Y=1|A=b]$)
  - laziness: if we hire the qualified from one group and random people from the other group, we can still achieve demographic parity.
- Equalized Odds – Predictive Parity [*separation and sufficiency*]
  - It may not help closing the gap between two groups



[1] Richard Berka, Hoda Heidaric, Shahin Jabbaric, Michael Kearnsc, and Aaron Rothc. 2017. **Fairness in Criminal Justice Risk Assessments: The State of the Art**.
[2] Alexandra Chouldechova. 2016. **Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments**. Big Data (2016)
[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. **Fairness Through Awareness**. 3rd Innovations in Theoretical CS Conference.
[4] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. **Inherent Trade-Offs in the Fair Determination of Risk Scores**. In ITCS

Individual Fairness

- Group Fairness → *Similar individuals could not be treated equally due to calibrations across groups to achieve group fairness measures*
- **Individual Fairness → treating similar individuals similarly**
  - Difference between individuals similar to difference in predictions
  - More fine-grained than any group-notion fairness: it imposes restriction on for each pair of i.

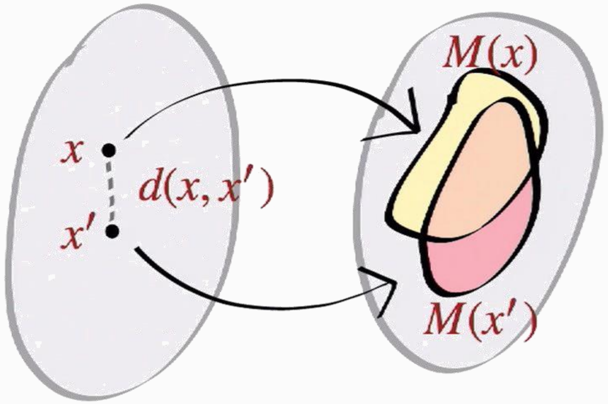
Our Dataset: $\boldsymbol{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_i^N$

Distance between $\boldsymbol{x_i}$pairs: $\boldsymbol{k}: \boldsymbol{V}\ \boldsymbol{x}\ \boldsymbol{V} \rightarrow \boldsymbol{R}$.

Mapping from $\boldsymbol{x_i}$ to probability distribution over outcomes $\boldsymbol{M}:\ \boldsymbol{V} \rightarrow \boldsymbol{\alpha A}$

Distance between distributions of outputs $\boldsymbol{D}$

Individual fairness $\boldsymbol{D(M(x), M(y))}\ \ =<\ \ \boldsymbol{k(x, y)}$



- ? How to define appropriate distance metrics for the specific problem and application?

Metric Learning | Graph Theory | Representation Learning

Dwork, C., et al.2012. **Fairness through awareness.** Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214-226
Verma, S., & Rubin, J. (2018). **Fairness definitions explained.** In 2018 ieee/acm fairware. IEEE.

Individual Fairness flaws

- Big expertise to establish a distance metric between individuals.
  - Metrics can still be implicit biased ☹
- Testing definitions relies on availability of “similar” individuals
  - Search space very large → e.g., the global population.
  - More work to narrow down the search space without impeding the accuracy

Graph Theory
Representation Learning
Semi/Self-Supervised Learning

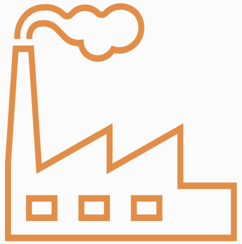
- Distance between data does not only depends on pairwise distances

  → Relationships among every all the data and topology *(cliques or communities on graphs)*
- Very difficult to find the proper metric (both d and M)
  - Specifically, M → unseen labels → Selective Labels / unobserved variables / substitutes labels

BSc / 1y.e.

MSc / 1y.e.

MSc / 0y.e.



Is individual A closer to B than C? How much?
→ very metric dependent d

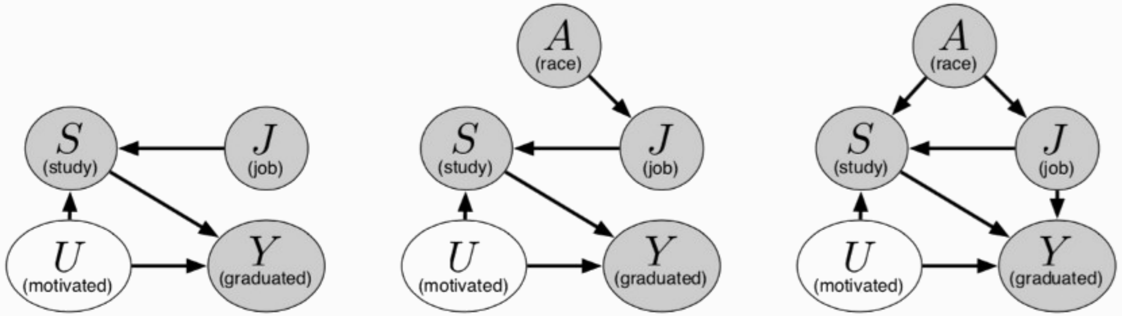
Is A closer to B than C regarding their predicted performance?
→ We don’t have real ground truth → Selective labels
→ Very metric dependent M

Dwork, C., et al.2012. **Fairness through awareness.** Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214-226
Kim, M. P., Reingold, O., & Rothblum, G. N. (2018). **Fairness through computationally-bounded awareness.** NIPS 2018

Counterfactual fairness

- *Group*
  - Observational fairness criteria
  - Cannot find the cause of the unfairness
- *Individual*
  - Limitation of finding the proper metric.

- <u>**Causality**</u>→ **Explaining the impact of bias via a <u>causal graph</u>**
  - Replacing A, other correlated features with it will also be influenced



<u>Causal graphs</u>: Acyclic graphs
- nodes representing attributes
- edges representing relationships

- Ideal idea? hard to reach a consensus in terms of
  - what the causal graph should look like?
  - which features to use even if we have such a graph?

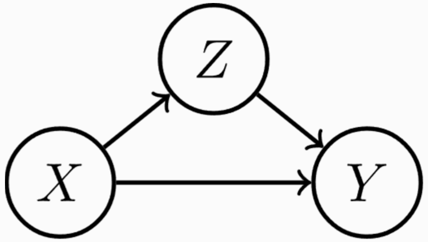
M.J. Kusner, J. Loftus, C. Russell and R. Silva, **Counterfactual fairness**, In Neural Information Processing Systems, (2017)
Barocas, S., Hardt, M., & Narayanan, A. (2017). **Fairness in machine learning.** Nips tutorial, 1, 2017
Shira Mitchell. 2018. **Reflection on quantitative fairness**. Web Book

Counterfactual fairness

- **Counterfactual** → *"Would I have been hired if I were non-black?" "Would I have avoided the traffic jam had I taken a different route this morning?"*
  - Decision does not depend on protected attribute

- The counterfactual $\boldsymbol{Y}_{\{X:=1, Z:=Z_{X:=0}\}}$ is the value that Y would obtain had X been set to 1 and had Z been set to the value Z would've assumed had X been set to 0

- Fair Causal graph → if Y don't depend on A, i.e., no A-Y way
  - Make decision only using non-descendants of A in the causal graph



- Difficult task of agreeing on which graph to build and validating it
- Impossible to test an existing classifier against **strict** causal definitions of fairness
- What should we do when not we are not able to built neither validate a causal graph?
  - **Counterfactual discrimination criteria → normative fairness criteria**

M.J. Kusner, J. Loftus, C. Russell and R. Silva, **Counterfactual fairness**, In Neural Information Processing Systems, (2017)
Barocas, S., Hardt, M., & Narayanan, A. (2017). **Fairness in machine learning.** Nips tutorial, 1, 2017
Shira Mitchell. 2018. **Reflection on quantitative fairness.** Web Book

Counterfactual fairness

- Notation of d(w), d(m) be the decision if the individual had been woman or men

- **Individual Counterfactual Fairness**

  $d_i(w) = d_i(m)$ for individual i and every other attribute remaining the same, i.e.,
  $P(\hat{Y}_{\{A \leftarrow a\}}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow b}(U) = y \mid X = x, A = a)$
  - negative answer to "*would the decision have been different if I were not black?*"

- **Counterfactual Demographic Parity**

  $E[d(w)] = E[d(m)]$ **i.e.,**
  $E[\hat{Y} \mid X = x, A = a] = E[\hat{Y} \mid X = x, A = b] \forall X \text{ and } \forall (a, b)$
  - negative answer to "*would the rates of hiring be different if everyone were black?*"

Related with *Conditional Demographic Parity*
$P(d = 1 \mid L = l, A = a) = P(d = 1 \mid L = l, A = b)$
which means $\hat{Y} \perp A \mid X$

- **Conditional Counterfactual Parity**

  $E[d(w) \mid X] = E[d(m) \mid X]$
  - "*would the rates of hiring be different if everyone were black?*" BUT stratified by some factors

- The easiest way to satisfy counterfactual demographic parity is : prediction only use non-descendants of A in the causal graph

M.J. Kusner, J. Loftus, C. Russell and R. Silva, **Counterfactual fairness**, In Neural Information Processing Systems, (2017)

Counterfactual in real world

"Race plays a significant role in admissions decisions. Consider the example of an Asian-American applicant who is male, is not disadvantaged, and has other characteristics that result in a 25% chance of admission. Simply changing the race of the applicant to white— and leaving all his other characteristics the same—would increase his chance of admission to 36%. Changing his race to Hispanic would increase his chance of admission to 77%. Changing his race to African-American would increase his chance of admission to 95%".

(150 Plaintiff's expert report of Peter S. Arcidiacono, Professor of Economics at Duke University)

- Logistic regression model against Harvard's past admissions decisions
- Conditional statistical parity is not satisfied

  $P(d=1 \mid L=l, A=a) = P(d=1 \mid L=l, A=a)$

150 Plaintiff's expert report of Peter S. Arcidiacono, Professor of Economics at Duke University

Fairness measurement in benchmarking dataset

- So, is the classifier fair? → Logistic regression on German Credit Dataset

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–

- Depends on the notion of fairness one wants to adopt.
  - More work is needed to clarify which definitions are appropriate to each particular situation

Context matters

German Credit Dataset. M. Lichman. 2013. UCI Machine Learning Repository. (2013). http://archive.ics. uci.edu/m
Verma, S., & Rubin, J. (2018). **Fairness definitions explained**. In 2018 ieee/acm fairware. IEEE. I

Summary of metrics

- Group Fairness
  - Independence, separation, sufficiency
  - Confusion matrix-related
  - Counterfactual parity
- Individual Fairness
  - Metrics
  - Individual counterfactual
- Counterfactual
  - Conceptually
  - Applied
- Many more…

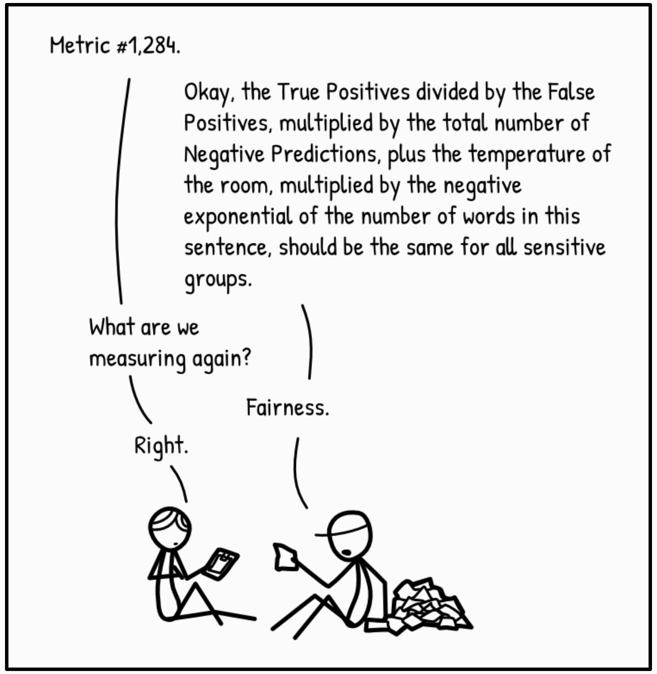


Table 1
A synthetic review of most of the notions of fairness.

