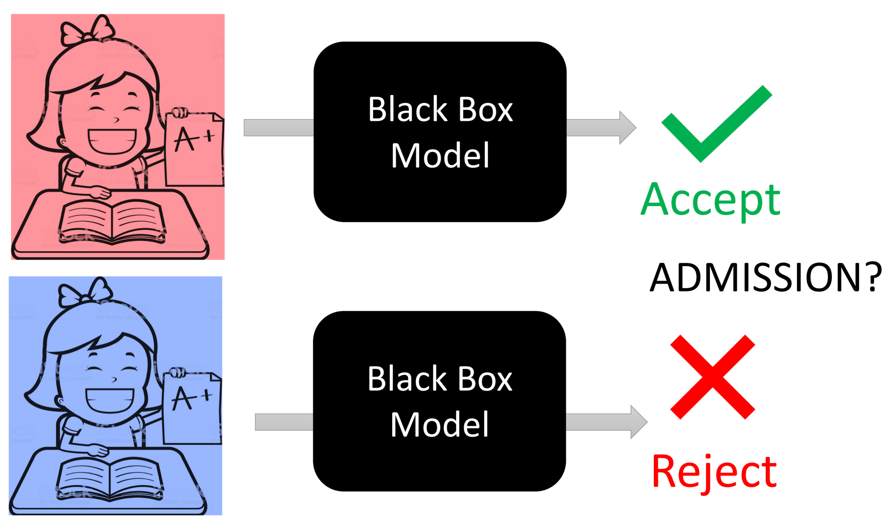
Carnegie Mellon

An Information-Theoretic Quantification of Discrimination with Exempt Features

Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, Pulkit Grover

Motivation



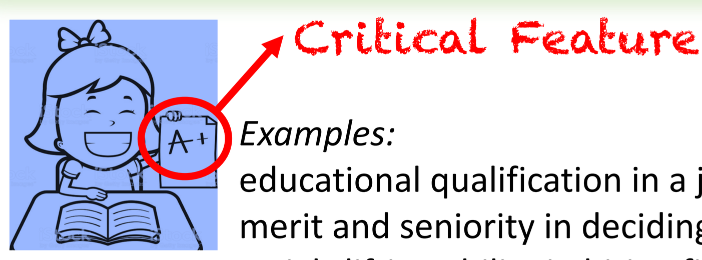
Fairness without Domain Knowledge?

Statistical Parity: Accepts unqualified members of protected group [Zemel et. al. '13][Hardt et. al.'16]

Equalized Odds: Agreement with true labels, may be affected by label bias [Hinnefeld '18] [Barocas & Selbst '16]

Decision may not place weight on critical features.

Goal: Quantify "non-exempt" discrimination while allowing for exemptions due to critical features



Exemptions in Law: Bona-Fide Occupational Qualification (BFOQ)

Examples: educational qualification in a job, merit and seniority in deciding salary, weightlifting ability in hiring firemen

Contributions

Quantification of "non-exempt" discrimination while allowing for exemptions due to critical features

Axiomatic approach, Counterexamples to existing works

E.g., unfair by Conditional Statistical Parity, but fair by Counterfactual Fairness

Lead us to examine Partial Information Decomposition (PID)

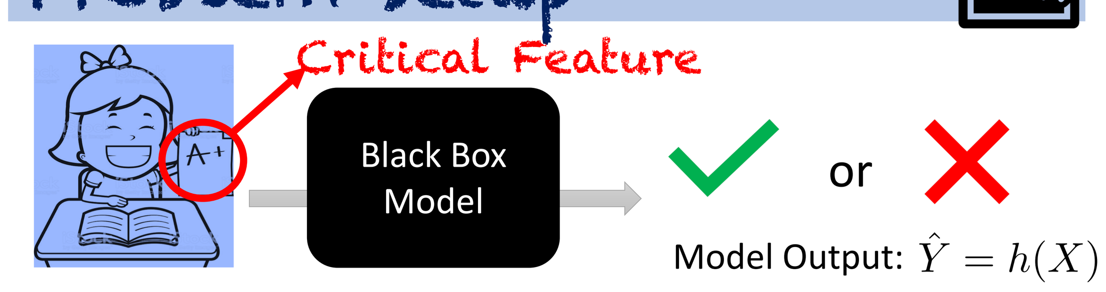
Conditional Mutual Information → Unique Information, Synergistic Information

Propose a novel "counterfactual" measure of non-exempt discrimination

Difficult to realize in practice?

