Fairness in Machine Learning

Peng Cui
Tsinghua University

Outline

- Part I: Background
- Part II: Definitions of fairness in machine learning
- Part III: Algorithms of fairness in machine learning
- Part IV: Fairness In Economics
- Part V: Fairness In Language Models

Algorithmic Fairness is *Unignorable*

- Algorithmic fairness has triggered heated debate in machine learning tasks
  - **Main target**: high-stake decision-making systems

Loan applications

Hiring processes

Criminal justice

Machine Learning Fairness

- **What is fairness?**
  - Discrimination towards subgroup or individual
- **Why ML models become unfair?**
  - There exists bias in the data
  - Machine learning models learn the bias in the data

Example: COMPAS

Black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism. (45 percent vs. 23 percent)

Example: UC Berkeley Gender Bias

- Total acceptance rate: men > women
- Acceptance rate in most departments: women > men

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

Legend:

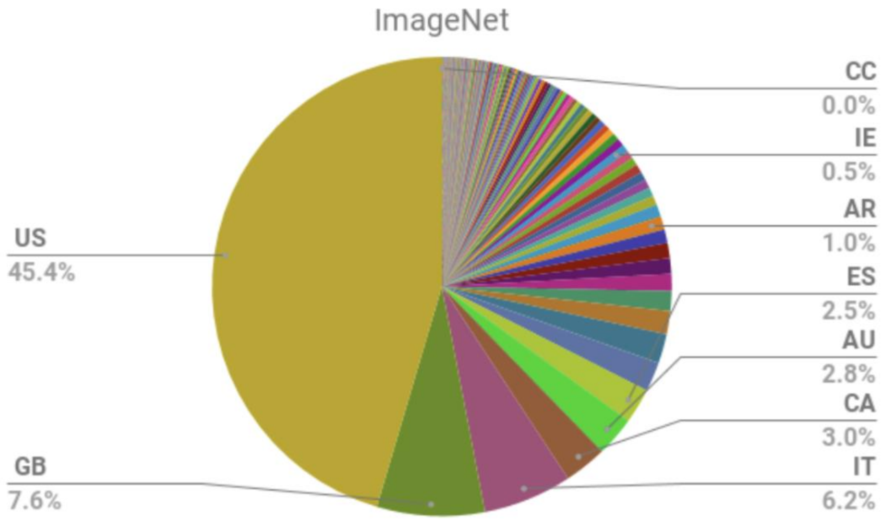
greater percentage of successful applicants than the other gender
greater number of applicants than the other gender

bold - the two 'most applied for' departments for each gender

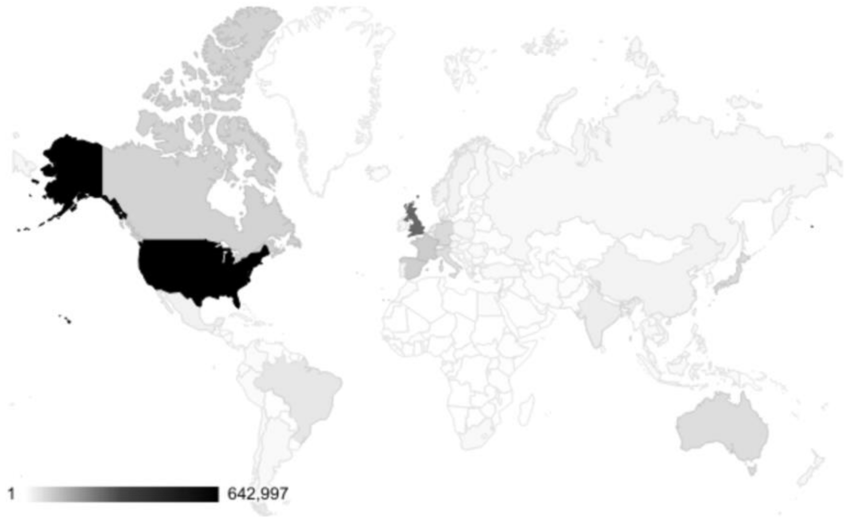
https://en.wikipedia.org/wiki/Simpson%27s_paradox

Example: Fraction of Each Country in Open Images and ImageNet Image Datasets

- US and Great Britain represent the top locations



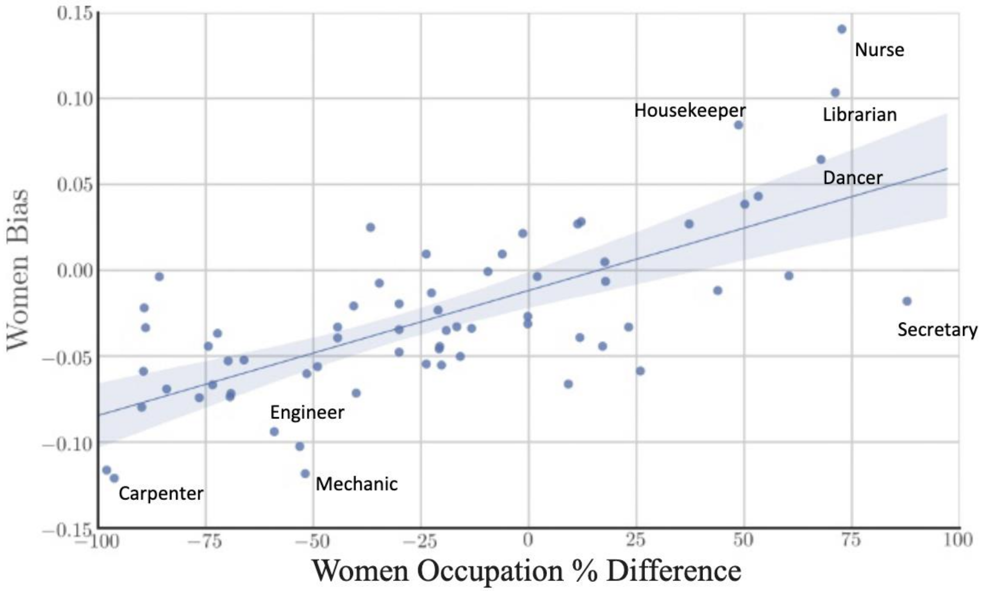
ImageNet



Open Images

Mehrabi, Ninareh, et al. A survey on bias and fairness in machine learning. CSUR, 2021.

Example: Biases in Word Embedding



Women’s occupation relative percentage vs. embedding bias in Google News vectors

Garg, Nikhil, et al. Word embeddings quantify 100 years of gender and ethnic stereotypes. PNAS, 2018.

Outline

- Part I: Background
- Part II: Definitions of fairness in machine learning
- Part III: Algorithms of fairness in machine learning
- Part IV: Fairness In Economics
- Part V: Fairness In Language Models

Basic Notations

- **Given individual features**
  - S: sensitive attributes, such as gender and race
    - Sometimes, we also use A to denote sensitive attributes
  - X: features
  - Y: outcomes
- **Example** ------ college admission case
  - S: gender
  - X: department choices, test scores, e.t.c.
  - Y: decision to admit a student
- **Target**: a fair $\hat{Y}$ that fits Y and satisfies some **fair constraints**.

Typical Fairness Notions

- **Individual fairness**
  - Fairness between individuals
- **Group fairness**
  - Fairness between subgroups
- **Causality-based fairness notions**
  - Using causal graph to characterize the unfair causal effect

Individual Fairness

- **General idea**
  - Similar individuals should be treated similarly
  $$D\big(f(x_1), f(x_2)\big) \leq d(x_1, x_2)$$
  - D: the distance in the outcome space
  - d: the distance in the feature space
- **Main issue**
  - The definition of function d is difficult.

Dwork, Cynthia, et al. Fairness through awareness. Proceedings of the 3rd innovations in theoretical computer science conference, 2012.

Group Fairness: Fairness through Unawareness

- **General idea**
  - Predict without sensitive attributes
  - $\hat{Y}$ is a function of X instead of (X, S)
- **Main issue**
  - Features X may be correlated with S
  - $\hat{Y}$ is still unfair
- **Example**
  - Zip code is strongly correlated with race
  - Prediction with zip code can still be unfair

Dwork, Cynthia, et al. Fairness through awareness. Proceedings of the 3rd innovations in theoretical computer science conference, 2012.

Group Fairness: Demographic Parity (DP)

- A predictor $\hat{Y}$ satisfies **demographic parity** if
  - The probabilities of positive predictions are the same regardless of whether the group is protected

$$P(\hat{Y} = 1 \middle| S = 0) = P(\hat{Y} = 1 \middle| S = 1)$$

- **Example**
  - The college admission rate should be the same for men and women
- **Issue**
  - Demographic parity often harms the utility that we might hope to achieve
  - The perfect predictor $\hat{Y}$ is not DP fair!

Dwork, Cynthia, et al. Fairness through awareness. Proceedings of the 3rd innovations in theoretical computer science conference, 2012.

Group Fairness: Equalized Odds (EO) and Equal Opportunity (EOpp)

- A predictor $\hat{Y}$ satisfies **equalized odds** if
  - Equal probabilities for both qualified/unqualified people across groups

$$P(\hat{Y}=1|S=0,Y=0)=P(\hat{Y}=1|S=1,Y=0)$$
$$P(\hat{Y}=1|S=0,Y=1)=P(\hat{Y}=1|S=1,Y=1)$$

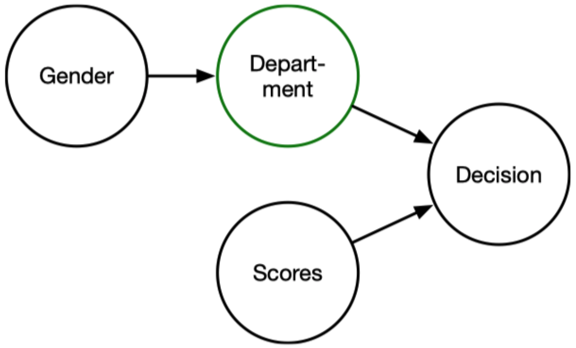
- A predictor $\hat{Y}$ satisfies **equal opportunity** if
  - Equal probabilities for qualified people across groups

$$P(\hat{Y}=1|S=0,Y=1)=P(\hat{Y}=1|S=1,Y=1)$$

Hardt, Moritz, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. NeurIPS, 2016.