Notion	Abbreviation	First Appeared
α-Protection	α-P	[159]
Indirect Discriminatory Measure	ELB	[72]
Decision Policy Discrimination	DPD	[131]
Prediction Dependency	PredD	[23]
Dataset Discrimination	DD	[97]
Discrimination Score	DS	[22]
Calders-Verwer Score	CVS	[22, 105]
Statistical Parity	SP	[50]
Mean Difference	MD	[24]
Area Under ROC Curve	AUC	[24]
Disparate Impact	DI	[56]
ϵ-Fairness	ϵ-F	[56]
η-Neutrality	η-N	[60]
Discrimination Correlation Indicator	DCI	[125]
Demographic Parity	DP	[76]
Equal Opportunity	EOp	[76]
Equal Odds	EOd	[76]
Fair Prediction Rule	FairPR	[124]
Indifference	Indiff	[94]
Total Causal Effect	TCE	[206]
Cross-Pair Group Fairness	CPGF	[14]
Hilbert-Schmidt Empirical Cross-Covariance	HSIC	[160]
Expected Statistical Parity	ESP	[37]
Expected Predictive Equality	EPE	[37]
Calibration	Calib	[37]
Balanced Loss	BL	[51]
False Positive Subgroup Fairness	FPSF	[107]
Proxy Discrimination	ProxD	[108]
Proxy Discrimination in Expectetion	PDE	[108]
P%-Rule	P-R	[113]
Normalised Disparate Impact	NDI	[137]
α-Discrimination	α-D	[189]
Value Unfairness	ValU	[194]
Absolute Unfairness	AbsU	[194]
Underestimation Unfairness	UeU	[194]
Overestimation Unfairness	OeU	[194]
Preferred Impact	PrefI	[199]
Preferred Treatment	PrefT	[199]
Disparate Mistreatment	DM	[197]
Absolute Value Difference	AVD	[13]
Squared Difference	SD	[13]
Balance	Bal	[28]
Relaxed Equal Odds with Calibration	REOC	[163]
Path Specific Effect	PSE	[143]
Natural Direct Effect	NDE	[143]
Mean Difference Discrimination Score	MDDS	[168]
k-way Discrimination Score	k-DS	[168]
Maximum Discrimination	MaxD	[168]
Discrimination In Prediction	DiscrP	[208]
Loss-Averse Statistical Parity	L-ASP	[5]
Loss-Averse Equal Opportunity	L-AEOp	[5]
Difference of Equal Opportunity	DEO	[33]
Hirschfeld-Gebelein-Rényi	HGR	[134]
Coefficient of Determination	Cod | R^2	[114]
Difference of Equal Opportunity	DEOp	[151]
Difference of Equal Odds	DEOd	[151]
Subgroup Risk	SR	[188]
Strong Demographic Parity	SDP	[91]
Strong Pairwise Demographic Disparity	SPDD	[91]

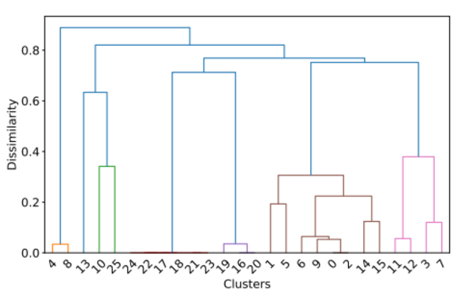
(*Continued*)

Table 1
(*Continued*)

Notion	Abbreviation	First Appeared
Group Fairness in Expectation	GFE	[59]
Prejudice Index	PI	[105]
Fair-Factorization	FF	[106]
Resilience to Random Bias	RRB	[58]
Normalised Discounted Difference	rND	[192]
Normalised Discounted Ratio	rRD	[192]
Normalised Discounted KL-divergence	rKL	[192]
Explanatory Conditional Discrimination	ECD	[210]
Expected Conditional Statistical Parity	ECSP	[37]
Individual Proxy Discrimination	IPD	[108]
Disparate Treatment	(DispT)	[197]
Disparity Amplification	DA	[78]
k-Neighbours Difference	k-ND	[126]
Fairness Lipschitz Property	FLP	[50]
Cross-Pair Individual Fairness	CPIF	[14]
Decision Boundary Covariance	DBC	[198]
Random Bias Individual Fairness	RBIF	[57]
Inconsistency Score	IS	[168]
(α, γ)-Approximately Metric-Fair	(α, γ)-AMF	[196]
Constant Relative Risk Aversion	CRRA	[81]
Rawlsian Equal Opportunity	R-EOP	[82]
Egalitarion Equal Opportunity	e-EOP	[82]
Generalised Entropy Index	GEI	[179]
Counterfactual Fairness	CF	[117]
ϵ, δ-Approximate Counterfactual Fairness	ϵ, δ-ACF	[171]
Counterfactual Direct Effect	CF-DE	[204]
Counterfactual Indirect Effect	CF-IE	[204]
Counterfactual Spurious Effect	CF-SE	[204]
Chebyshev Demographic Parity	CDP	[207]
Maximum Mean Discrepancy	MMD	[68]
Fairness Ramp-Constraint	FRC	[65]
δ-fairness	δ-F	[96]
Impartiality Score	IS	[94]
Formal Equality of Opportunity	FEO	[94]
Full Substantive Equality of Opportunity	F-SEO	[94]
Log-Linear Interaction	LLI	[190]
Markov Decision Fairness	MDF	[88]
Approximate-Choice Markov Decision Fairness	α-CF	[88]
Approximate-Action Markov Decision Fairness	α-AF	[88]
Indirect Influence	II	[1]
ϵ-Loss Fair	ϵ-LF	[49]
α-MultiCalibration	α-MC	[80]
Covariance Constraint	CC	[149]
Metric MultiFairness	MMC	[109]
ϵ-Loss General Fair	ϵ-LGF	[153]
Mutual Information	MI	[186]
Kullback-Leibler Divergence	KL-D	[186]
Wasserstein Distance	WD	[201]
Path Specific Counterfactual Fairness	PSCF	[26]

Lim Swee Kiat. Retrieved December 2021. Machines go Wrong. https://machinesgonewrong.com/fairness/
Oneto, L. (2020). Learning fair models and representations. Intelligenza Artificiale, 14(1), 125-152.DOI 10.3233/IA-190034
Castelnovo, A., Crupi, R., Greco, G., & Regoli, D. (2021). The zoo of Fairness metrics in Machine Learning. arXiv

Metrics clarification



- **Theory:** Formal criteria aforementioned:
  - A⊥S|X – A⊥S – A⊥S|Y – A⊥Y|S
- **Applied:** Majumder, S. et al (2021)
  - 26 classification metrics → 7 clusters
  - 4 dataset metrics → 3 clusters

RQ1: Do current fairness metrics agree with each other?

No → 51% agreement

RQ2: Can we group (cluster) fairness metrics based on similarity?

Yes → minimizing intra-cluster disagreement

RQ4: Can we achieve fairness based on all the metrics at the same time?

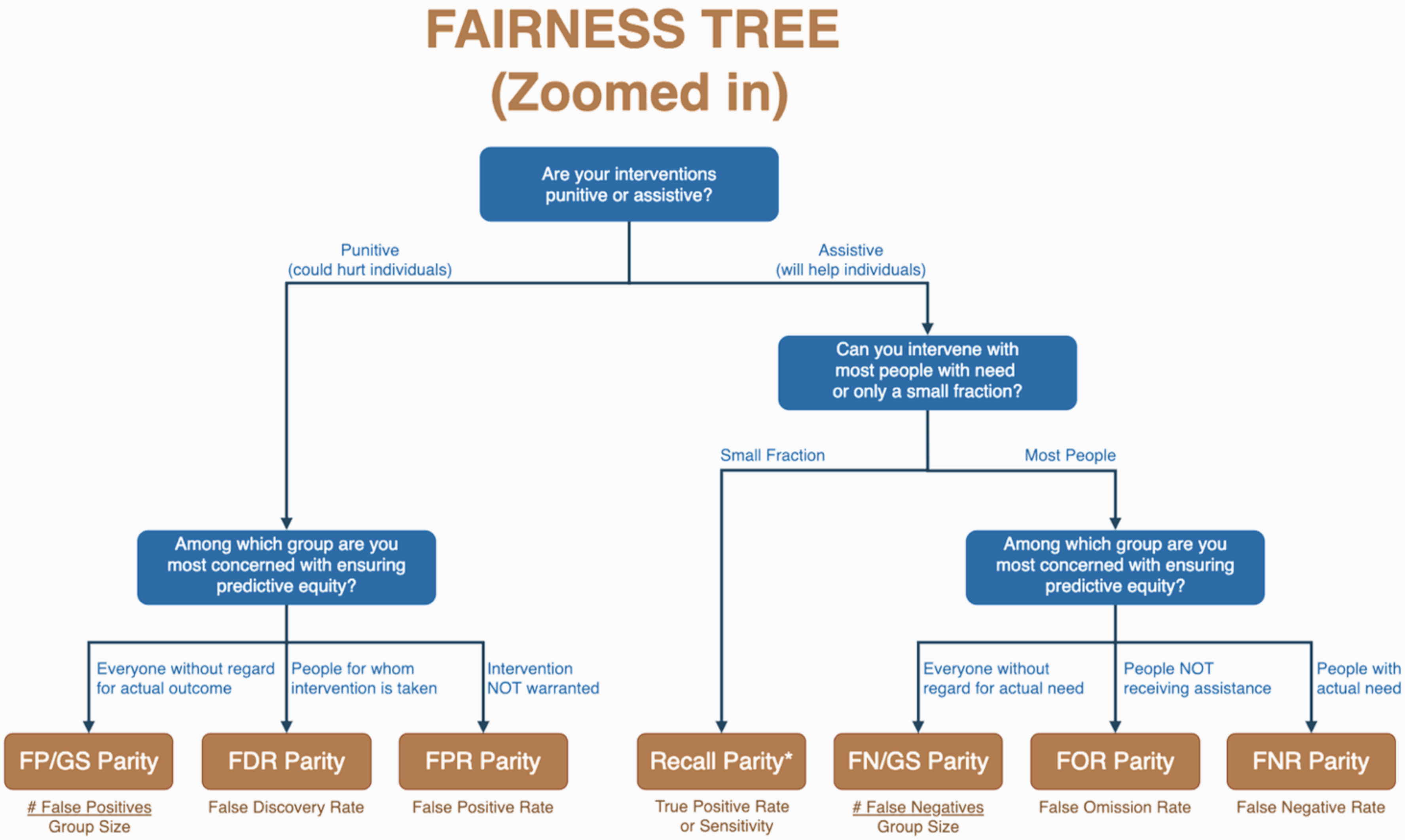
No. Each cluster and metric measure on thing, sometimes opposite