Observational measures of "non-exempt" discrimination (impossibility, utility and limitations)

Problem Setup

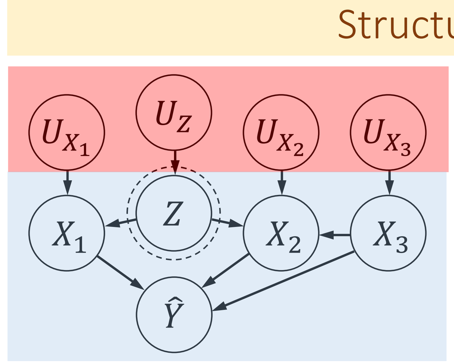


Model Output: $\hat{Y}=h(X)$

Total Features: $X=(X_c,X_g)$ — Critical, Non-Critical/General

Protected Attribute: Z

Structural Causal Model



Unobserved Latent Social Factors (U's) *Mutually Independent*

Observables $X, Z, \hat{Y}$ *Directed Acyclic Graph*

Z has no parents

$\hat{Y}=f(Z,U_X)$

$Z \perp U_X$

Counterfactual Causal Influence (CCI) [Kusner et. al.'17][Datta et. al.'17][Russell et. al.'17]

$$E_{Z,Z',U_X}\|f(Z,U_X)-f(Z',U_X)\|_2$$

Information-Theoretic Equivalent of CCI

Total Discrimination: $I(Z;W)$

$$W=[f(Z,u_a),f(Z,u_b),\dots]\quad \forall u \quad \Pr(U_X=u)>0$$

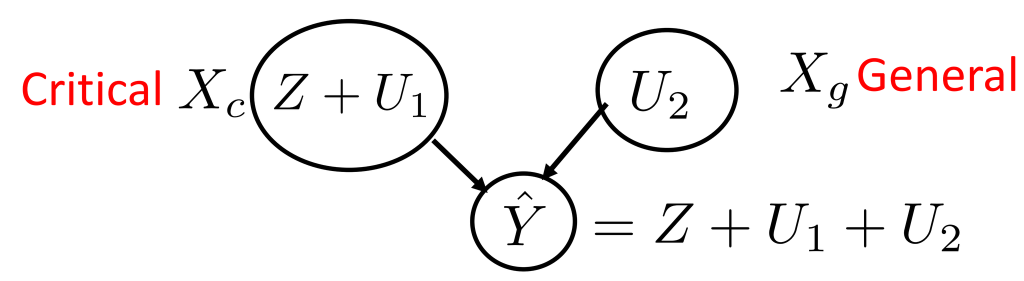
Thought Experiments

Total Discrimination: $I(Z;W)$ → M_{NE} Non-Exempt, M_E Exempt

Candidate Measures of Non-Exempt Discrimination

Candidate 1: $I(Z;\hat{Y})$

Counterexample: Hiring actors for a male role (BFOQ Defense)



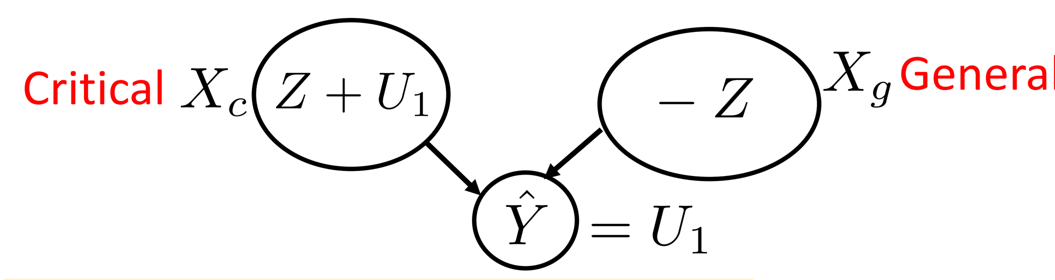
$Z - X_c - \hat{Y}$

M_{NE} should be 0, but ...