Drawbacks of Group Fairness

- **Drawback**: Cannot distinguish detailed fair and unfair parts of the problem.



Causal graph of fair college admission case

$120 + 20 = 140 \neq 110 = 30 + 80$

	Department A		Deparment B	
	Male	Female	Male	Female
Number of applicants	400	100	100	400
Number of accepted	120	30	20	80
Acceptance rate(%)	30	30	20	20

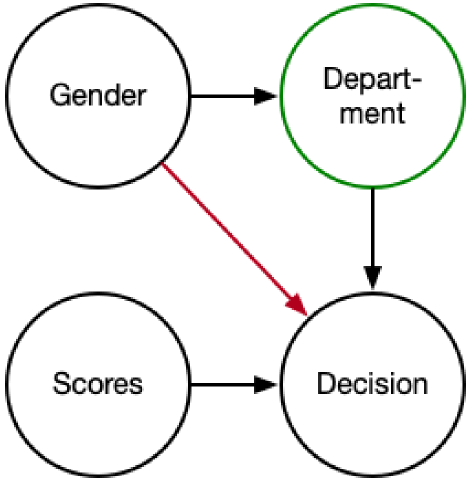
Same! Same!

Toy data of fair college admission case

- **Example** ------ fair college admission case
  - Total acceptance rates: male > female (Unfair in DP fairness)
  - Acceptance rates in different department: male == female!

Drawbacks of Group Fairness

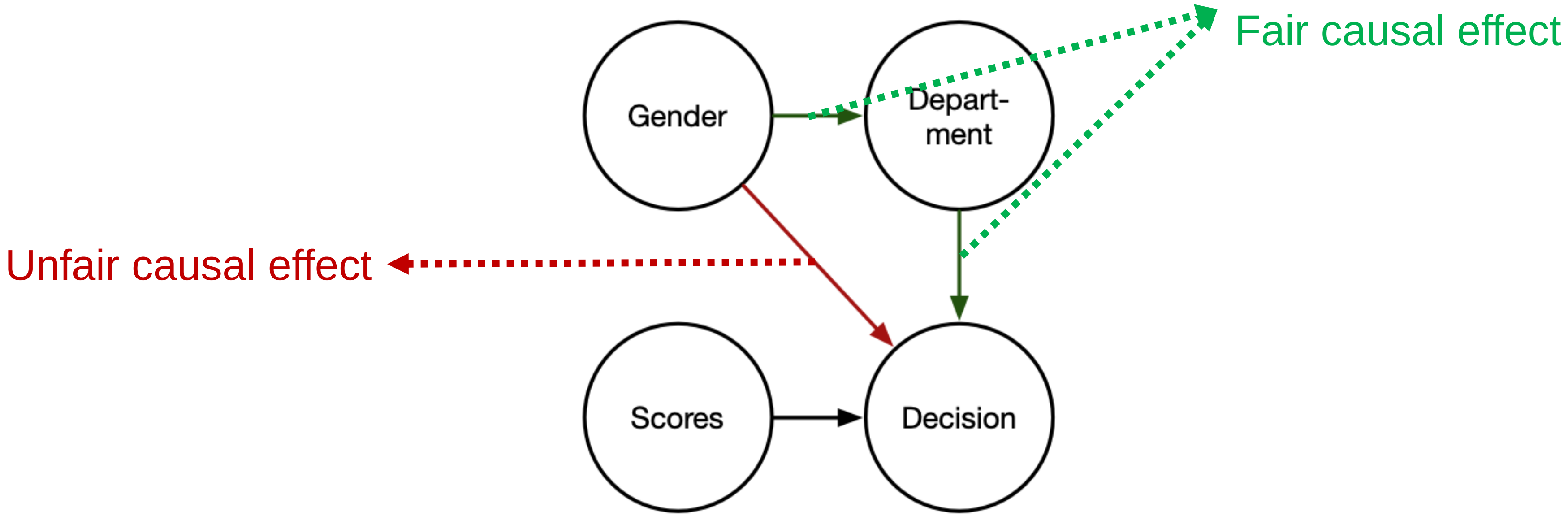
- **Drawback**: Cannot distinguish detailed fair and unfair parts of the problem.
- **Example** ------ unfair college admission case
  - Historical outcome Y is biased towards gender
  - Perfect predictor $\hat{Y} = Y$ satisfies EO constraint.
  - But it is not fair!



Causal graph of unfair college admission case

Causality-Based Fairness Notions

- **Causality-based fairness notions**
  - General idea: the unfair causal effect from S to Y should be zero.

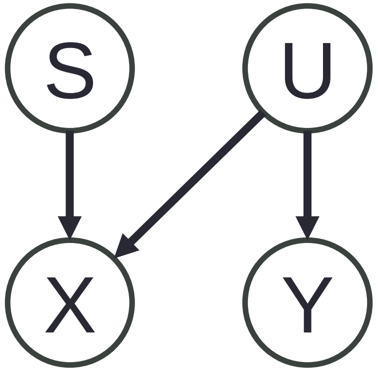


Causality-Based Fairness Notions: Counterfactual Fairness

- **Mathematical formulation**: for any x, y, a, a'

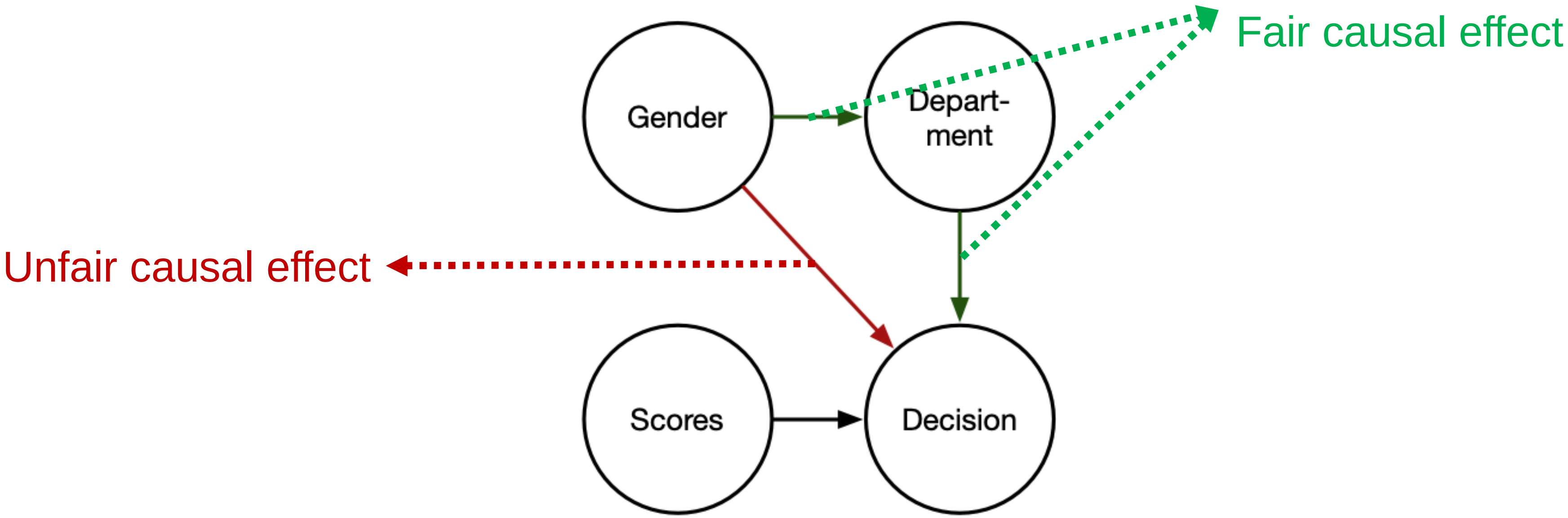
$$P\left(\hat{Y}_{S\leftarrow s}(U) = y \mid X = x, S = s\right) = P\left(\hat{Y}_{S\leftarrow s'}(U) = y \mid X = x, S = s\right)$$

  - It measures the **total causal effect** from S to Y
- **Implementation**
  - Predict via U and non-descendants of S
- **Example: the red car**
  - S: race, X: prefer red cars, U: aggressive driving, Y: accident rate
  - Counterfactually fair: predicting with U



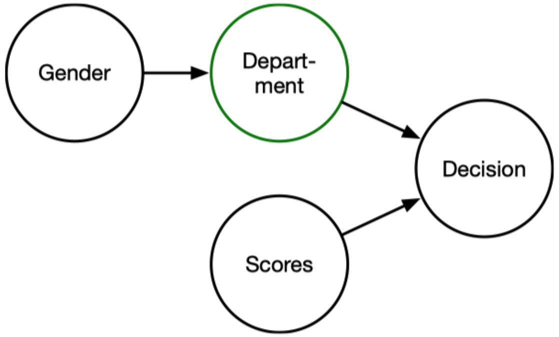
Drawbacks of Causality-Based Fairness Notions

- **Not scalable**
  - Need causal structure assumption
  - Need assumptions on fair and unfair paths

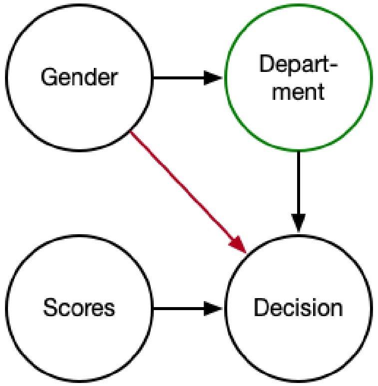


Group Fairness: Conditional Fairness

- **Fair variables**
  - Pre-decision covariates, which are irrelevant in assessing the fairness of decision-making algorithms
- **Example**
  - Department choice in the college admission case
- **Conditional fairness**
  - Outcome ⊥ sensitive attributes | fair variables
  - Explanation in College admission case:
    - In each departments, the acceptance rate should be equal.



Fair college admission case



Unfair college admission case

Xu, Renzhe, et al. Algorithmic decision making with conditional fairness. KDD, 2020.

Group Fairness: Conditional Fairness

- **Given individual features**
  - S: sensitive attributes, such as gender and race
  - X: features
    - F: fair variables
    - O: other variables
  - Y: outcomes
- **Target**: learn $\hat{Y}$ that fits Y and satisfies $\hat{Y} \perp S \mid F$.
- **Special cases**
  - $F = \emptyset \Longrightarrow$ demographic parity ($\hat{Y} \perp S$).
  - $F = Y \Longrightarrow$ equalized odds ($\hat{Y} \perp S \mid Y$).