Again, choose depends on the context

Cluster Id	MID	Metrics	Datasets: Adult	Compas	German	Health	Bank	Student	Titanic	Metric Type
0	C3	false_omission_rate_difference	Unfair	Fair	Fair	Unfair	Fair	Fair	Unfair	Mis-classification
0	C7	false_omission_rate_ratio	Unfair	Fair	Fair	Unfair	Fair	Unfair	Unfair	
0	C11	error_rate_difference	Unfair	Fair	Fair	Unfair	Fair	Fair	Fair	
0	C12	error_rate_ratio	Unfair	Fair	Fair	Unfair	Fair	Fair	Fair	
		Percentage of agreement	**100%**	**100%**	**100%**	**100%**	**100%**	**75%**	**50%**	
1	C10	average_abs_odds_difference	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	Differential Fairness
1	C25	differential_fairness_bias_amplification	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
		Percentage of agreement	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	
2	C16	generalized_entropy_index	Fair	Unfair	Fair	Fair	Fair	Fair	Unfair	Individual Fairness
2	C19	theil_index	Unfair	Unfair	Fair	Unfair	Unfair	Fair	Unfair	
2	C20	coefficient_of_variation	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	
		Percentage of agreement	**67%**	**100%**	**67%**	**67%**	**67%**	**67%**	**100%**	
3	C4	false_discovery_rate_difference	Fair	Fair	Fair	Fair	Fair	Fair	Unfair	Mis-classification
3	C8	false_discovery_rate_ratio	Fair	Fair	Fair	Fair	Fair	Unfair	Unfair	
		Percentage of agreement	**100%**	**100%**	**100%**	**65%**	**100%**	**50%**	**100%**	
4	C0	true_positive_rate_difference	Unfair	Unfair	Fair	Unfair	Unfair	Fair	Unfair	Confusion Matrix Based Group Fairness
4	C1	false_positive_rate_difference	Fair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
4	C2	false_negative_rate_difference	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
4	C5	false_positive_rate_ratio	Fair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
4	C6	false_negative_rate_ratio	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	
4	C9	average_odds_difference	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
4	C14	disparate_impact	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	
4	C15	statistical_parity_difference	Unfair	Unfair	Unfair	Unfair	Unfair	Fair	Unfair	
		Percentage of agreement	**75%**	**100%**	**88%**	**100%**	**100%**	**75%**	**100%**	
5	C17	between_all_groups_generalized_entropy_index	Fair	Fair	Fair	Fair	Fair	Fair	Fair	Between Group Individual Fairness
5	C18	between_group_generalized_entropy_index	Fair	Fair	Fair	Fair	Fair	Fair	Fair	
5	C21	between_group_theil_index	Fair	Fair	Fair	Fair	Fair	Fair	Fair	
5	C22	between_group_coefficient_of_variation	Fair	Fair	Fair	Fair	Fair	Fair	Unfair	
5	C23	between_all_groups_theil_index	Fair	Fair	Fair	Fair	Fair	Fair	Fair	
5	C24	between_all_groups_coefficient_of_variation	Fair	Fair	Fair	Fair	Fair	Fair	Unfair	
		Percentage of agreement	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**67%**	
6	C13	selection_rate	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Unfair	Intermediate Metric
		Percentage of agreement	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	
		Percentage of metrics marking dataset as unfair	**58%**	**54%**	**34%**	**65%**	**50%**	**23%**	**77%**	

Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., & Menzies, T. (2021). Fair Enough: Searching for Sufficient Measures of Fairness. preprint arXiv:2110.13029.

Metrics clarification



Saleiro, P., et al. (2018). Aequitas: A bias and fairness audit toolkit. arXiv:1811.05577
http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

Impossibility Theorem

Why different definitions are not compatible?

Inherent Trade-off of fairness

Fairness limitations

- Accuracy VS Fairness
- Group Fairness Impossibility Theorem
- Group VS Individual

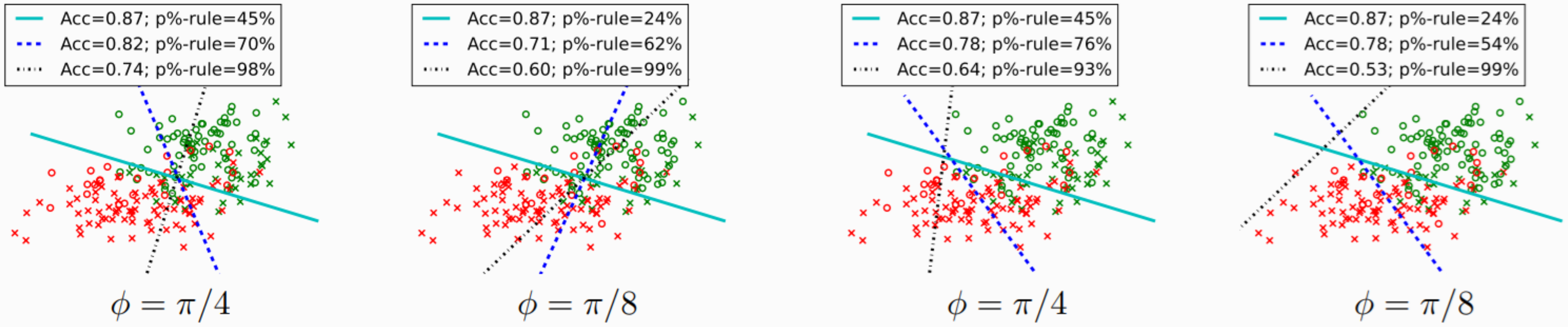
Accuracy vs Fairness Tradeoff

Impose constraints on the accuracy with fairness metrices leads to not aligned objectives

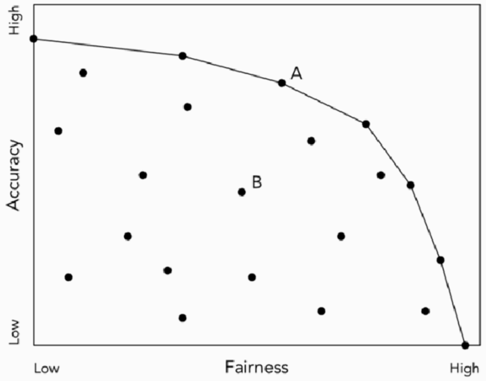
Tradeoff depends on how “similar” Y and A are → e.g., if aligned, then linear penalty

The more aligned, the more one will penalize the other

We will have solutions in the pareto front



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



$$p\%rule = \min\left(\frac{P\{\hat{Y} = 1 \mid A = a\}}{P\{\hat{Y} = 1 \mid A = b\}}, \frac{P\{\hat{Y} = 1 \mid A = b\}}{P\{\hat{Y} = 1 \mid A = a\}}\right) \geq \frac{p}{100}$$

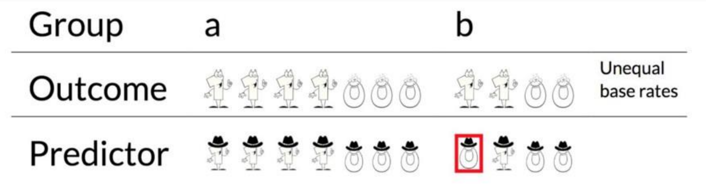
Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. IJIS, 36(4), 1619-1643.
Menon, A. K., & Williamson, R. C. (2018, January). The cost of fairness in binary classification. In Conference on Fairness, Accountability and Transparency (pp. 107-118). PMLR
Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, April). **Fairness constraints: Mechanisms for fair classification.** In Artificial Intelligence and Statistics . PMLR.

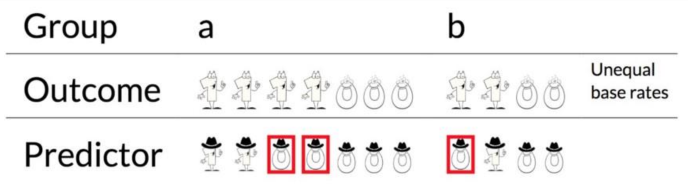
Formal criteria's impossibility theorem

Independence	*Separation*	*Sufficiency*
A⊥S	A⊥S | Y	A⊥Y | S

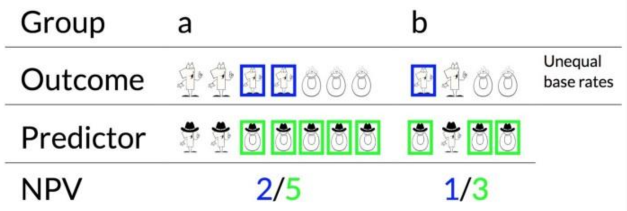
¬⊥ → dependent || ⊥→ Independent
Demographic Parity - DP
Equalized odds - EO
Predictive Parity - PP

- Independence vs sufficiency – DP vs PP
  - If **A¬⊥Y** → either DP or PP, but NOT BOTH
- Independence vs Separation – DP vs EO
  - If **Y¬⊥ A** && **Y¬⊥ S** → either DP or EO, but NOT BOTH
- Separation vs sufficiency – EO vs PP
  - If $P(a, s, y) > 0 \; \forall \; AxSxY$ (all events in the joint distribution of have positive probability) AND
  - If **A¬⊥Y**→ either EO or PP, but NOT BOTH
  - If predictor satisfy EO, PP requires equal PPV, and therefore need equal base rates → Not usually happen
  - i.e., If different base rates P(Y=1 | A=a) ≠ P(Y=1 | A=b) → either EO or PP, but NOT BOTH