$I(Z;\hat{Y})>0$

Candidate 2: $I(Z;\hat{Y}|X_c)$

Counterexample: College admissions (U_1 is the true ability of a candidate)



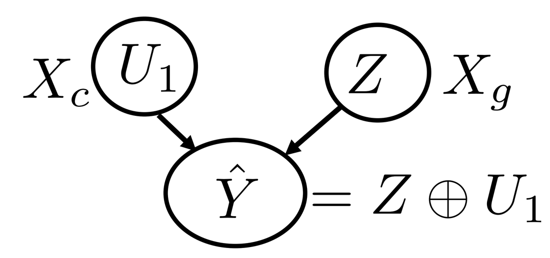
Counterfactually fair

M_{NE} should be 0, but ...

$I(Z;U_1|Z+U_1)>0$

Candidate 3: $Uni(Z:\hat{Y}\backslash X_c)$

Counterexample: Expensive Housing ad shown to high-income Race A and low-income Race B (largely irrelevant to latter)



$Z,U_1\sim Bern(1/2)$

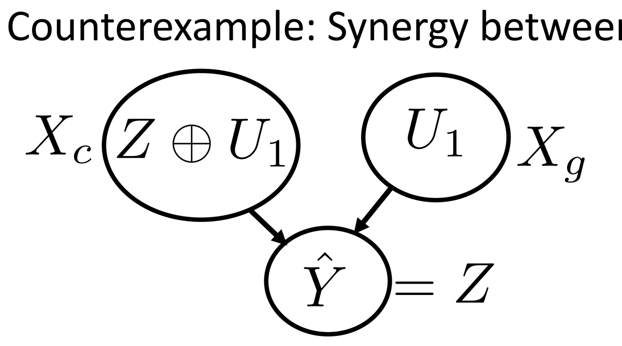
Masked Discrimination High-income Race B

M_{NE} should not be 0, but ...

$Uni(Z:\hat{Y}\backslash X_c)=0$

Candidate 4: Path-Specific Causal Influence

Counterexample: Synergy between critical and general



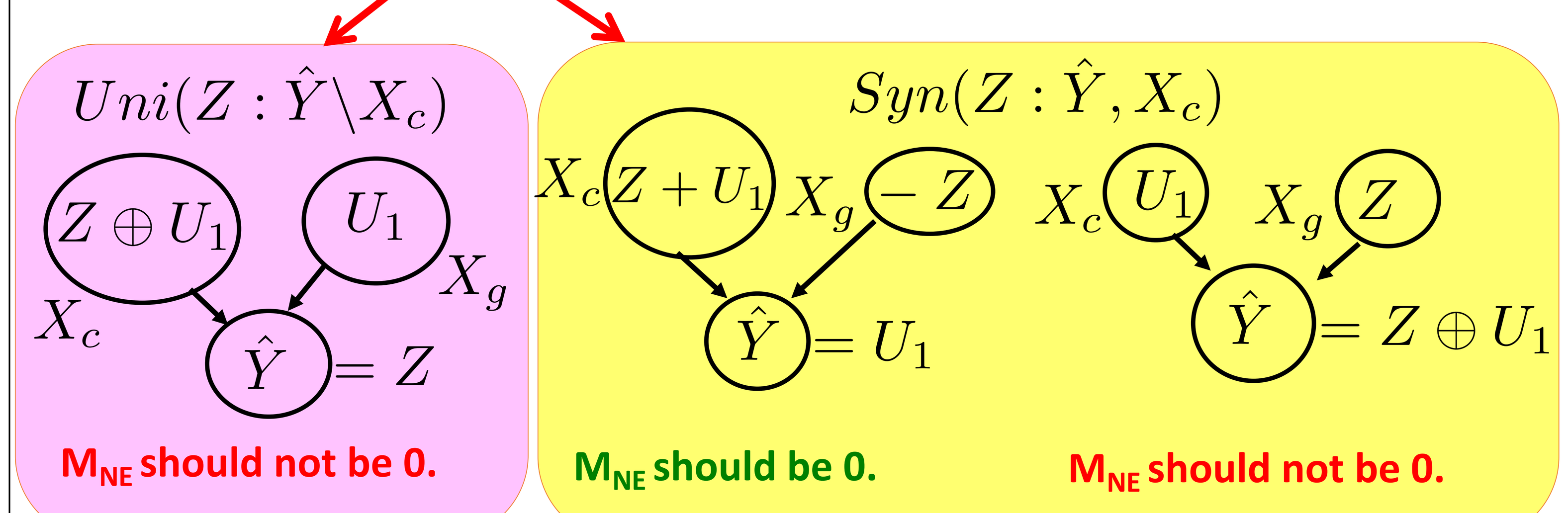
$Uni(Z:\hat{Y}\backslash X_c)>0$

M_{NE} should not be 0,but ...