Xu, Renzhe, et al. Algorithmic decision making with conditional fairness. KDD, 2020.

Group Fairness: Subgroup Fairness

- Consider the following setting with two sensitive attributes
  - Suppose the fractions of white men, white women, black men, and black women are ¼

$\hat{Y}$	Men	Women
White people	0	1
Black people	1	0

  - $\hat{Y}$ is **DP fair** if considering **only one sensitive attribute** (men vs women or while vs black)
  - $\hat{Y}$ is **unfair** if considering **both sensitive attributes**
- We need to take all sensitive attributes into account!

Kearns, Michael, et al. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. ICML, 2018.

Inherent Trade-Off between Different Fairness Notions

- Different fairness notions may **contradict** with each other
  - Equalized odds and calibration can be satisfied at the same time when
    1. Base rate equals: $P(Y = 1|S = 0) = P(Y = 1|S = 1)$, or
    2. The prediction is perfect: $\hat{Y} = Y$

- We need to choose proper fairness notions in different applications!

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. arXiv, 2016.

Outline

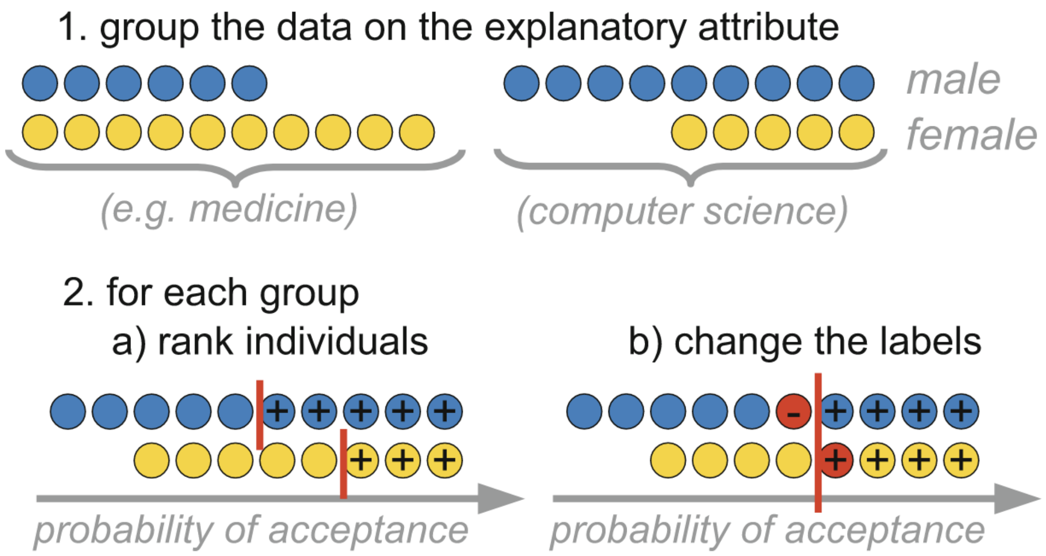
- Part I: Background
- Part II: Definitions of fairness in machine learning
- Part III: Algorithms of fairness in machine learning
- Part IV: Fairness In Economics
- Part V: Fairness In Language Models

Categories of Algorithms

- Most methods are on **group fairness**
  - Pre-processing methods
  - In-processing methods
  - Post-processing methods

Pre-Processing

- **Key idea**
  - Change the distribution of $P(X|S=0)$ and $P(X|S=1)$
- **Method**: local massaging



Kamiran, Faisal, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and information systems, 2013.

In-Processing: Optimization

- **General idea**
  - Formulate the problem as a general constrained optimization problem
  - Then solve it
- **Fairness target**
  - Consider any fairness targets that can be formulated as a form of linear constraints

$$\mathbf{M}\boldsymbol{\mu}(h) \leq \mathbf{c},$$

  - $\mu(h)$ is a vector of conditional moments of the form

$$\mu_j(h) = \mathbb{E}\big[\, g_j(X, A, Y, h(X)) \;\big|\; \mathcal{E}_j \,\big] \quad \text{for } j \in \mathcal{J},$$

  - This includes common group fairness notions, including DP, EO, and EOpp.

Agarwal, Alekh, et al. A reductions approach to fair classification. International conference on machine learning. PMLR, 2018.

In-Processing: Optimization

- **Overall target**

$$\min_{h \in \mathcal{H}} \mathrm{err}(h) \quad \text{subject to} \quad \mathbf{M}\boldsymbol{\mu}(h) \leq \mathbf{c}.$$

- **How to solve it?**
  - More generally, consider randomized classifiers $Q \in \Delta$, where Δ is a distribution on $\mathcal{H}$

$$\mathrm{err}(Q) = \textstyle\sum_{h \in \mathcal{H}} Q(h)\, \mathrm{err}(h)$$

  - The new optimization target:

$$\min_{Q \in \Delta} \mathrm{err}(Q) \quad \text{subject to} \quad \mathbf{M}\boldsymbol{\mu}(Q) \leq \mathbf{c},$$

  - This is actually a linear programming on Q! (although the space of Q is large)

Agarwal, Alekh, et al. A reductions approach to fair classification. International conference on machine learning. PMLR, 2018.

In-Processing: Optimization

- **Overall target**

$$\min_{h \in \mathcal{H}} \mathrm{err}(h) \quad \text{subject to} \quad \mathbf{M}\boldsymbol{\mu}(h) \leq \mathbf{c}.$$

- **Optimization**
  - Consider the Lagrangian function

$$L(Q, \boldsymbol{\lambda}) = \widehat{\mathrm{err}}(Q) + \boldsymbol{\lambda}^{\top}\left(\mathbf{M}\widehat{\boldsymbol{\mu}}(Q) - \widehat{\mathbf{c}}\right).$$

  - Strong duality holds for this problem

$$\min_{Q \in \Delta} \max_{\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{K}|}, \, \|\boldsymbol{\lambda}\|_1 \leq B} L(Q, \boldsymbol{\lambda}), \qquad \text{(P)}$$

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{K}|}, \, \|\boldsymbol{\lambda}\|_1 \leq B} \min_{Q \in \Delta} L(Q, \boldsymbol{\lambda}). \qquad \text{(D)}$$

Agarwal, Alekh, et al. A reductions approach to fair classification. International conference on machine learning. PMLR, 2018.

In-Processing: Optimization

- **Optimization**

$$\min_{Q\in\Delta} \max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{|\mathcal{K}|},\, \|\boldsymbol{\lambda}\|_1\le B} L(Q,\boldsymbol{\lambda}), \tag{P}$$

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{|\mathcal{K}|},\, \|\boldsymbol{\lambda}\|_1\le B} \min_{Q\in\Delta} L(Q,\boldsymbol{\lambda}). \tag{D}$$

- **Method**
  - Iteratively optimize Q and λ

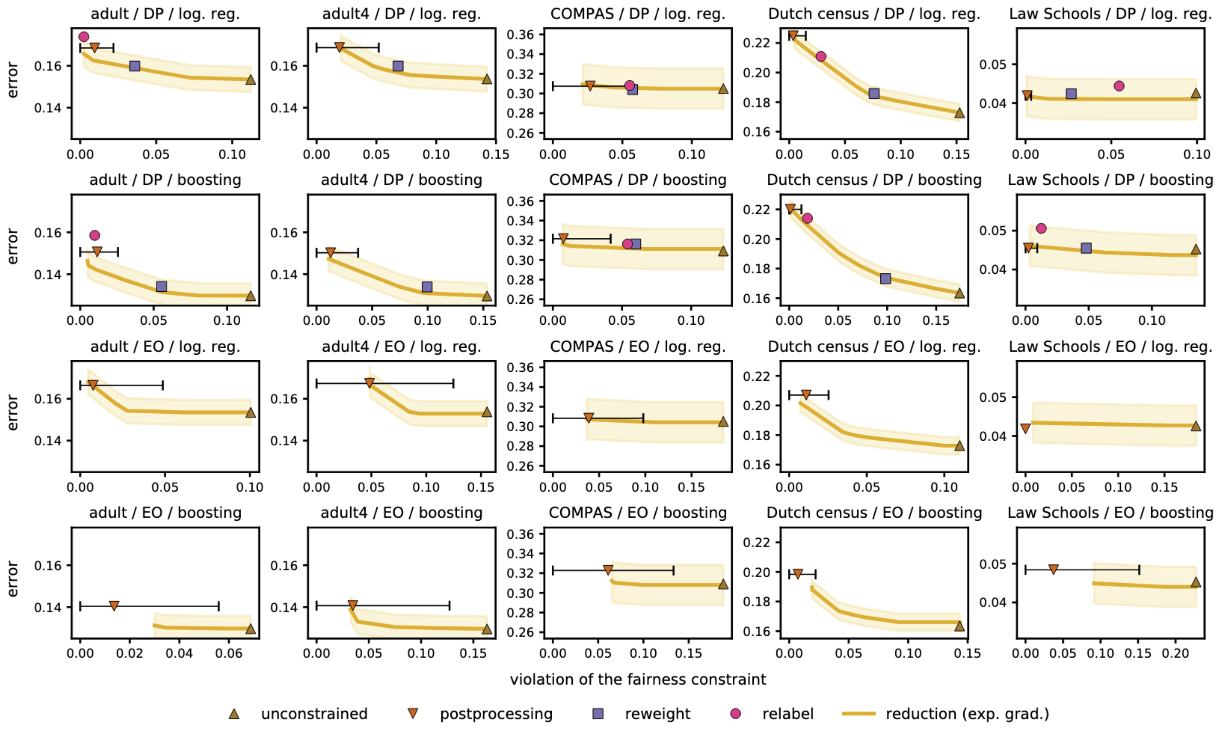
Algorithm 1 Exp. gradient reduction for fair classification