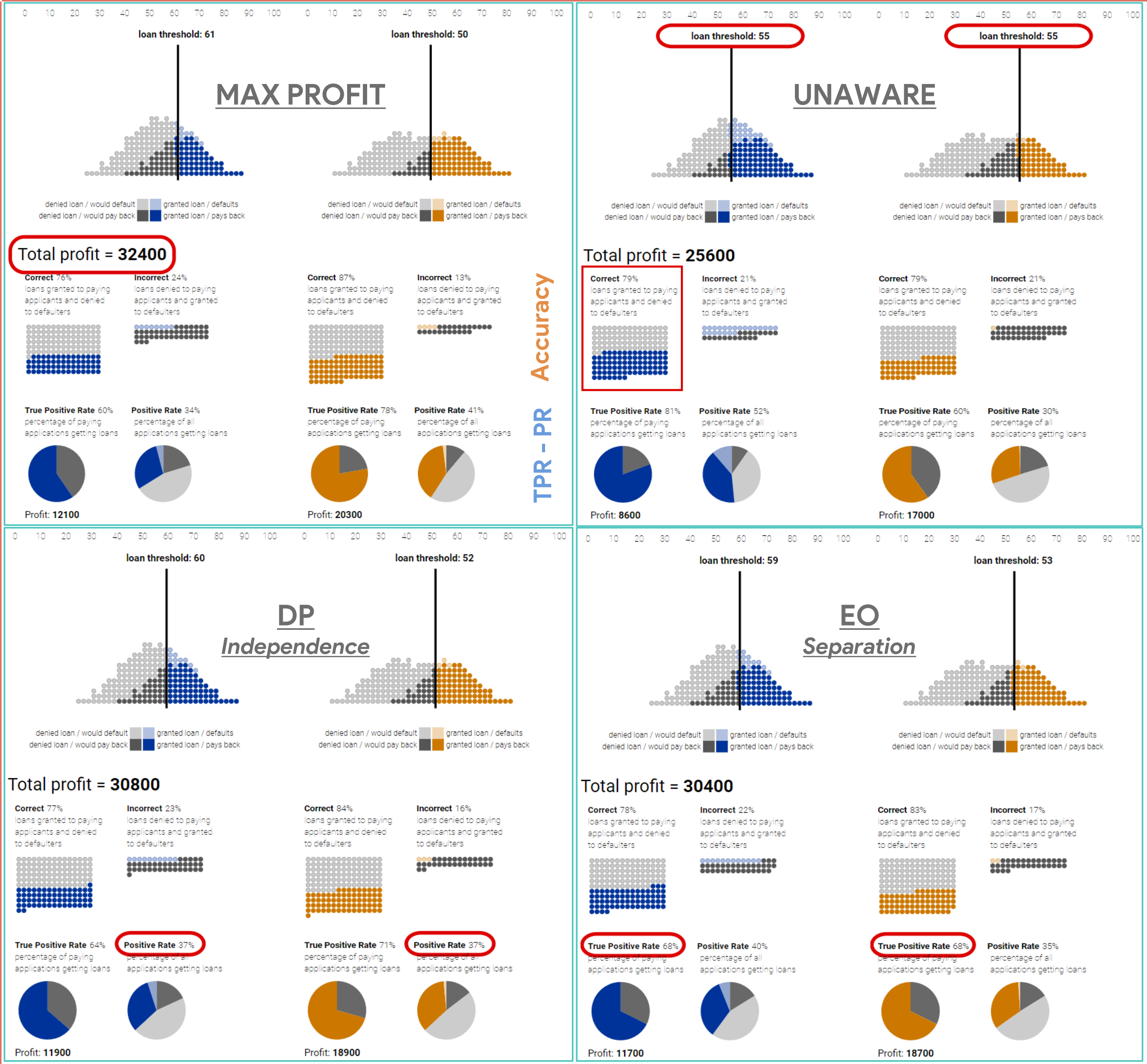


Make 2 FP to achieve EO
Equal TPR and TNR between groups



Negative Predictive Parity violated
Not possible to preserve NPV without sacrificing EO/PP

J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, Innovations in Theoretical Computer Science Conference
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163
Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. Nips tutorial, 1, 2017



Formal criteria's relationship

$$\underbrace{P(y|s,A)}_{\text{Predictive Parity}} \times \underbrace{P(s|A)}_{\text{Demographic Parity}} = \underbrace{P(s|y,A)}_{\text{Equalized odds}} \times \underbrace{P(y|A)}_{\text{Base Rate}}$$

Proofs based on Positive Predicted Value, TPR and FPR

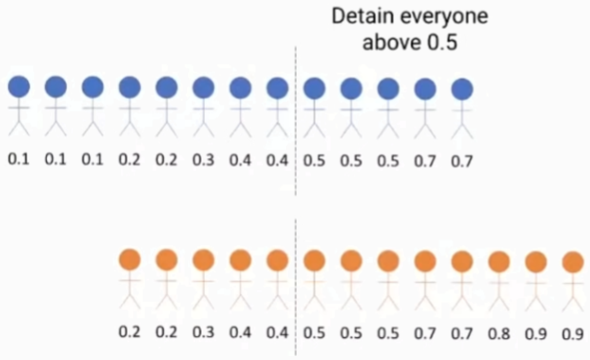
If **unequal base rates** && **not perfect classifier**
→ **Sufficiency** implies that **Error parity Fails**

Loan granting: 2 groups with different base rates

- Maximize profit → violate TPR and PR
- Unaware → orange gets fewer loans - also violate TPR and PR
- Demographic Parity (PR) → Violates TPR (EO)
- Equalized odds (EO) → Violates PR (DP)

Martin Wattenberg, Fernanda Viégas, and Moritz Hardt Attacking discrimination with smarter ML.
https://research.google.com/bigpicture/attacking-discrimination-in-ml/

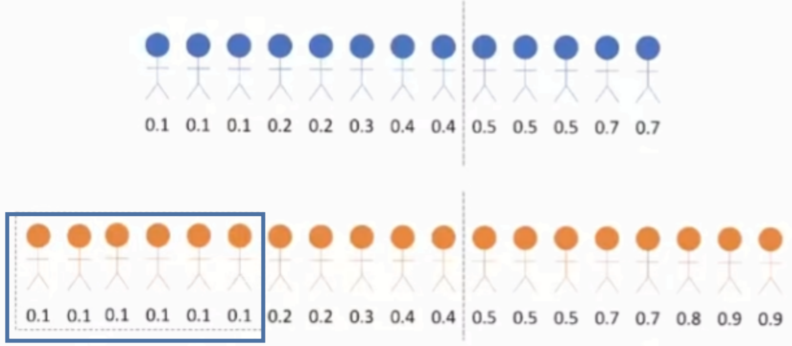
Metrics not sufficient on their own



Detention rate	False pos. rate
38%	25%
61%	42%

— Impendence and error rate parity [EO, FPR] **violated**

?



Statistical fairness criteria on their own cannot be a proof of fairness, just a piece of it

Detention rate	False pos. rate
38%	25%
~~61%~~ 42%	~~42%~~ 26%

Garg, P., Villasenor, J., & Foggo, V. (2020). **Fairness metrics: A comparative analysis**. In 2020 IEEE Big Data. IEEE.
del Barrio, E., Gordaliza, P., & Loubes, J. M. (2020). Review of mathematical frameworks for fairness in machine learning. arXiv
Castelnovo, A., Crupi, R., Greco, G., & Regoli, D. (2021). The zoo of Fairness metrics in Machine Learning. arXiv preprint arXiv:2106.00467
Chiappa, S., & Isaac, W. S. (2018). A causal bayesian networks viewpoint on fairness. In IFIP International Summer School on Privacy and Identity Management. Springer, Cham.Oneto, L., & Chiappa, S. (2020). Fairness in Machine Learning. ArXiv, abs/2012.15816.
Martin Wattenberg, Fernanda Viégas, and Moritz Hardt Attacking discrimination with smarter ML. https://research.google.com/bigpicture/attacking-discrimination-in-ml/
Moritz Hardt - MLSS 2020, Tübingen. https://youtu.be/Igq_S_7IfOU?t=4056
http://www-student.cse.buffalo.edu/~atri/algo-and-society/support/notes/fairness/index.html

Imposing fairness

How to plug chosen fairness definition into the training on ML algorithms?

How to satisfy Fairness criteria

Pre-processing

- From feature space to a **representation**→ Independence S⊥A
- Model learned from this representation will be fair [*Data processing inequality* Information Theory]
- Model agnostic
- Information loss in latent space

In-processing

- **Fairness constraints** in the optimization process
- Powerful → fairness during the optimization process
- Loss of generality → each type of model and specific task uses its own regularize

Post-processing

- Taking a trained classifier → adjust it depending on the sensitive attribute and randomness
- independence is achieved
- Works for black-box models and no re-training needed
- Useful when no access to training data, complex-no access to training pipeline
- Not that efficient due to the same reasons

Lots of them… again

- **Method family**
  - Pre
  - In
  - Post
- Task
  - Binary classification
  - Multiclassification
  - Regression
- Protected attribute
  - Binary
  - Categorical
  - Numerical

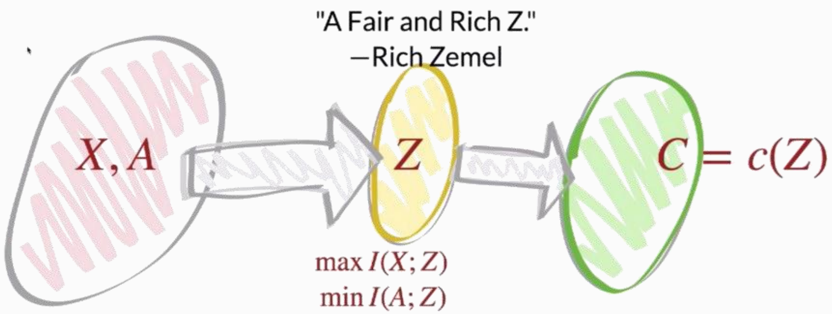
Table 2
A synthetic review of most of the papers available in the literature