Path-Specific examination qualifies exempt.

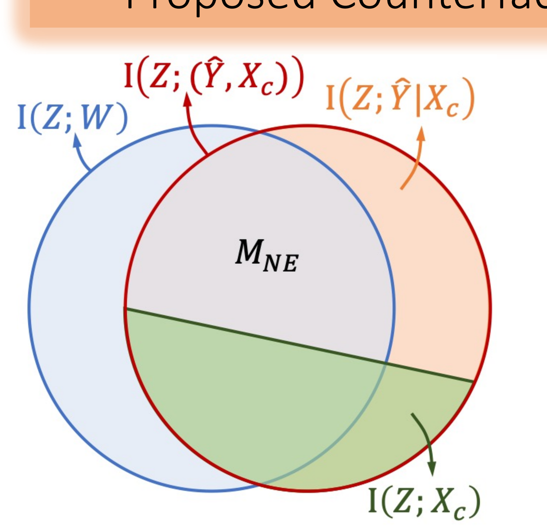
Understanding Scenarios where $I(Z;\hat{Y}|X_c)>0$

$I(Z;\hat{Y}|X_c)$ Partial Information Decomposition (PID) [Bertschinger et al. '14]



$I(Z;W)>0$ CCI present | $I(Z;W)=0$ No CCI | $I(Z;W)>0$ CCI present

Proposed Counterfactual Measure of Non-Exempt Discrimination



Information-theoretic sub-volume between $I(Z;\hat{Y}|X_c)$ and $I(Z;W)$

$$Uni(Z:W\backslash X_c)-Uni(Z:W\backslash \hat{Y},X_c)$$

satisfies all desirable properties

Main Results

Desirable Properties

Property 1: M_{NE} should be 0 if $X_c=X$.

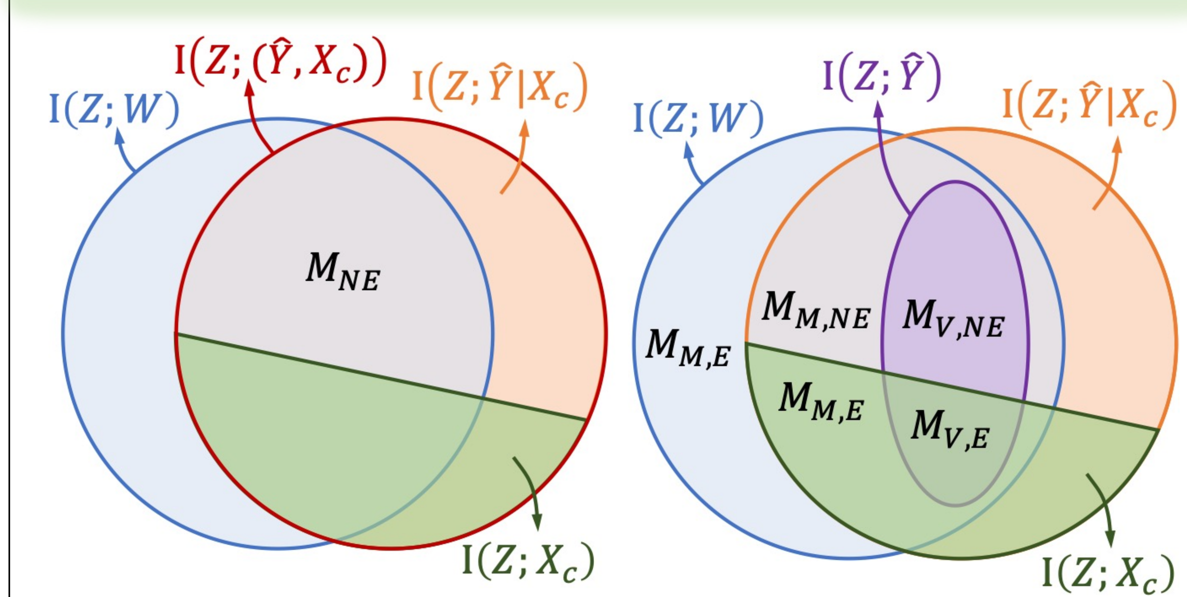
Property 2: M_{NE} should be greater than 0 if $Uni(Z:\hat{Y}\backslash X_c)>0$.