Input: training examples $\{(X_i, Y_i, A_i)\}_{i=1}^n$
fairness constraints specified by g_j, $\mathcal{E}_j$, $\mathbf{M}$, $\widehat{\mathbf{c}}$
bound B, accuracy ν, learning rate η
Set $\boldsymbol{\theta}_1 = \mathbf{0} \in \mathbb{R}^{|\mathcal{K}|}$
for $t = 1, 2, \ldots$ **do**
  Set $\lambda_{t,k} = B \frac{\exp\{\theta_k\}}{1+\sum_{k'\in\mathcal{K}}\exp\{\theta_{k'}\}}$ for all $k \in \mathcal{K}$
  $h_t \leftarrow \textsc{Best}_h(\boldsymbol{\lambda}_t)$
  $\widehat{Q}_t \leftarrow \frac{1}{t}\sum_{t'=1}^t h_{t'}$, $\overline{L} \leftarrow L\left(\widehat{Q}_t, \textsc{Best}_{\boldsymbol{\lambda}}(\widehat{Q}_t)\right)$
  $\widehat{\boldsymbol{\lambda}}_t \leftarrow \frac{1}{t}\sum_{t'=1}^t \boldsymbol{\lambda}_{t'}$, $\underline{L} \leftarrow L\left(\textsc{Best}_h(\widehat{\boldsymbol{\lambda}}_t), \widehat{\boldsymbol{\lambda}}_t\right)$
  $\nu_t \leftarrow \max\left\{L(\widehat{Q}_t, \widehat{\boldsymbol{\lambda}}_t) - \underline{L},\ \ \overline{L} - L(\widehat{Q}_t, \widehat{\boldsymbol{\lambda}}_t)\right\}$
  if $\nu_t \le \nu$ **then**
    Return $(\widehat{Q}_t, \widehat{\boldsymbol{\lambda}}_t)$
  end if
  Set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta\left(\mathbf{M}\widehat{\boldsymbol{\mu}}(h_t) - \widehat{\mathbf{c}}\right)$
end for

Agarwal, Alekh, et al. A reductions approach to fair classification. International conference on machine learning. PMLR, 2018.

In-Processing: Optimization

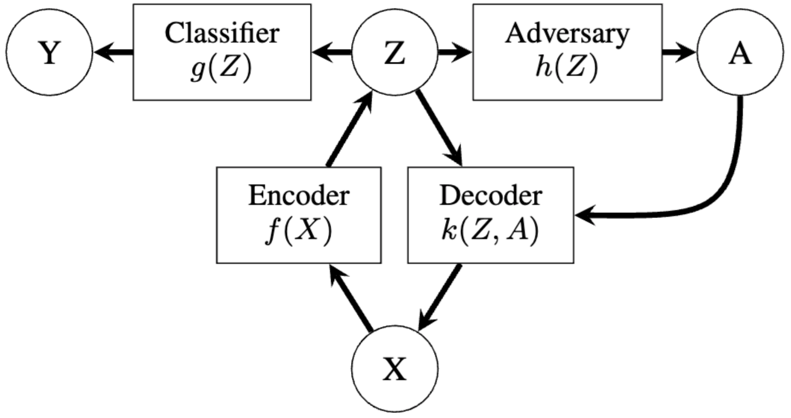
- **Experimental results**
  - Measure the trade-off between fairness and performance



Agarwal, Alekh, et al. A reductions approach to fair classification. International conference on machine learning. PMLR, 2018.

In-Processing: Learning Fair Representations for DP

- **Target**: demographic parity $\hat{Y} \perp S$
- **High-level idea**: the property of the representation Z
  - Z should can reconstruct X
  - Z should be able to predict Y
  - Z should not be able to predict S (A in the figure)



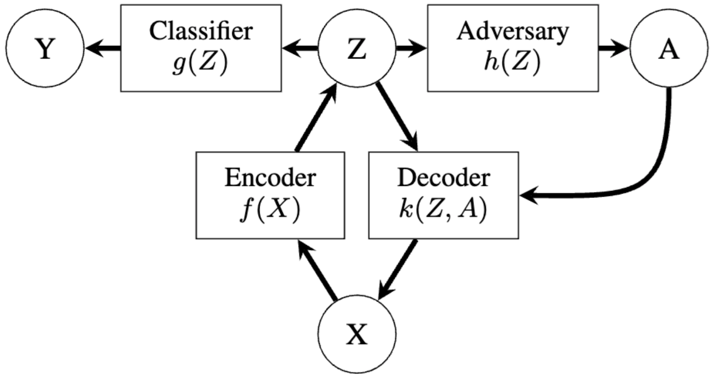
Madras, David, et al. Learning adversarially fair and transferable representations. International Conference on Machine Learning. PMLR, 2018.

In-Processing: Learning Fair Representations for DP

- **Target**: demographic parity $\hat{Y} \perp S$
- **Loss function**

$$\underset{f,g,k}{\text{minimize}}\ \underset{h}{\text{maximize}}\ \mathbb{E}_{X,Y,A}\left[L(f,g,h,k)\right]$$

$$\begin{aligned} L(f,g,h,k) = {} & \alpha L_C(g(f(X,A)),Y) \\ & + \beta L_{Dec}(k(f(X,A),A),X) \\ & + \gamma L_{Adv}(h(f(X,A)),A) \end{aligned}$$

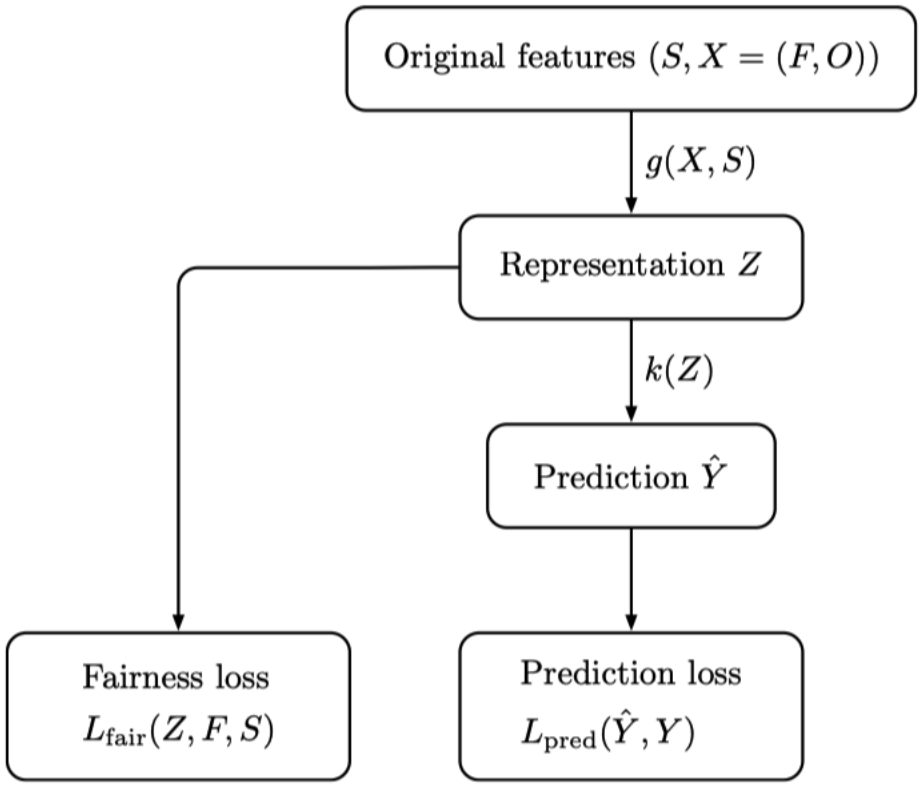


- L_C: prediction loss
- L_{Dec}: reconstruction loss
- L_{Adv}: fairness loss

Madras, David, et al. Learning adversarially fair and transferable representations. International Conference on Machine Learning. PMLR, 2018.

In-Processing: Learning Fair Representations for CF

- **Target:** conditional fairness $\hat{Y} \perp S \mid F$
- **Framework**
  - $g: (S, X) \to Z$, representation function
  - $k: Z \to \hat{Y}$, prediction function
- **Total loss function**
  - Prediction loss $L_{\text{pred}}(\hat{Y}, Y)$
  - Fairness loss $L_{\text{fair}}(Z, F, S)$
  - $L = L_{\text{pred}}\left(\hat{Y}, Y\right) + \lambda \cdot L_{\text{fair}}(Z, F, S)$
- **Challenge**
  - $Z \perp S \mid F \to L_{\text{fair}}(Z, F, S)$



Xu, Renzhe, et al. Algorithmic decision making with conditional fairness. KDD, 2020.

In-Processing: Learning Fair Representations for CF

- **Motivation**

Characterization of conditional independence (Daudin, 1980)

The random variables Z, S are independent conditional on F ($Z \perp S \mid F$) if and only if, for any function $u \in L^2_S$, $\tilde{h} \in \mathcal{E}_{ZF}$,

$$\mathbb{E}[u(S) \cdot \tilde{h}(Z, F)] = 0,$$

where

$$\begin{aligned} L^2_S &= \left\{u(S) \mid \mathbb{E}[u^2] < \infty\right\}, \\ L^2_{ZF} &= \left\{h(Z, F) \mid \mathbb{E}[h^2] < \infty\right\}, \\ \mathcal{E}_{ZF} &= \left\{\tilde{h}(Z, F) \in L^2_{ZF} \mid \mathbb{E}[\tilde{h}|F] = 0\right\}. \end{aligned}$$

- **Simplify**
  - Sensitive attribute S is binary

Xu, Renzhe, et al. Algorithmic decision making with conditional fairness. KDD, 2020.