Paper	Method Family	Task	Protected Attribute	Notion of Fairness	Theoretical Results	Experimental Results	Comparison Against	Code Available
[16]	PreP, InP	BC, MC, R	B, C	DP, EOp		✓		
[164]	InP	BC	B, C	MMD		✓	[76, 197]	
[194]	InP	BC, MC, R	B, C	ValU, AbsU, UeU, OeU	✓			
[137]	PreP	BC, MC	B	SP, NDI	✓	✓		
[149]	InP	BC	B, C	CC	✓	✓	[197]	
[163]	PostP	BC	B	REOC	✓	✓	[76, 197]	
[1]	PostP	BC, MC	B, C, N	II		✓	[83]	✓
[25]	PreP	BC	B, C	α-P, DP	✓	✓	[202]	
[37]	InP	BC	B, C	ESP, ECSP, EPE	✓			
[107]	InP	BC	B, C	SP, FPSF	✓	✓	[2]	✓
[80]	InP	BC	B, C, N	α-MC	✓			✓
[69]	InP	BC	B	ESP, EPE				
[208]	PreP, PostP	BC	B	DiscrP	✓	✓	[76, 197]	
[4]	InP	BC, MC	B, C	EOd	✓			✓
[49]	PreP, InP	BC	B, C	ϵ-LF	✓	✓	[76, 197]	✓
[2]	PostP	BC	B, C	DP, EOd	✓	✓	[76, 99]	
[64]	InP	MAB	B, C	FLP				
[78]	InP	BC, MC, R	B, C, N	DA		✓		
[109]	PostP	BC	B, C, N	MMC				✓
[129]	InP	BC	B, C	EOd	✓			
[128]	PreP	BC	B	DP, EOp, EOd	✓	✓	[52]	✓
[145]	InP	BC, MC	B, C	DP, EOd	✓	✓		
[203]	PreP, InP	BC, MC, R	B, C	DP, EOd, EOp	✓	✓	[16]	
[133]	PreP, InP	BC, MC	B, C	DP, EOd		✓	[2, 128, 145, 198, 203]	
[143]	InP	BC, MC, R	B, C	PSE, NDE		✓		
[168]	PostP	BC, MC, R	B, C, N	MDDS, k-DS, MaxD, IS		✓	[97, 103, 202]	
[167]	PreP, InP	BC	B	MDDS, IS		✓	[52, 103, 123, 168, 202]	
[196]	InP	BC, MC, R	B, C, N	(α, γ)-AMF	✓			
[185]	Prep, InP	BC, MC, R	B, C	DP, EOd		✓		
[67]	PreP	BC	B	DI	✓	✓	[56]	
[63]	PreP, InP	BC, MC, R	B, C	MI, EOd	✓			
[5]	PostP	BC	B	L-ASP, L-AEOP		✓		
[39]	InP	BC, MC	B, C	DP, EOp	✓	✓		✓
[40]	InP	BC, MC	B, C	DP, EOp	✓	✓		✓
[81]	InP	BC, MC, R	B, C, N	CRRA		✓		
[114]	InP	BC, MC, R	B, C, N	CoD	✓	✓		
[186]	PreP	BC, MC, R	B, C, N	MI, KL-D		✓		
[178]	PreP	BC, MC, R	B, C, N	MI	✓	✓		
[82]	InP	BC, MC, R	B, C	R-EOP, e-EOP		✓	[81]	
[179]	InP	BC, MC, R	B, C	GEI	✓	✓	[197]	
[201]	PreP	BC, MC, R	B, C	WD	✓	✓		
[59]	InP, PostP	BC, MC, R	B, C	GFE	✓	✓		
[144]	InP	BC, MC, R	B, C	PSE	✓	✓		
[33]	PostP	BC	B	DEO	✓	✓	[49, 76, 197]	
[86]	InP	BC	B	ϵ-LF	✓			
[110]	PostP	BC	B	α-MC	✓	✓		✓
[134]	InP	BC, MC, R	B, C, N	HGR	✓	✓	[13, 49]	
[147]	PostP	BC	B	EOd, EOp		✓	[76]	
[153]	PreP, InP	BC, MC, R	B, C, N	ϵ-LGF	✓	✓	[197, 198]	
[151]	InP	BC	B, C	DEOp, DEOd		✓		
[152]	PreP, InP	BC, MC, R	B, C	DP	✓	✓	[52, 128]	
[188]	InP	BC, MC, R	B, C, N	SR	✓	✓	[49]	
[26]	InP	BC, MC, R	B, C	PSCF, MMD	✓	✓		
[91]	InP, PostP	BC	B, C, R	SDP, SPDD, WD	✓	✓	[76]	
[200]	InP	BC, MC	B, C	DBC, DI, DM		✓	[37, 49, 76, 98, 103]	
[138]	PreP	BC	B	SP, DI, FLP	✓			
[89]	PostP, InPro	BC, MC	B	EOd	✓	✓		
[97]	PreP	BC	B, C	DD		✓		✓
[158]	PostP	BC	B, C	α-P		✓		
[23]	PreP	BC	B, C	PredD		✓	[97]	
[22]	PreP, InP, PostP	BC	B	DS (CVS)		✓		✓
[100]	InP, PostP	BC	B	DS		✓	[22, 23, 97]	✓
[98]	PreP	BC	B, C	DS		✓	[23, 97]	✓
[126]	PreP	BC, MC	B, C, N	k-ND		✓		✓
[210]	PreP	BC	B	ECD		✓	[23, 98]	✓
[72]	PreP	BC	B, C	ELB		✓		✓
[105]	InP	BC	B, C	PI		✓	[22]	
[50]	InP	BC	B,C,N	FLP, SP	✓			
[71]	PreP	BC	B, C	ELB		✓		✓
[101]	InP	BC	B, C	DS		✓	[22, 99, 100]	✓
[99]	PreP	BC	B	DS	✓	✓	[22, 100]	✓
[131]	PreP	BC	B, C	DPD		✓		✓
[73]	PostP	BC	B, C	α-P		✓		✓
[24]	InP	BC, MC, R	B, C	MD, AUC		✓		
[106]	InP	BC	B	FF		✓	[22]	
[202]	PreP, InP	BC	B	SP		✓	[97, 103]	
[102]	PreP, PostP	BC	B	ECD		✓	[23, 98]	✓
[132]	PreP	BC	B, C	α-P		✓	[22]	✓
[74]	PreP	BC	B, C	α-P		✓		✓
[56]	PreP	BC	B	DI, ϵ-F	✓	✓	[97, 103, 202]	
[123]	PreP	BC, MC, R	B, C, N	DP, MMD		✓	[202]	
[52]	PreP, InP	BC	B	SP		✓		✓
[60]	InP	BC, MC, R	B, C, N	η-N	✓	✓	[22, 104]	
[75]	PreP	BC	B, C	α-P	✓	✓		✓
[125]	InP	BC, MC	B, C	DCI		✓		✓
[55]	PreP	BC	B, C	SP, DI		✓	[202]	✓
[57]	PreP, InP	BC	B	DP, RBIF		✓	[202]	
[76]	PostP	BC	B, C	EOp, EOd	✓	✓		
[65]	InP	BC	B, C	FRC		✓	[198]	✓
[96]	InP	BC	B, C	δ-F	✓			✓
[95]	InP	BC	B, C	δ-F	✓	✓		✓
[58]	PostP	BC, MC, R	B, C, N	RRB	✓	✓	[97, 100, 202]	
[124]	PreP	BC, MC, R	B, C, N	FairPR		✓		
[94]	InP	BC, MC, R	B, C, N	FEO, F-SEO		✓		
[190]	PostP	BC, MC	B, C	LLI		✓		✓
[206]	PreP	BC	B, C	TFE		✓	[56, 210]	✓
[198]	InP	BC	B, C	DBC		✓	[98, 103]	
[51]	InP	BC, MC	B, C	BL	✓	✓		
[14]	InP	BC, MC, R	B	CPIF, CPGF		✓		
[88]	InP	BC	B, C	MDF, α-CF, α-AF	✓			
[93]	PreP	BC	B, C, R	FairPR		✓		✓
[108]	InP	BC, MC	B, C	ProxD, PDE, IPD				
[113]	PreP, InP	BC, MC, R	B, C, N	P-R	✓	✓	[24, 56, 202]	✓
[117]	InP	BC, MC, R	B, C	CF	✓	✓		
[171]	InP	BC, MC, R	B, C	ϵ, δ-ACF		✓		
[189]	InP, PostP	BC	B	α-D	✓			
[199]	InP, PostP	BC, MC	B, C	PrefI, PrefT		✓		
[197]	InP	BC	B, C	DM		✓	[76]	
[192]	InP	BC	B	rND, rKL, rRD		✓		✓
[207]	PreP	BC	B	CDP	✓	✓	[56, 210]	✓
[13]	InP	BC	B	AVD, SD		✓	[76, 197]	
[28]	PreP	C	B	Bal	✓	✓		
[160]	PreP, InP	BC, MC, R	B, C, N	HSIC		✓		

Oneto, L. (2020). Learning fair models and representations. Intelligenza Artificiale, 14(1), 125-152.DOI 10.3233/IA-190034

Pre-processing: Fair Representation Learning

- Approaches
  - Awareness
  - Representation Learning
  - Re-weighting
  - Resampling → Over/Under – SMOTE, etc



- Z → Latent representation
  - $\max_{Z=g(X)} I(X;Z)$
  - subject to $I(A;Z) < e$
  - S⊥A

$$\alpha Loss_{similarity} + \beta Loss_{fairness} + \gamma Loss_{prediction}$$

$D = \{(a_i, x_i, y_i)\}_{i=1}^{N}$
$x_i \in R^d$
$g: R^d \to R^r$ i.e., $g(x_i) = z_i$
$z_i \in R^z$
$z_i \perp a_i$
$Z \perp A$

If model involved [hybrid]:
$f(g(X))$

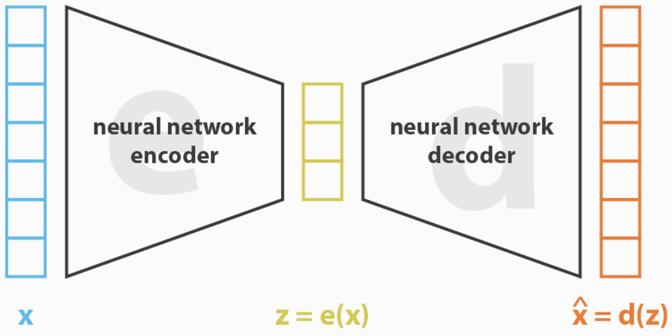
- Strict approach → Optimizes only Statistical Parity or Individual Fairness
  - Info of Y not used
- No need to access A at test time nor Y at representation time
- If Y is used → **hybrid** approach with potential better results [S⊥A|Y and Y⊥A|S]

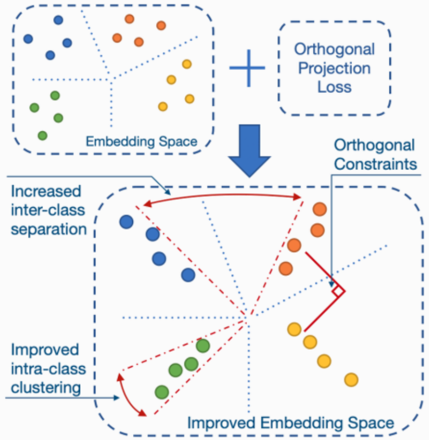
Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. 2013,. **Learning fair representations**. In International conference on machine learning
Cynthia Dwork,et al. 2012**. Fairness Through Awareness**. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference
F. Kamiran and T.G.K. Calders. 2012. **Data preprocessing techniques for classification without discrimination.** Knowledge and Information Systems 33

Pre-processing: Fair Representation Learning

Lots of works using NN
max I(A, g(X)) while *min I(A,g(X))* and may *max(g(X),Y)*

$$Loss_C = |x - x'|^2 - \lambda\, Loss_A(z)$$





$$Loss_C = \alpha|x - x'|^2 + \lambda\, Loss_A(Z_A) + \beta L\perp$$

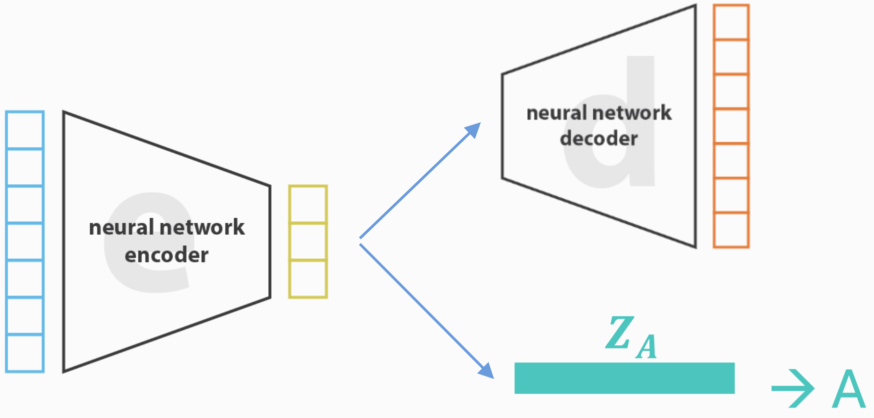
$$\alpha Loss_{similarity} + \beta Loss_{fairness} + \gamma Loss_{prediction}$$

`aif360.algorithms.preprocessing`.LFR

class `aif360.algorithms.preprocessing.LFR`(*unprivileged_groups, privileged_groups, k=5, Ax=0.01, Ay=1.0, Az=50.0, print_interval=250, verbose=0, seed=None*) [source]

Learning fair representations is a pre-processing technique that finds a latent representation which encodes the data well but obfuscates information about protected attributes [2]. .. rubric:: References

[2] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning Fair Representations.” International Conference on Machine Learning, 2013.