Property 3: M_{NE} should be greater than 0 in the canonical example of masked discrimination.

Property 4: M_{NE} should be 0 if $I(Z;W)=0$.

Theorem 1: Our proposed measure satisfies all four properties.

Theorem 2: Total discrimination $I(Z;W)$ decomposes into four non-negative components: visible exempt, visible non-exempt, masked exempt, masked non-exempt.



Observational Measures: Impossibility, Utility, Limitation

Theorem 3: No observational measure can satisfy Properties 3 and 4 together.

$Uni(Z:\hat{Y}\backslash X_c)$	Satisfies all properties except the property of non-exempt masked discrimination (Prop. 3).
$I(Z;\hat{Y}\|X_c)$	Captures masked discrimination but gives false-positives under cancellation (Prop. 4).
$I(Z;\hat{Y}\|X_c,X')$	Satisfies only Prop. 1, others are satisfied partially with some counterexamples.

A Simple Case Study

Goal: Decide whether to show ads for an editorial job
Protected attribute Z: native English speaker or not

X_1: a score based on online writing samples — **Critical**

X_2: a score based on browsing history, e.g., interest in English websites as compared to websites of other languages

X_3: a preference score based on geographical proximity.

$X_1=Z+U_1$ $X_2=Z+U_2$ $X_3=U_3$

Ground Truth: $Y=\mathcal{I}(S\geq 1)$ where $S=X_1+X_2+X_3$

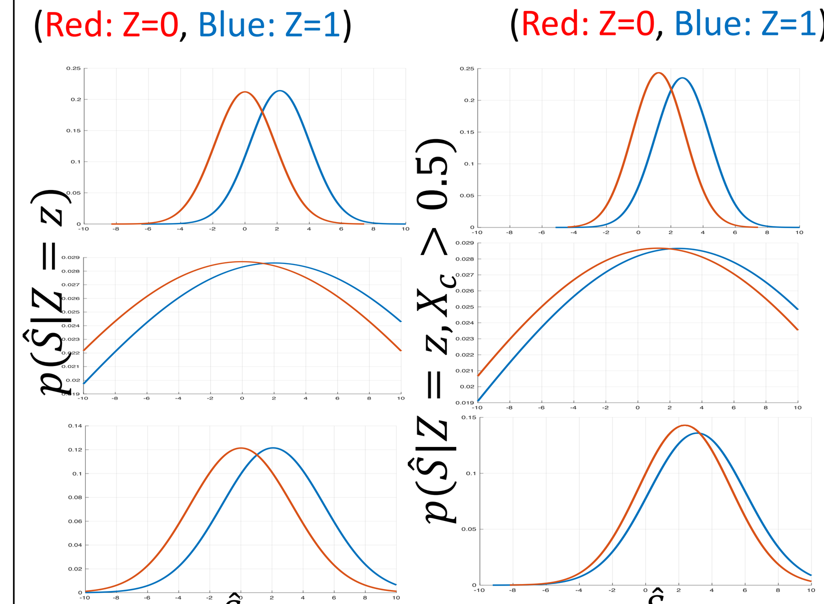
Equal Weight for all Features

$Z\sim \text{Bern}(1/2)$

$U_1,U_2,U_3\sim\mathcal{N}(0,1)$

Histograms of predicted scores for all candidates (Red: Z=0, Blue: Z=1)

Histogram of predicted scores for those satisfying critical necessity (Red: Z=0, Blue: Z=1)



Loss 1 (No fairness): equal weight on all features (Acc:98.5%)

Loss 2 (MI): low weight on correlated features X_1 and X_2 (Acc:70.2%)

Loss 3 (CMI): low weight on X_2, the correlated feature in X_g (Acc:80.8%)

Electrical & Computer ENGINEERING

Ack: (1) K&L Gates Presidential Fellowship in Ethics and Computational Technologies (2) NSF