In-Processing: Learning Fair Representations for CF

Derivable Conditional Fairness Regularizer

$$L_{\text{fair}}(Z,F,S) = \sup_h Q(h)$$

$$= \sup_h \left(C - \mathbb{E}\left[P(1-S|F)|h(Z,F)-S|\right]\right).$$

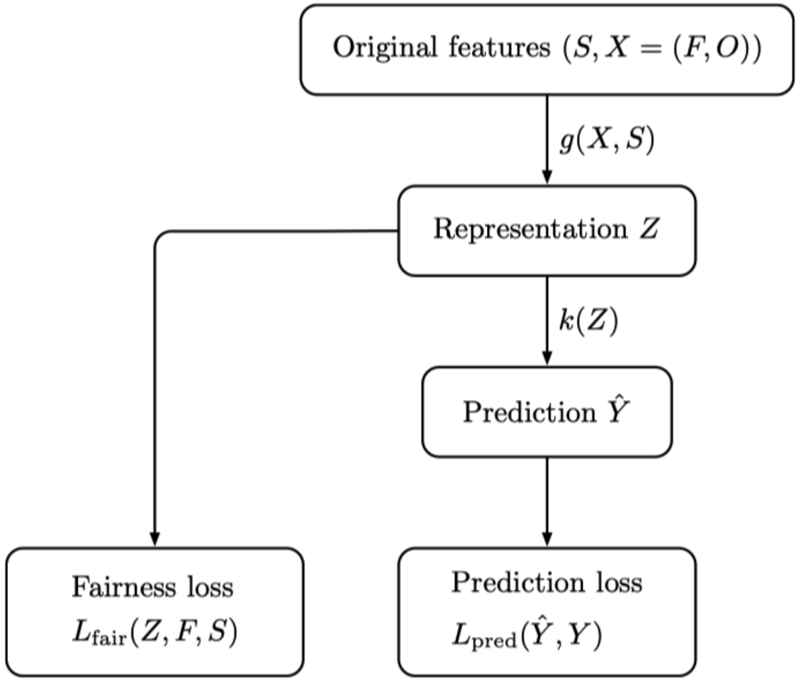
Here C is a constant.

- **Explanation**
  - $Q(h)$: weighted L1 loss when using $h(Z,F)$ to predict S.
- **Threoretic guarantee**
  - $L_{\text{fair}}(Z,F,S) = 0 \Leftrightarrow Z \perp S \mid F$

In-Processing: Learning Fair Representations for CF

- **Optimization target**

$$\min_{g,k} \sup_{h} \left(L_{\mathrm{pred}}(k(g(X,S)), Y) + \lambda Q(h)\right)$$



- **Special case**
  - Demographic parity ($F = \emptyset$)
    - This method becomes the same with the algorithm for DP

Xu, Renzhe, et al. Algorithmic decision making with conditional fairness. KDD, 2020.

In-Processing: Learning Fair Representations for CF

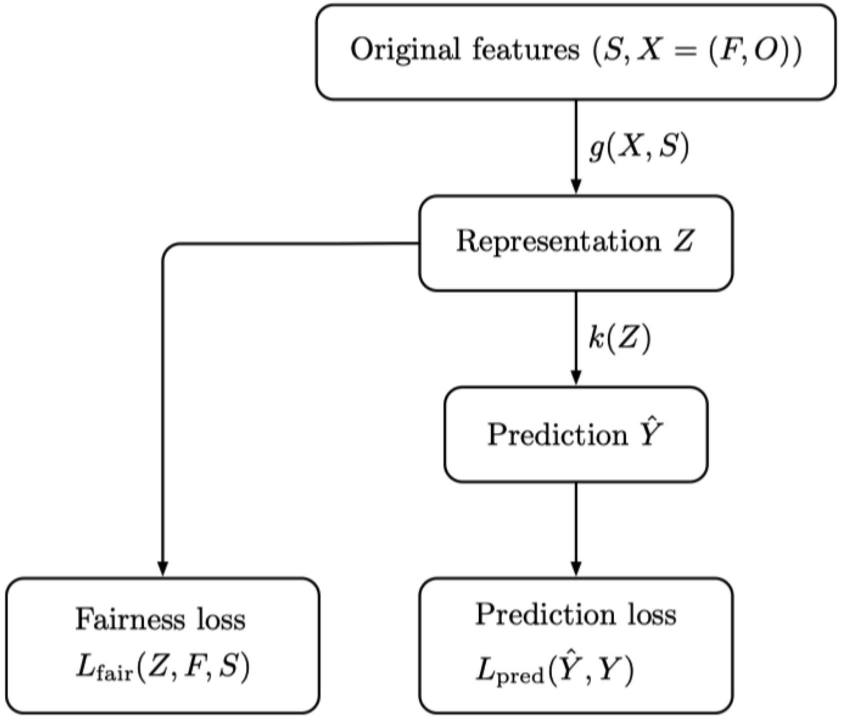
- **Practical Implementation**
  - L1 loss is difficult to optimize → L2 loss.

$$L_{\mathrm{fair}}(Z,F,S) = C - \sup_{h} \mathbb{E}\left[P(1-S|F)|h(Z,F)-S|\right]$$

$$\Downarrow$$

$$L'_{\mathrm{fair}}(Z,F,S) = C - \sup_{h} \mathbb{E}\left[P(1-S|F)(h(Z,F)-S)^2\right]$$

- **Theoretical guarantee**
  - $L'_{\mathrm{fair}}(Z,F,S) \geq L_{\mathrm{fair}}(Z,F,S)$



Xu, Renzhe, et al. Algorithmic decision making with conditional fairness. KDD, 2020.

In-Processing: Learning Fair Representations for CF

- **Results**
  - Plot the accuracy-fairness trade curve.
  - The method could effectively balance fairness and performance

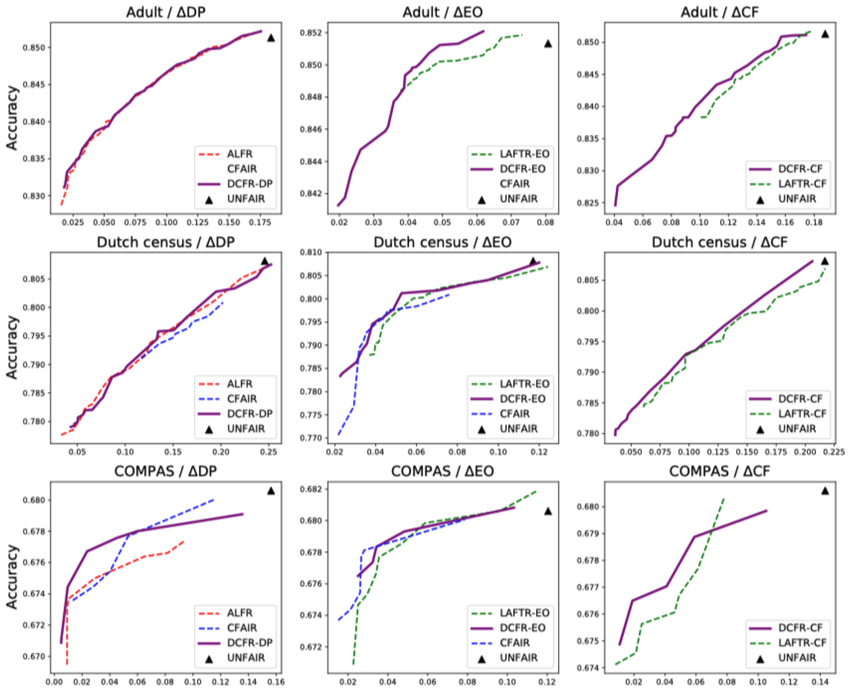


Figure: Accuracy-fairness trade-off curve. The upper-left corner is preferred.

Xu, Renzhe, et al. Algorithmic decision making with conditional fairness. KDD, 2020.

Post-Processing

- **General idea**: post-hoc adjustment of existing method
- **Procedure**
  - For binary classification, learn a Bayes-optimal score function $R(x, s) \in [0,1]$
  - Set different threshold for different subgroups
$$\hat{Y}(x, s) = \mathbb{I}(R(x, s) \geq T_s)$$
  - Here T_s is the threshold on subgroup s

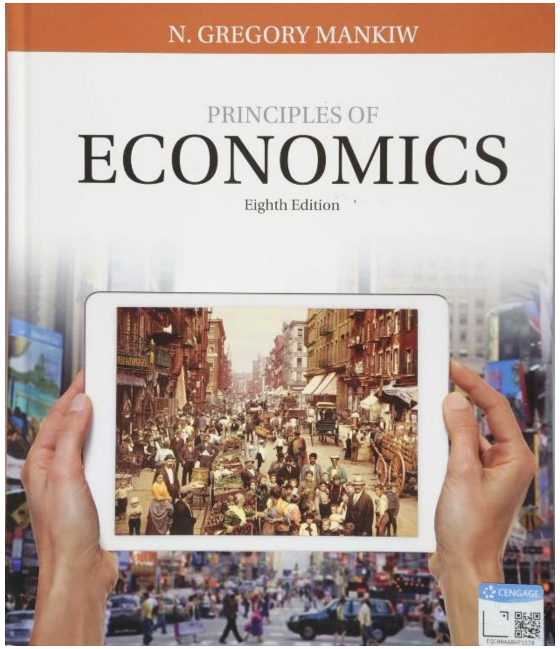
Hardt, Moritz, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. NeurIPS, 2016.

Outline

- Part I: Background
- Part II: Definitions of fairness in machine learning
- Part III: Algorithms of fairness in machine learning
- Part IV: Fairness In Economics
- Part V: Fairness In Language Models