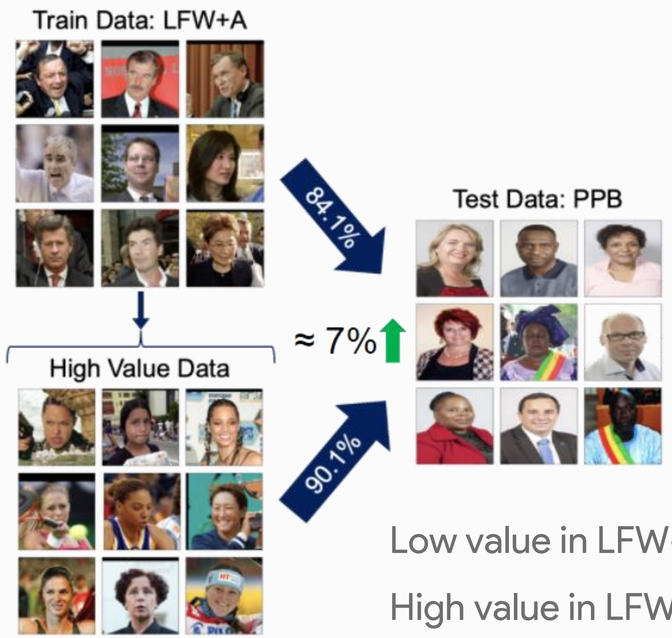


Bai, H.,et al.(2020). Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. preprint arXiv:2012.09382
FRLTradeoffs: https://blog.ml.cmu.edu/2020/02/28/inherent-tradeoffs-in-learning-fair-representations/

Pre-processing: Reweighting

- Weight the examples (group, label) to ensure fairness in classification
- Unbalanced learning-related → e.g., Fair-SMOTE
- Advanced example → SHAPLEY values

Domain adaptation: gender detection



Low value in LFW+A - males – overrepresented

High value in LFW+A –women – underrepresented

`aif360.algorithms.preprocessing`**.Reweighing**

class `aif360.algorithms.preprocessing.Reweighing`(*unprivileged_groups, privileged_groups*) [source]

Reweighing is a preprocessing technique that Weights the examples in each (group, label) combination differently to ensure fairness before classification [4].

References

[4] F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, 2012.

Ghorbani, A., & Zou, J. (2019, May). Data shapley: Equitable valuation of data for machine learning. In ICML. PMLR
Joymallya Chakraborty, et al. 2021. Bias in Machine Learning Software: Why? How? What to Do?. 29th ESEC/FSE 2021. ACM

In-processing

aif360.algorithms.inprocessing.PrejudiceRemover

class aif360.algorithms.inprocessing.PrejudiceRemover(*eta=1.0, sensitive_attr='', class_attr=''*) [source]

Prejudice remover is an in-processing technique that adds a discrimination-aware regularization term to the learning objective [6].

References

[6] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer," Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2012.

- Add penalty to objective function during learning → Regularizer
- Prior work: **Prejudice remover** (Kamishima et al., 2012)
  - Prejudice remover regularizer: Based on the degree of indirect prejudice (PI)

Mutual Information between Y and S

$$\boldsymbol{PI} = \sum_{(y,a)\in D} \hat{P}[y,s] \ln \frac{\hat{P}[y,s]}{\hat{P}[y]\hat{P}[s]}$$

S: *protected/sensitive attribute*

Prejudice remover regularizer

$$\mathrm{R}_{\mathrm{PR}}(\mathcal{D}, \boldsymbol{\Theta}) = \sum_{(\mathbf{x}_i, s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y|\mathbf{x}_i, s_i; \boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}$$

$$\sum_{(y_i, \mathbf{x}_i, s_i)} \boxed{\ln \mathcal{M}[y_i|\mathbf{x}_i, s_i; \boldsymbol{\Theta}]} + \boxed{\eta\, \mathrm{R}_{\mathrm{PR}}(\mathcal{D}, \boldsymbol{\Theta})} + \boxed{\frac{\lambda}{2} \sum_{s\in\mathcal{S}} \|\mathbf{w}_s\|_2^2},$$

Logistic Regression

Prejudice remover regularizer

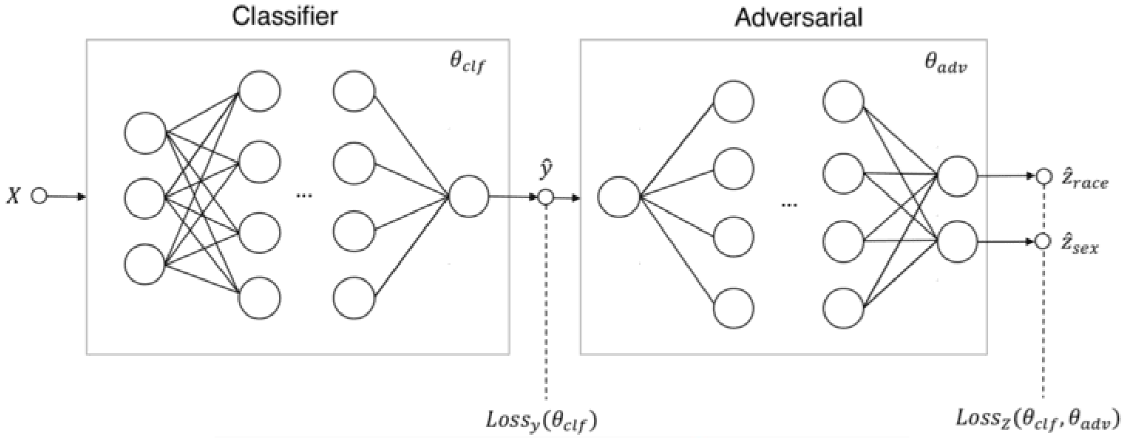
L2 Regularization

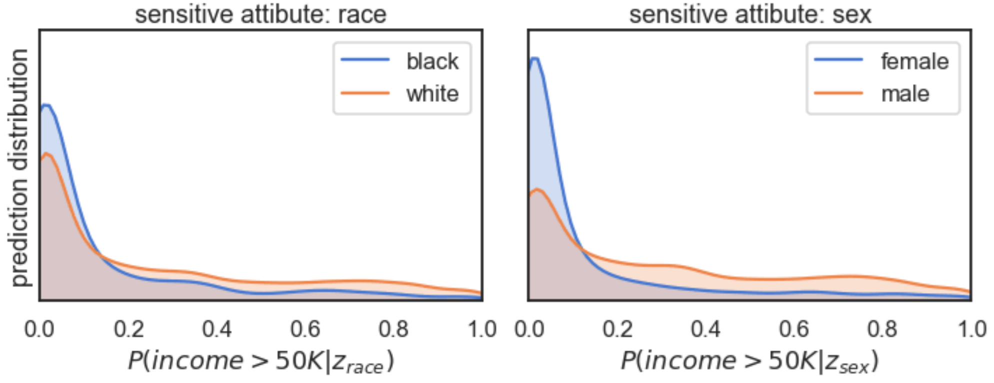
Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. Joint ECML-KDD.

In-processing: Adversarial debiasing

- Make the best possible predictions while ensuring that A cannot be derived from them
  - Demographic Parity
    - Adversary gets $\hat{Y}$
  - Equality Of Odds
    - Adversary gets $\hat{Y}$ and Y
  - Equality Of Opportunity
    - On a given class y → restrict adversary's training set to X where $Y = y$

$$\min_{\theta_{clf}}[Loss_y(\theta_{clf}) - \lambda Loss_Z(\theta_{clf}, \theta_{adv})]$$





Training iteration #1

Prediction performance:
- ROC AUC: 0.90
- Accuracy: 84.9

Satisfied p%-rules:
- race: 44%-rule
- sex: 35%-rule

aif360.algorithms.inprocessing.**AdversarialDebiasing**

class `aif360.algorithms.inprocessing.AdversarialDebiasing`*(unprivileged_groups, privileged_groups, scope_name, sess, seed=None, adversary_loss_weight=0.1, num_epochs=50, batch_size=128, classifier_num_hidden_units=200, debias=True)*
[source]

$$p\%rule = \min\left(\frac{P\{\hat{Y} = 1 \mid A = a\}}{P\{\hat{Y} = 1 \mid A = b\}}, \frac{P\{\hat{Y} = 1 \mid A = b\}}{P\{\hat{Y} = 1 \mid A = a\}}\right) \geq \frac{p}{100}$$

Zhang, B. H., et al (2018). Mitigating unwanted biases with adversarial learning. 2018 AAAI/ACM AI, Ethics, and Society (pp. 335-340). https://arxiv.org/pdf/1801.07593.pdf
Towards fairness in ML with adversarial networks. Stijn Tonk. 27 April 2018. URL: https://godatadriven.com/blog/towards-fairness-in-ml-with-adversarial-networks/

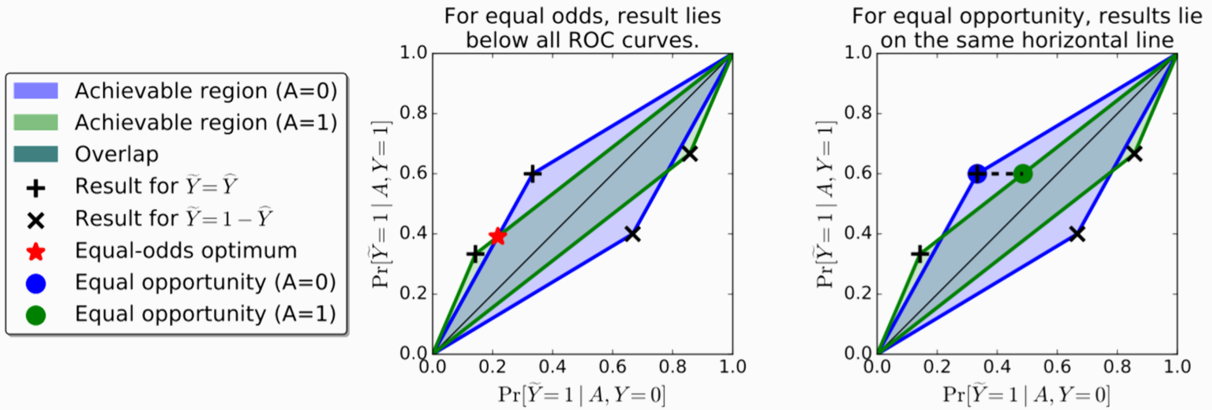
Post-processing

- Deal with output predictions of the model
  - Useful in black-box models or if we don't have access to the train pipeline → NO retraining
  - **Find a proper threshold** using the output for each group
  - Require A to be available in testing → compliance risk

`aif360.algorithms.postprocessing`.**EqOddsPostprocessing**

class `aif360.algorithms.postprocessing.EqOddsPostprocessing`(*unprivileged_groups, privileged_groups, seed=None*) [source]

Equalized odds postprocessing is a post-processing technique that solves a linear program to find probabilities with which to change output labels to optimize equalized odds [8] [9].



`aif360.algorithms.postprocessing`.**RejectOptionClassification**

class `aif360.algorithms.postprocessing.RejectOptionClassification`(*unprivileged_groups, privileged_groups, low_class_thresh=0.01, high_class_thresh=0.99, num_class_thresh=100, num_ROC_margin=50, metric_name='Statistical parity difference', metric_ub=0.05, metric_lb=-0.05*) [source]

Reject option classification is a postprocessing technique that gives favorable outcomes to unpriviliged groups and unfavorable outcomes to priviliged groups in a confidence band around the decision boundary with the highest uncertainty [10].