Fairness is Important in Economics

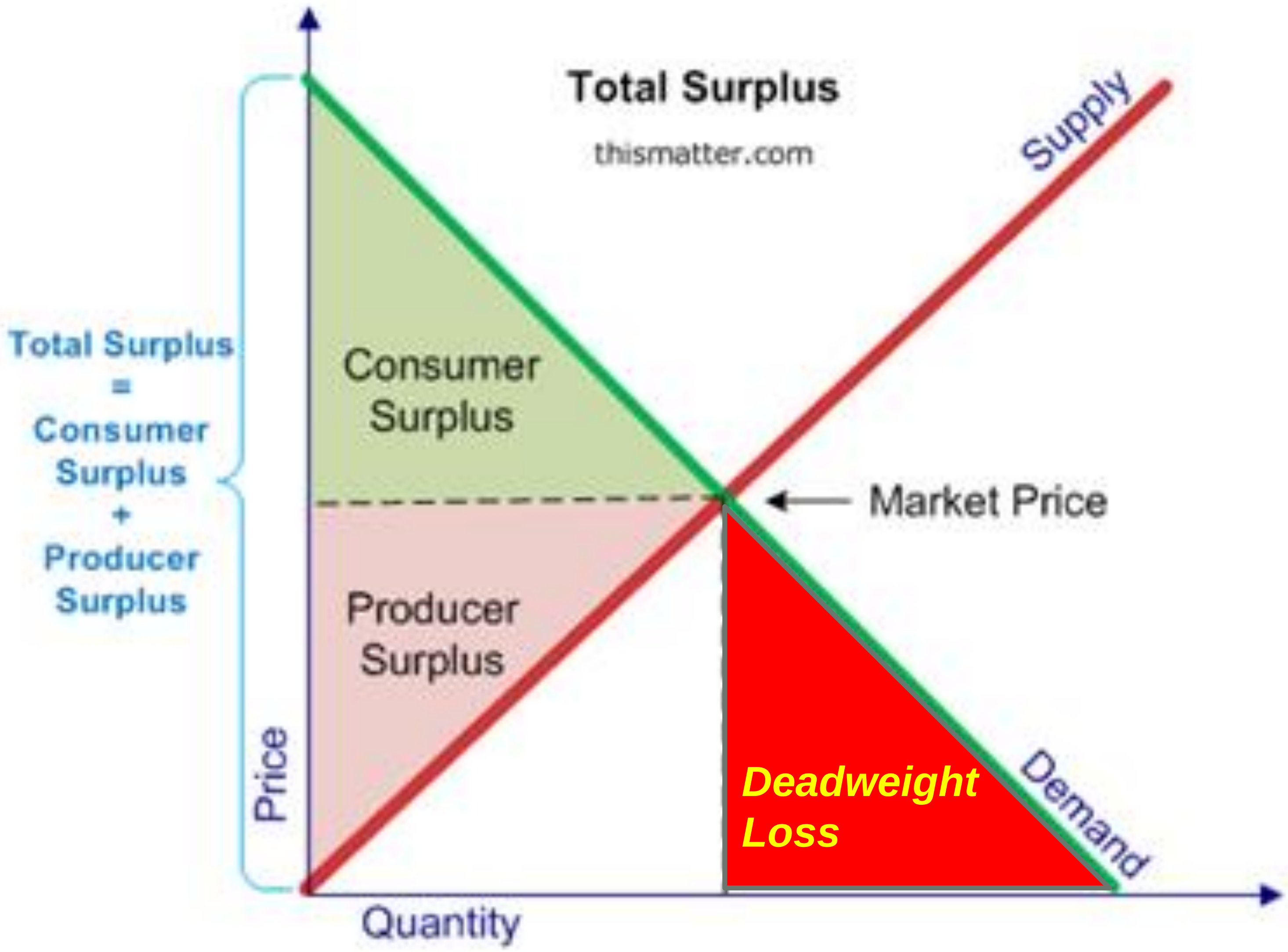
- In the famous book *Principles of Economics*
  - “People Face Trade-offs” is the first principle in Ten Principles of Economics mentioned
  - **Efficiency and equality** is an important trade-off faced by society



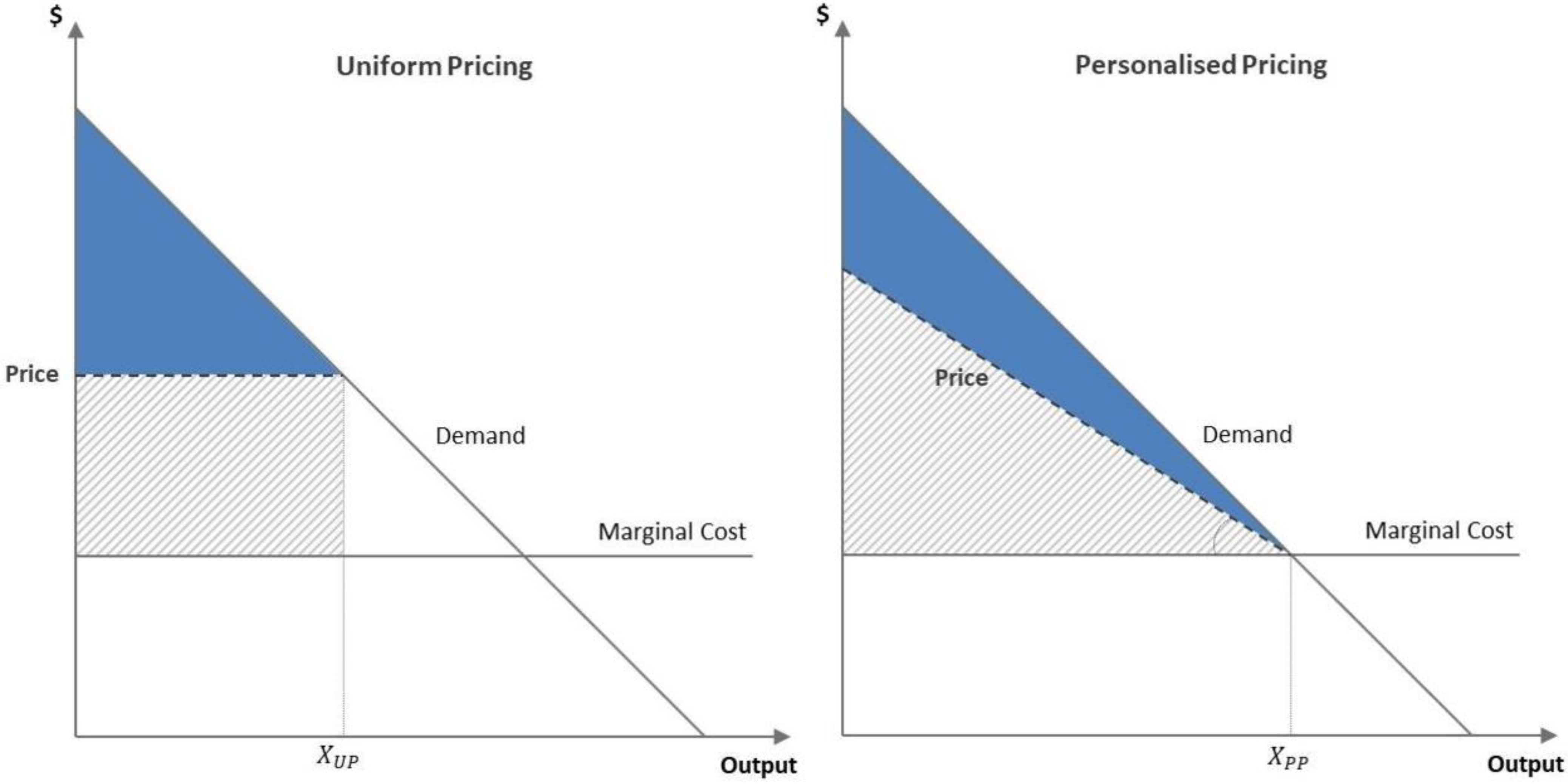
Fairness in Economics

- Fairness is important in many scenarios
- The profits should be allocated to different agents in a fair way
- **Some applications**
  - Markets
  - Resource allocation
  - ...

Market: Consumer v.s. Producer



Personalized Pricing



Regulation Instruments over Personalized Pricing

- **Target**
  - To design effective policy instruments to balance benefits between consumers and producers
- **Challenge**
  - Improper regulatory policies may be harmful to consumers.
    - Example --- 6 people with willingness to pay $1, 2, 3, 5, 6, 7

Market segments	Optimal pricing strategy	Producer surplus	Consumer surplus	Total surplus
{1, 2, 3, 5, 6, 7}	$5	$15	$3	$18
{1}, {2, 3, 5, 6, 7}	$1, $5	$16	$3	$19
{1}, {2, 3}, {5, 6, 7}	$1, $2, $5	$20	$4	$24
{1}, {2}, {3}, {5}, {6}, {7}	$1, $2, $3, $5, $6, $7	$24	$0	$24

Dubé, Jean-Pierre, and Sanjog Misra. Personalized pricing and consumer welfare. Journal of Political Economy, 2023.

Problem Setup

- **Basic setup**
  - A single monopoly sell a single product to various consumers with fixed marginal cost c
- **Willingness to pay**
  - V: consumers' willingness to pay, drawn from the demand distribution F
  - The monopoly could precisely estimate consumers' willingness to pay and make personalized prices accordingly.
  - A consumer with willingness to pay V buys the product $\Leftrightarrow$ V exceeds the charged price

Xu, Renzhe, et al. Regulatory instruments for fair personalized pricing. Proceedings of the ACM Web Conference, 2022.

Problem Setup

- **Assumption on the demand distribution**
  - monotone hazard rate distribution (uniform, exponential, logistic)
  - strongly regular (some power law)

- **Explanation**
  - Assumption on the 'tail' of the demand distribution

Xu, Renzhe, et al. Regulatory instruments for fair personalized pricing. Proceedings of the ACM Web Conference, 2022.

Overview of Results

- **Two regulatory policies**
  - $\boldsymbol{\epsilon}$-difference fair: $p_u - p_l \leq \epsilon$
  - $\boldsymbol{\gamma}$-ratio fair: $\frac{p_u - c}{p_l - c} \leq \gamma$
- **Theoretical analysis of the two policies**
  - For common demand distributions
    - Stricter constraints → increasing consumer surplus, decreasing producer surplus
    - Stricter constraints → drop on total surplus
    - ϵ-difference achieves better consumer-producer trade-off.

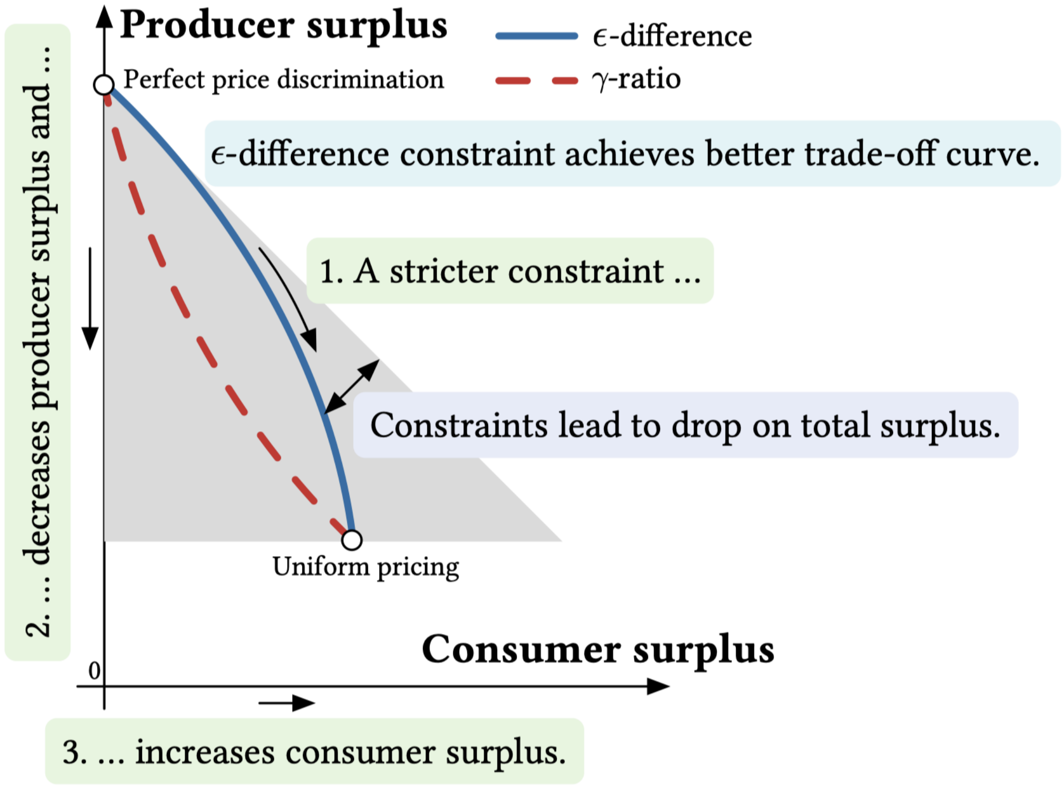
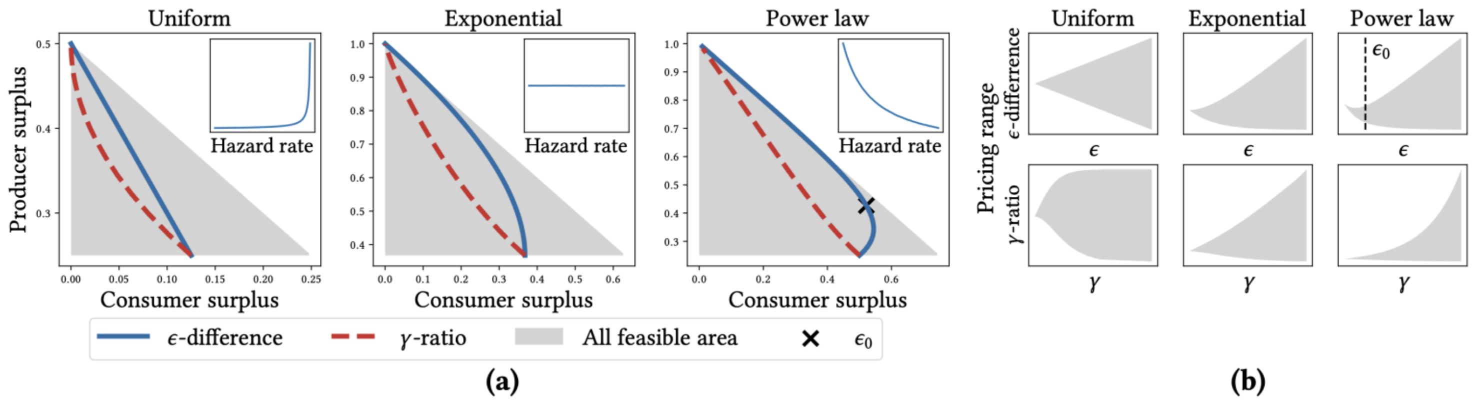


Figure 1: Graphical explanations of our major findings.

Xu, Renzhe, et al. Regulatory instruments for fair personalized pricing. Proceedings of the ACM Web Conference, 2022.

Experiments

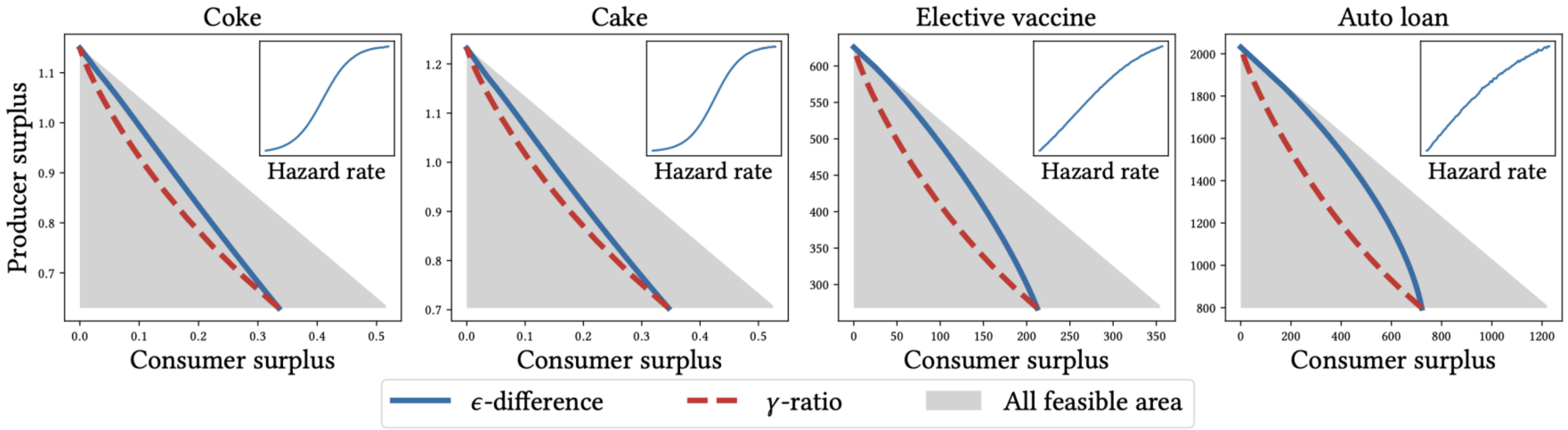
- **Simulation**
  - Uniform / exponential / power-law demand distributions
- **Results**
  - Balancing consumer surplus and producer surplus
  - Drop on total surplus
  - ϵ-difference constraint vs γ-ratio constraint



Xu, Renzhe, et al. Regulatory instruments for fair personalized pricing. Proceedings of the ACM Web Conference, 2022.

Experiments

- **Real-world datasets**
  - Coke and cake
    - Demand distribution has monotone hazard rate (MHR)
  - Elective vaccine and auto loan
    - Demand distribution has MHR from the long-run trend, though existing fluctuations in short-run



Xu, Renzhe, et al. "Regulatory instruments for fair personalized pricing." *Proceedings of the ACM Web Conference 2022*. 2022.

Outline

- Part I: Background
- Part II: Definitions of fairness in machine learning
- Part III: Algorithms of fairness in machine learning
- Part IV: Fairness In Economics
- Part V: Fairness In Language Models

Taxonomy of Social Biases in NLP

Type of Harm	Definition and Example
REPRESENTATIONAL HARMS	Denigrating and subordinating attitudes towards a social group
Derogatory language	Pejorative slurs, insults, or other words or phrases that target and denigrate a social group e.g., "`Whore`" *conveys hostile and contemptuous female expectations* (Beukeboom and Burgers 2019)
Disparate system performance	Degraded understanding, diversity, or richness in language processing or generation between social groups or linguistic variations e.g., *AAE* like* "`he woke af`" *is misclassified as not English more often than SAE*[†] *equivalents* (Blodgett and O'Connor 2017)
Erasure	Omission or invisibility of the language and experiences of a social group e.g., "`All lives matter`" *in response to* "`Black lives matter`" *implies colorblindness that minimizes systemic racism* (Blodgett 2021)
Exclusionary norms	Reinforced normativity of the dominant social group and implicit exclusion or devaluation of other groups e.g., "`Both genders`" *excludes non-binary identities* (Bender et al. 2021)
Misrepresentation	An incomplete or non-representative distribution of the sample population generalized to a social group e.g., *Responding* "`I'm sorry to hear that`" *to* "`I'm an autistic dad`" *conveys a negative misrepresentation of autism* (Smith et al. 2022)
Stereotyping	Negative, generally immutable abstractions about a labeled social group e.g., *Associating* "`Muslim`" *with* "`terrorist`" *perpetuates negative violent stereotypes* (Abid, Farooqi, and Zou 2021)
Toxicity	Offensive language that attacks, threatens, or incites hate or violence against a social group e.g., "`I hate Latinos`" *is disrespectful and hateful* (Dixon et al. 2018)

Type of Harm	Definition and Example
ALLOCATIONAL HARMS	Disparate distribution of resources or opportunities between social groups
Direct discrimination	Disparate treatment due explicitly to membership of a social group e.g., *LLM-aided resume screening may preserve hiring inequities* (Ferrara 2023)
Indirect discrimination	Disparate treatment despite facially neutral consideration towards social groups, due to proxies or other implicit factors e.g., *LLM-aided healthcare tools may use proxies associated with demographic factors that exacerbate inequities in patient care* (Ferrara 2023)

Gallegos, Isabel O., et al. Bias and fairness in large language models: A survey. Computational Linguistics, 2024.

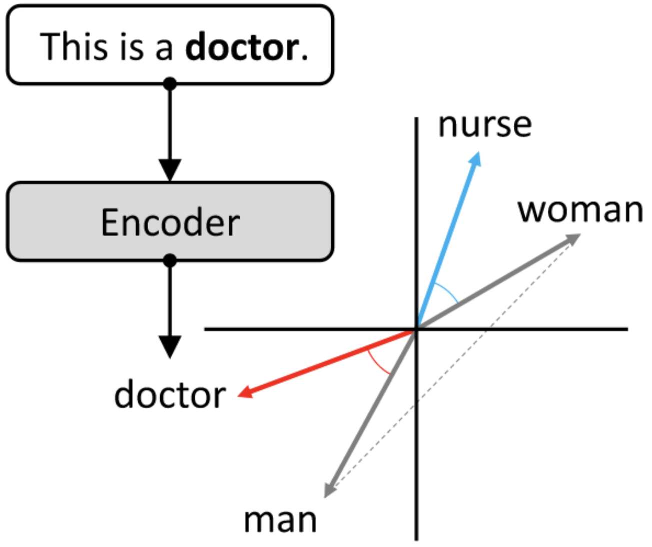
Taxonomy of Metrics for Model Bias Evaluation

- Embedding-based metrics
  - Use vector hidden representations
- Probability-based metrics
  - Use model-assigned token probabilities
- Generated text-based metrics
  - Use model-generated text continuations

Gallegos, Isabel O., et al. Bias and fairness in large language models: A survey. Computational Linguistics, 2024.