Nengfeng Zhou, et al.. 2021. Bias, Fairness, and Accountability with AI and ML Algorithms. arXiv:2105.06558
F. Kamiran, A. Karim, and X. Zhang, 2012 "Decision Theory for Discrimination-Aware Classification," IEEE International Conference on Data Mining
G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, 2017 "On Fairness and Calibration," Conference on Neural Information Processing Systems
M. Hardt, E. Price, and N. Srebro, 2016 "Equality of Opportunity in Supervised Learning," Conference on Neural Information Processing Systems

More prominent approaches

Causality

Domain-specific
Images
Text
Graphs

Discriminatory Transfer
Multitask Fairness

XAI
Interpretability

Game theoretical
approaches

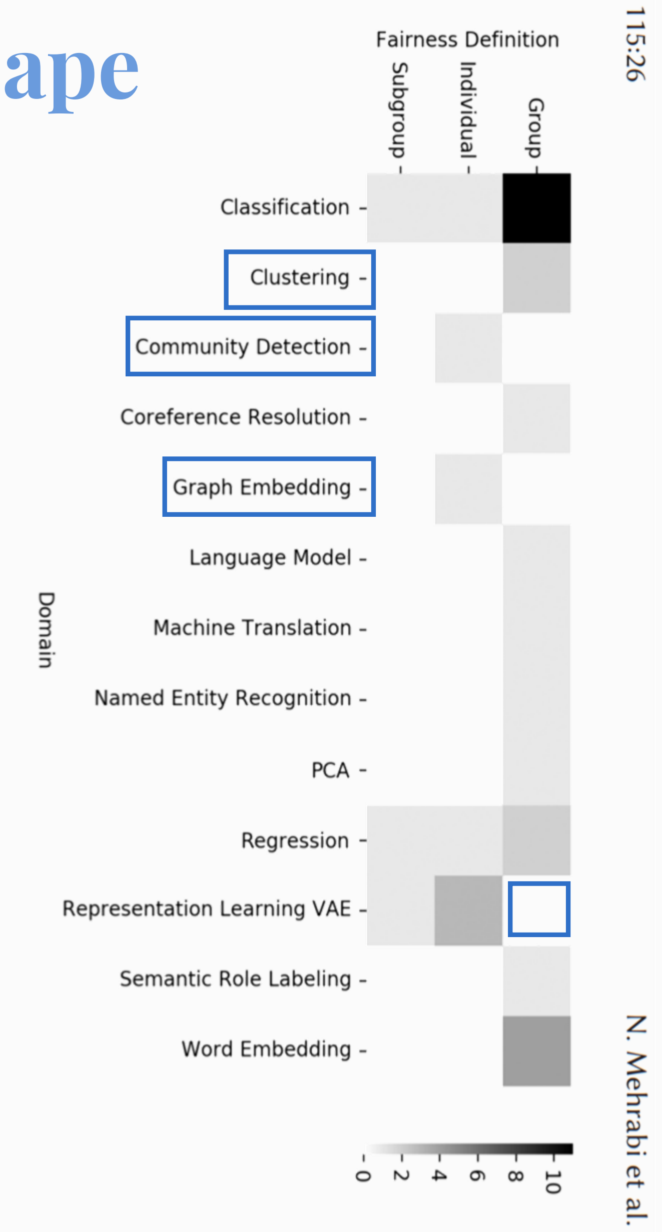
Current situation

Quick view on graphs & causality

Current landscape

Table 2. List of Papers Targeting and Talking about Bias and Fairness in Different Areas

Area	Reference(s)
Classification	[25, 49, 57, 63, 69, 73, 75, 78, 85, 102, 118, 143, 150, 151, 155]
Regression	[1, 14]
PCA	[133]
Community detection	[101]
Clustering	[8, 31]
Graph embedding	[22]
Causal inference	[82, 95, 111, 112, 123, 156, 160, 161]
Variational auto encoders	[5, 42, 96, 108]
Adversarial learning	[90, 152]
Word embedding	[20, 58, 165] [23, 162]
Coreference resolution	[130, 164]
Language model	[21]
Sentence embedding	[99]
Machine translation	[52]
Semantic role labeling	[163]
Named Entity Recognition	[100]



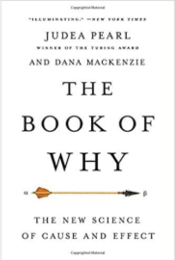
Mehrabi, N., et al. (2021). **A survey on bias and fairness in machine learning**. ACM Computing Surveys (CSUR), 54(6), 1-35

Graphs & Fairness

What fairness need? *Defining – detecting – imposing - apply*	**How can Graphs help?**
Capture Individual similarity	– Natural node pairwise distance – Structural similarity – Role similarity – Graph Representation Learning *(for Nodes & Edges & Graphs)*
Capture Group Structure-Behavior	– Community detection – Inherent data structure in graphs – Structural Analysis (e.g., Laplacian)
Capture deeper relationships between data	– Node – Edge - classification – Missing link prediction – Message passing – Information Flow – Rewiring – Changing graph structure
Different label bias problems	– Semi-Supervised Learning *i.e., help with labels we cannot see*
Causality	– Strong theory behind graphs – GNN → SCM
Applied to social problems	– Network is the natural structure of data – Also, everything can be modeled as a graph
XAI	– Interpretable by design – Friendly straightforward graph explanations – Great XAI graph-based

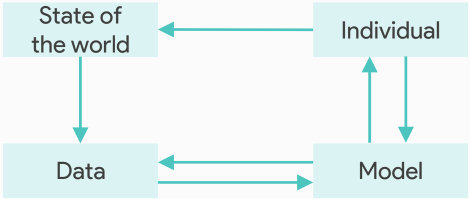
Yuan, H., Yu, H., Gui, S., & Ji, S. (2020). Explainability in graph neural networks: A taxonomic survey. arXiv preprint arXiv:2012.15445
Zecevic, M., Dhami, D. S., Velickovic, P., & Kersting, K. (2021). Relating graph neural networks to structural causal models. arXiv preprint arXiv:2109.04173
R. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec. 2019 GNNExplainer: Generating Explanations for Graph Neural Networks, NeurIPS
Bose, A., & Hamilton, W. (2019). Compositional fairness constraints for graph embeddings. ICML. PMLR.

Causality



- Previous definitions relies on **Joint probabilities of (X,Y,S,A)**
  - Reactive vision: take everything as given about the world as it is → Observational
- Can we capture social context? **Let's use causal models**
  - How changes in variables propagate in a system, be it natural, engineered or social
  - What should we do when there's no direct effect?

Exploit Structural Causal Model properties to look for biases Neal, B. (2020)



Definition 4.2 (Structural Causal Model (SCM)) *A structural causal model is a tuple of the following sets:*

1. *A set of endogenous variables V*
2. *A set of exogenous variables U*
3. *A set of functions f, one to generate each endogenous variable as a function of other variables*

$$M: \quad \begin{aligned} B &:= f_B(A, U_B) \\ C &:= f_C(A, B, U_C) \\ D &:= f_D(A, C, U_D) \end{aligned}$$

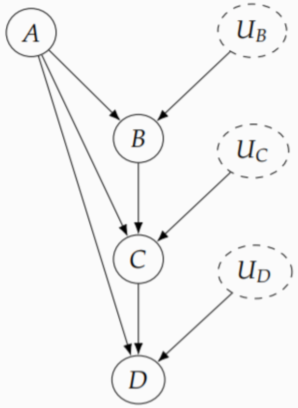
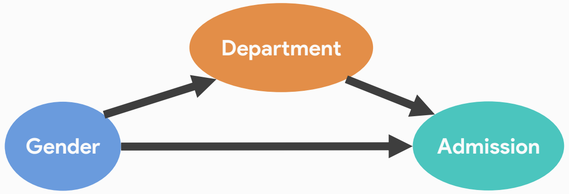


Figure 4.8: Graph for the structural equations in Equation 4.24.



Causal fairness criteria and path-specific effects

J. Pearl, 2009 Causality: Models, Reasoning and Inference, 2nd ed. New York, NY, USA: Cambridge University Press,
Neal, B. (2020). Introduction to causal inference from a ML perspective. *Book (draft)*. https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf
Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness.
Loftus, J. R., Russell, C., Kusner, M. J., & Silva, R. (2018). Causal reasoning for algorithmic fairness
Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). Survey on Causal-based Machine Learning Fairness Notions. arXiv preprint arXiv:2010.09553.
Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning
Zhang, J., & Bareinboim, E. (2018, April). Fairness in decision-making—the causal explanation formula. In Thirty-Second AAAI
Wu, Y. (2020). Achieving Causal Fairness in Machine Learning
S. Chiappa. 2019, Path-specific counterfactual fairness. Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)
Chiappa, S., & Isaac, W. S. (2018,). A causal bayesian networks viewpoint on fairness. In IFIP International Summer School on Privacy and Identity Management
Fairness – Moritz Hardt – Part 2 – MLS2020 - https://www.youtube.com/watch?v=9oNVFQ9lIPc&t=1449s

Causality: examples

- **Counterfactual fairness:**
  - Outcome probability in factual world = the counterfactual world
  - How would the world have to be different for a desirable output to occur?
  - *What would have happened if I were different?*
- **Causal Representation Learning**
- **Algorithmic Recourse**
  - → Causality +XAI → explanations + recommendations
  - **Actionable** feedback about how to change the outcomes of ML models
  - *"To have your loan approved, you would need to increase your income by $10,000 per year"*

*"**Counterfactuals explain** complex models with the **use of examples**...*
*...while **recourse** tries to **find actions** that leads to a better outcome"* Annabelle Redelmeier

	Counterfactuals	**Recourse**
Optimization function	Loss function	Cost function
Algorithm solves for…	Vectors/Individuals (x)	Actions (δ)
Ultimate goal	Explain a model	Solve for actions to achieve "recourse"

Karimi, A. H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv:2010.04050
Karimi, A. H., Schölkopf, B., & Valera, I. (2021,). Algorithmic recourse: from counterfactual explanations to interventions. In Proceedings of the 2021 ACM Conference FAccT
A (deeper) look at counterfactuals in explainable AI April 29th, 2021 Annabelle Redelmeier Norwegian Computing Center (Norsk Regnesentral)