Taxonomy of Metrics for Model Bias Evaluation: Embedding-Based Metrics

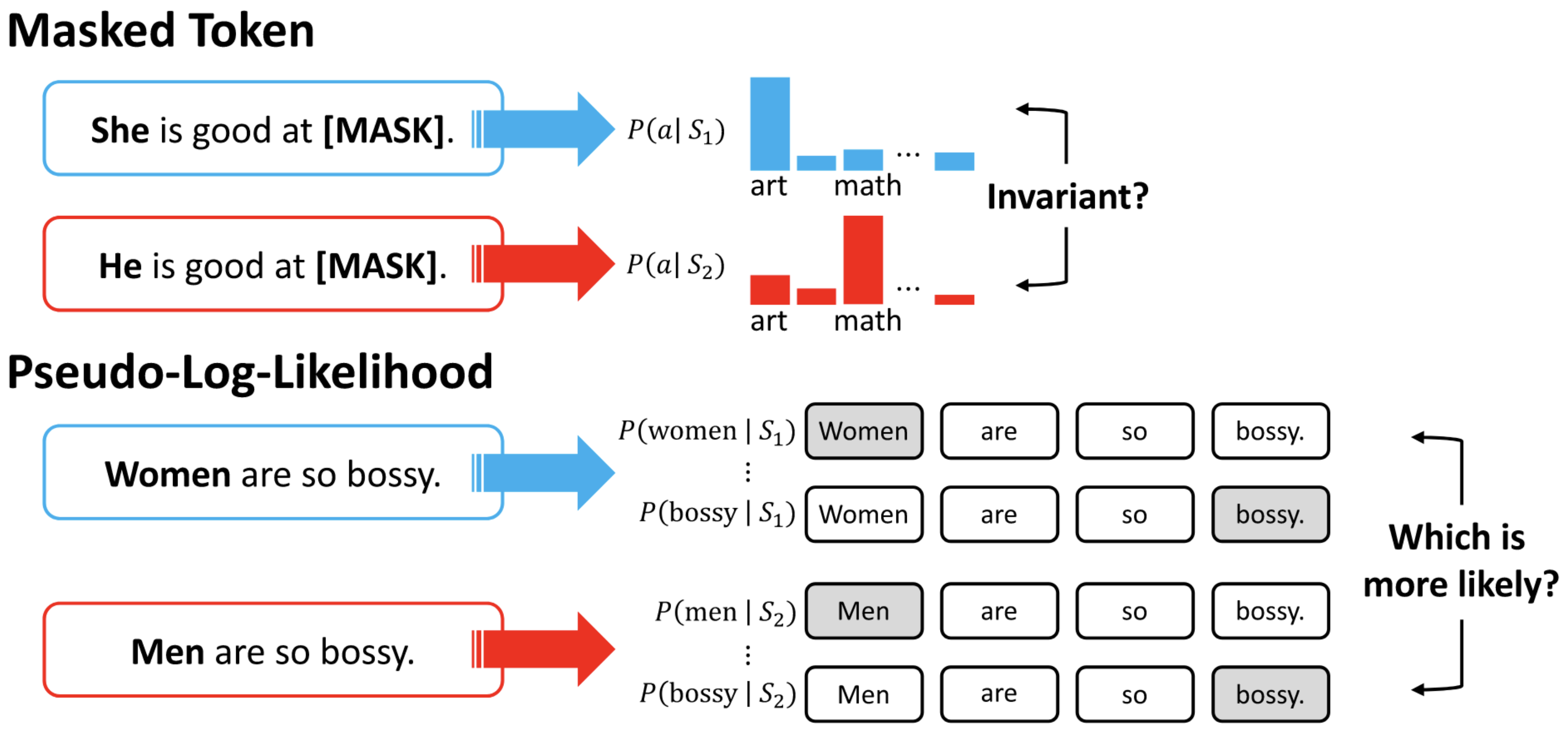
- Word embedding metrics
  - E.g., computing cosine distances between neutral and gendered words
- Sentence embedding metrics
  - Use the embedding of sentences instead of words



Gallegos, Isabel O., et al. Bias and fairness in large language models: A survey. Computational Linguistics, 2024.

Taxonomy of Metrics for Model Bias Evaluation: Probability-Based Metrics

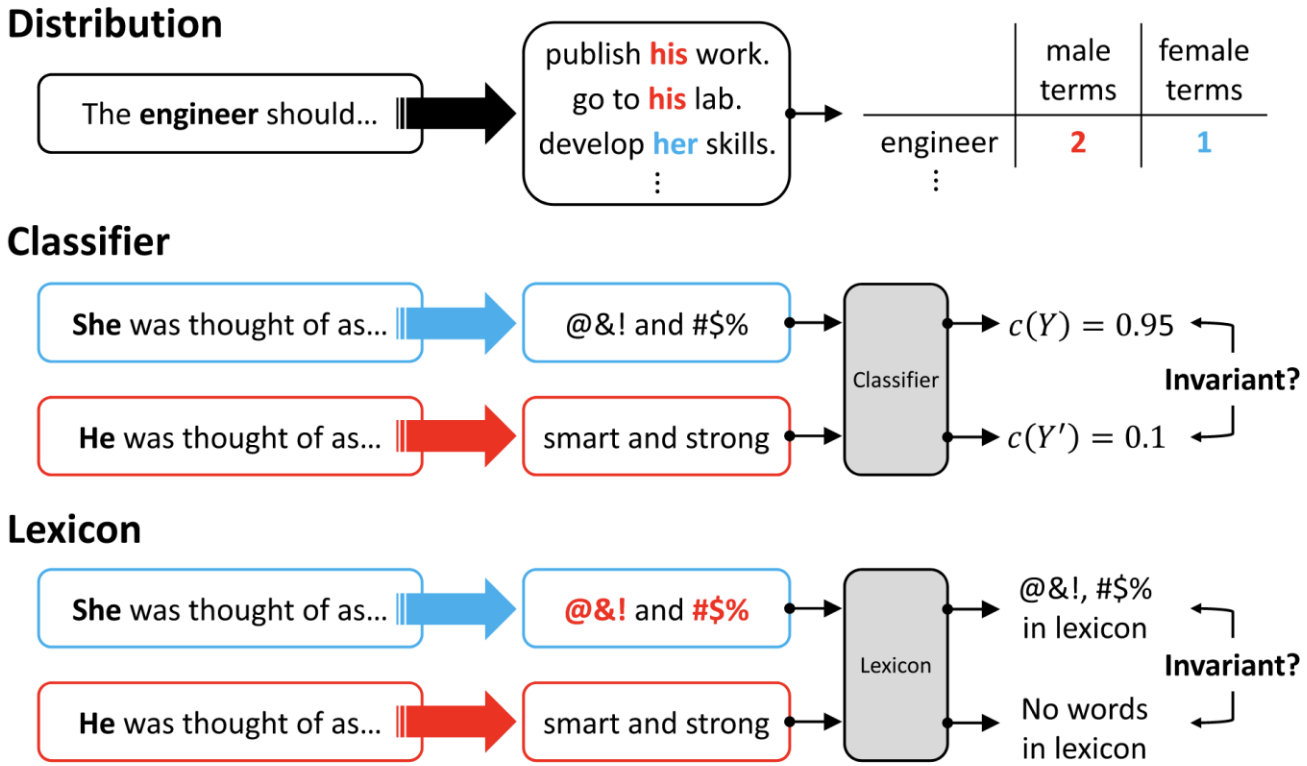
- Masked token methods
- Pseudo-log-likelihood methods



Gallegos, Isabel O., et al. Bias and fairness in large language models: A survey. Computational Linguistics, 2024.

Taxonomy of Metrics for Model Bias Evaluation: Generated Text-Based Metrics

- More useful when LLM is a black box (we do not have embeddings or probabilities)
- Three types
  - Distribution metrics
  - Classifier metrics
  - Lexicon metrics



Gallegos, Isabel O., et al. Bias and fairness in large language models: A survey. Computational Linguistics, 2024.

Taxonomy of Techniques for Bias Mitigation

Mitigation Stage	Mechanism
PRE-PROCESSING (§ 5.1)	Data Augmentation (§ 5.1.1)
	Data Filtering & Reweighting (§ 5.1.2)
	Data Generation (§ 5.1.3)
	Instruction Tuning (§ 5.1.4)
	Projection-based Mitigation (§ 5.1.5)
IN-TRAINING (§ 5.2)	Architecture Modification (§ 5.2.1)
	Loss Function Modification (§ 5.2.2)
	Selective Parameter Updating (§ 5.2.3)
	Filtering Model Parameters (§ 5.2.4)
INTRA-PROCESSING (§ 5.3)	Decoding Strategy Modification (§ 5.3.1)
	Weight Redistribution (§ 5.3.2)
	Modular Debiasing Networks (§ 5.3.3)
POST-PROCESSING (§ 5.4)	Rewriting (§ 5.4.1)

Gallegos, Isabel O., et al. Bias and fairness in large language models: A survey. Computational Linguistics, 2024.

Recommended Papers for In-depth Exploration

[1] Dwork C, Hardt M, Pitassi T, et al. Fairness Through Awareness[C/OL]//Innovations in Theoretical Computer Science Conference. 2012[2019-08-17]. http://arxiv.org/abs/1104.3913.

[2] Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning[C/OL]//Advances in Neural Information Processing Systems. 2016[2019-07-18]. http://arxiv.org/abs/1610.02413.

[3] Agarwal A, Beygelzimer A, Dudík M, et al. A Reductions Approach to Fair Classification[C/OL]//International Conference on Machine Learning. 2018[2020-08-10]. http://arxiv.org/abs/1803.02453.

[4] Kusner M J, Loftus J R, Russell C, et al. Counterfactual Fairness[C/OL]//Advances in Neural Information Processing Systems. 2018[2019-12-21]. http://arxiv.org/abs/1703.06856.

Thanks!

Peng Cui
cuip@tsinghua.edu.cn
http://pengcui.thumedialab.com