Libraries

Libraries

IBM Research Trusted AI

AI Fairness 360

FAIRENSICS

Fairlearn

FairKit

Aequitas

Bias & Fairness Audit

Datasets

Benchmarking datasets

- Big amount of tabular dataset in all domains

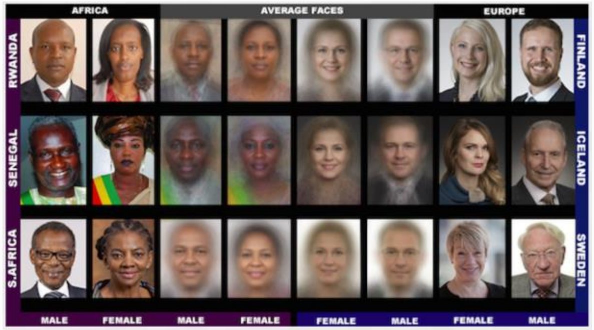


- Every dataset may have intrinsic bias

Images

Text

School Effectiveness	[66]	15362	9	Ethnicity, Gender	R
Heart Disease	[90]	303	75	Age, Gender	MC, R
German Credit	[85]	1K	20	Age, Gender/Marital-Stat	MC
Census/Adult Income	[112]	48842	14	Age, Ethnicity, Gender, Native-Country	BC
Contraceptive Method Choice	[121]	1473	9	Age, Religion	MC
Law School Admission	[187]	21792	5	Ethnicity, Gender	R
Arrhythmia	[70]	452	279	Age, Gender	MC
Communities & crime	[169]	1994	128	Ethnicity	R
Wine Quality	[154]	4898	13	Color	MC, R
Heritage Health	[146]	≈60K	≈20	Age, Gender	MC, R
Stop, Question & Frisk	[45]	84868	≈100	Age, Ethnicity, Gender	BC, MC
Bank Marketing	[142]	45211	17-20	Age	BC
Diabetes US	[181]	101768	55	Age, Ethnicity	BC, MC
Student Performance	[38]	649	33	Age, Gender	R
CelebA Faces	[122]	≈200K	40	Gender Skin-Paleness, Youth	BC
xAPI Students Perf.	[6]	480	16	Gender, Nationality, Native-Country	MC
Chicago Faces	[127]	597	5	Ethnicity, Gender	MC
Credit Card Default	[195]	30K	24	Age, Gender	BC
COMPAS	[119]	11758	36	Age, Ethnicity, Gender	BC, MC
MovieLens	[77]	100K	≈20	Age, Gender	R
Drug Consumption	[54]	1885	32	Age, Ethnicity, Gender, Country	MC
Student Academics Perf.	[87]	300	22	Caste, Gender	MC
NLSY	[148]	≈10K		Birth-date, Ethnicity, Gender	BC, MC, R
Diversity in Faces	[140]	1 M	47	Age, Gender	MC, R



Pilot Parliaments Benchmark

Retiring Adult:
New Datasets for Fair Machine Learning

Frances Ding* UC Berkeley
Moritz Hardt* UC Berkeley
John Miller* UC Berkeley
Ludwig Schmidt* Toyota Research Institute

Quy, T. L., Roy, A., Iosifidis, V., & Ntoutsi, E. (2021). A survey on datasets for fairness-aware machine learning. arXiv
Oneto, L. (2020). Learning fair models and representations. Intelligenza Artificiale, 14(1), 125-152
Barocas, S., Hardt, M., & Narayanan, A. (2017). **Fairness in machine learning**. Nips tutorial, 1, 2017
Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., & Menzies, T. (2021). Fair Enough: Searching for Sufficient Measures of Fairness. preprint arXiv:2110.13029.
http://gendershades.org/overview.html - https://nips.cc/media/neurips-2021/Slides/26854.pdf

History and conceptual point of view

What should we learn from the past fairness research?
What other conceptual concerns should we consider?

Fairness beginning: 60's & 70's

Shout out to pioneers

1966	1968	1971	1971	1973	1976
Guion	Cleary	Thorndike	Darlington	Cole	Peterson and Novick

- 60's: start to quantify bias
- 70's: From unfairness to Fairness
  - FP & FN rates
  - Fair use of the test, rather than the scores themselves
- Mid 70's: halt ☹, Why?
  - **No** analyses to **unequivocally indicate fairness**
  - **No clear** procedures to **avoid unfairness**
  - **Disagreement in views of fairness** view between professionals and general public

  "***Fairness actually obscure*** *the fundamental problem, which is to find some rational basis for* ***providing compensatory treatment for the disadvantaged***" (Melvin R Novick et al. 1976)
- Rediscovered by ML around 13 year ago (Calders et al. 2009)

What should we learn?

- DON'T reinvent the wheel
- DON'T forget actual objective
  → compensatory treatment to disadvantaged

- DON'T get stacked in discussions far from real-world problems
- DON'T be far from **practical needs** of society, politics & law
- Work in political and law implication
- Relating fairness debates to ethical theories and value systems

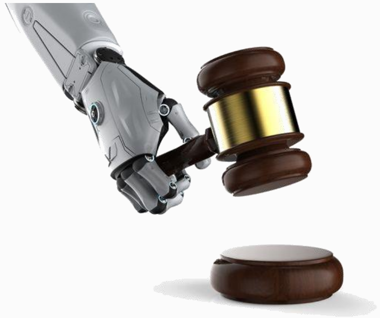
- ML Fairness community should be more aware of our own implicit cultural biases

Hutchinson, B., & Mitchell, M. 2019. **50 years of test (un) fairness: Lessons for machine learning**. FAccT 2019
Nancy S Cole and Michael J Zieky. 2001. The new faces of fairness. Journal of Educational Measurement 38, 4
Rebecca Zwick and Neil J Dorans. 2016. Philosophical Perspectives on Fairness in Educational Assessment. In Fairness in Educational Assessment and Measurement
T. Anne Cleary. 1966. Test bias: Validity of the Scholastic Aptitude Test for Negro and white students in integrated colleges
Calders, Kamiran, and Pechenizkiy, "Building Classifiers with Independency Constraints," in In Proc. IEEE ICDMW, 2009, 13–18
Kamiran and Calders, "Classifying Without Discriminating," in Proc. 22Nd International Conference on Computer, Control and Communication, 2009.

Fair ML and law

*"Careful attention should be paid to **legal and public concerns about fairness**. The experiences of the test fairness field suggest that in the coming years, **courts may start ruling on the fairness of ML models**. Therefore, **If technical definitions of fairness stay too far** from the public's perceptions of fairness, then the **political will to use scientific contributions** in advance of public policy **may be difficult to obtain**"*

Hutchinson, B., & Mitchell, M. 2019.
50 years of test (un) fairness: Lessons for machine learning. FAccT 2019



Other cultural and conceptual challenges

Even we are looking for bias, **we are inducing bias**	**CONTEXT MATTERS** Quantitative techniques + policy-level questions	Make methods flexible to **adapt to each situation, context and use**
PUBLIC'S NOTION OF FAIRNESS Explicitly connect fairness criteria to different **socio-cultural and philosophical values**	Try to **unify fairness** definition and framework	Politics and law **implication**
Remind: Fairness and unfairness are related but different concepts	Make Fair ML research **accessible** to general public, other researchers	**From equality to equity** Give each one the resources that each one need to reach to the same point

Example of conceptual bias: **Why groups should be treated as discrete categories?**

- Most definitions of protected attribute-group relies on **categoric division → implicit cultural bias & unstable social construct**
- Other possibility: intersectional modelling → **Protected attribute as continuous variables**
  - Quantify fairness along one dimension (e.g., age) conditioned on another dimension (e.g., skin tone)

e.g., Use Computer vision clustering of skin tones instead of pre-defined ethnics

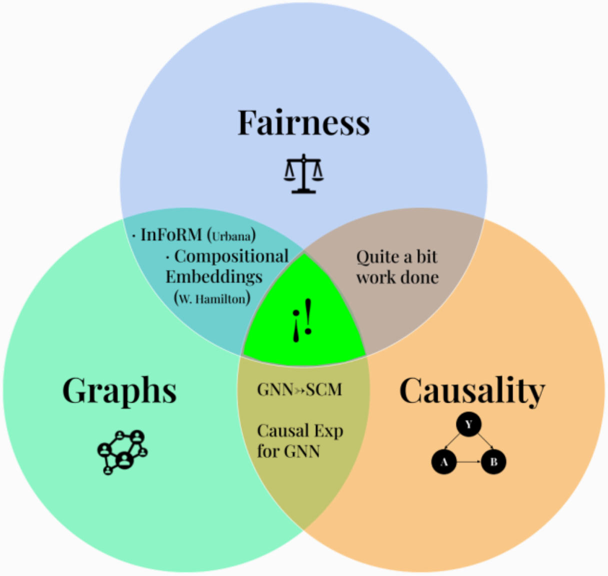
Wrapping up

Conclusion

- **Don't feel overwhelmed** by the big amount methods and measures!
  - Method depends on task, and technical context
  - Definitions and metrics depends on the context
  - Development and relationship of the measures with ethics → Now you choose context – experts – social and ethical analysis

- More work needed in **ethical-cultural aspect**
  - Equity → Considering individual resources
  - Continual protected attributes
  - Social-Law-Political needs close relationship

- **Technical takeaways**
  - Beyond observational → Causality
  - Deep structural data relationship → Graphs



Bibliography

More references in each slide

- M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in Neural Information Processing Systems (2016).
- Cynthia Dwork,et al. 2012. Fairness Through Awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference
- Alexandra Chouldechova. 2016. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data.
- Verma, J. Rubin, Fairness definitions explained, IEEE/ACM International Workshop on Software Fairness (2018) 1–7.
- Richard Berka, Hoda Heidaric, Shahin Jabbaric, Michael Kearnsc, and Aaron Rothc. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art.
- Alexandra Chouldechova. 2016. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data (2016)
- Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In ITCS
- Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness: Lessons for Machine Learning.
- M.J. Kusner, J. Loftus, C. Russell and R. Silva, Counterfactual fairness, In Neural Information Processing Systems, (2017)
- Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. Nips tutorial, 1, 2017
- Shira Mitchell. 2018. Reflection on quantitative fairness. Web Book
- Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., & Menzies, T. (2021). Fair Enough: Searching for Sufficient Measures of Fairness. arXiv preprint arXiv:2110.13029.
- L. Oneto, Learning fair models and representations, Intelligenza Artificiale 14 (1) (2020) 151–178.
- Castelnovo, A., Crupi, R., Greco, G., & Regoli, D. (2021). The zoo of Fairness metrics in Machine Learning. arXiv
- Franco, D., Navarin, N., Donini, M., Anguita, D., & Oneto, L. (2022). Deep fair models for complex data: Graphs labeling and explainable face recognition. Neurocomputing, 470
- A.F. Winfield, K. Michael, J. Pitt, V. Evers, Machine ethics: the design and governance of ethical ai and autonomous systems, Proceedings of the IEEE 107 (2019) 509–517
- D. Gunning, Explainable artificial intelligence (xai), Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2)
- Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., & Menzies, T. (2021). Fair Enough: Searching for Sufficient Measures of Fairness. preprint arXiv:2110.13029.
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671
- Lim Swee Kiat. Retrieved December 2021. Machines go Wrong. https://machinesgonewrong.com/fairness/
- Manuel Gomez Rodriguez et al. (2020). Human-Centric Machine Learning Feedback loops, Human-AI Collaboration and Strategic Behavior [Link]. Web
- Corbett-Davies & Goel. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning
- Mehrabi, N., et al. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35
- 2017. CS 294: Fairness in Machine Learning. https://fairmlclass.github.io (2017). Online; accessed February 2018

Talk in the scope of the project:

Achieving Fair, Accountable and Transparent Machine Learning Models through Graph Theory and Causality

Thesis in Progress by PhD Student Adrián Arnaiz Rodríguez

PhD Nuria Oliver

PhD Francisco Escolano

PhD Manuel Gómez Rodríguez

ellis ALICANTE unit

Thank you!

Q's & feedback?

adrian@ellisalicante.org

@arnaiztech

AdrianArnaiz

European Union
European Regional
Development Fund

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

GENERALITAT VALENCIANA
Conselleria d'Innovació, Universitats, Ciència i Societat Digital

Sabadell Fundación