Fairness and Explainability: Bridging the Gap Towards Fair Model Explanations


Yuying Zhao

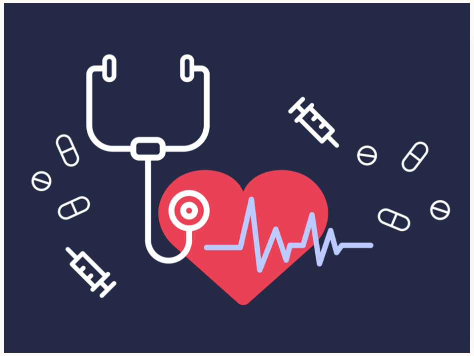

Yu Wang


Tyler Derr

yuying.zhao@vanderbilt.edu
https://yuyingzhao.github.io/

Background: Bias and Fairness in ML

Healthcare



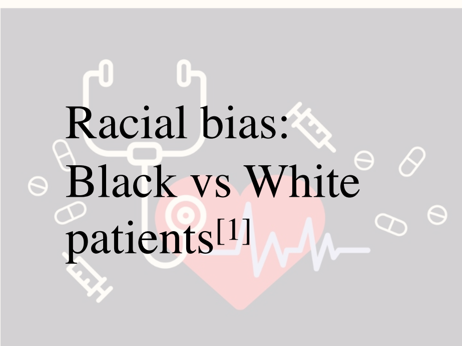
Education

Finance

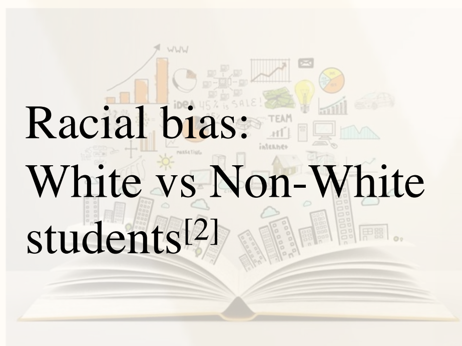
Utility Performance

Fairness Concerns

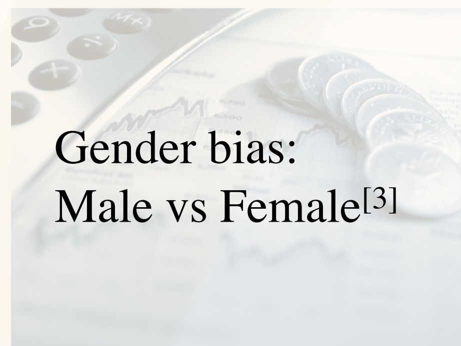
Healthcare



Education



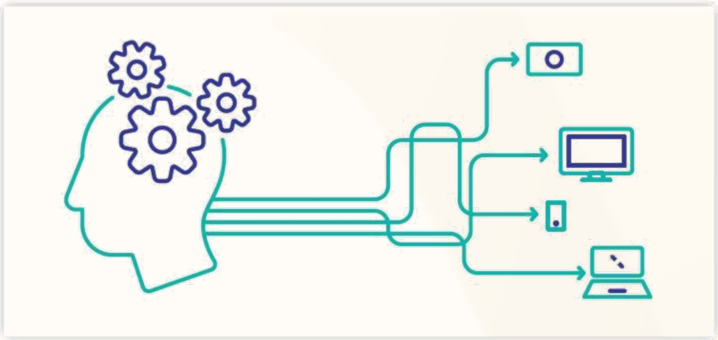
Finance



[1] Obermeyer, Ziad, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science. 2019.
[2] Anderson, Henry, et al. "Assessing the Fairness of Graduation Predictions." EDM. 2019.
[3] Zhang, Yukun, et al. "Fairness assessment for artificial intelligence in financial industry." NeurIPS. 2019.

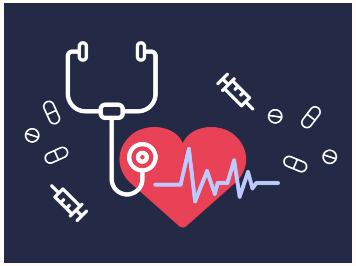
Background: Model Explainability

Explainability



- Why should I trust the model?
- Why did a model make a certain decision?

Healthcare



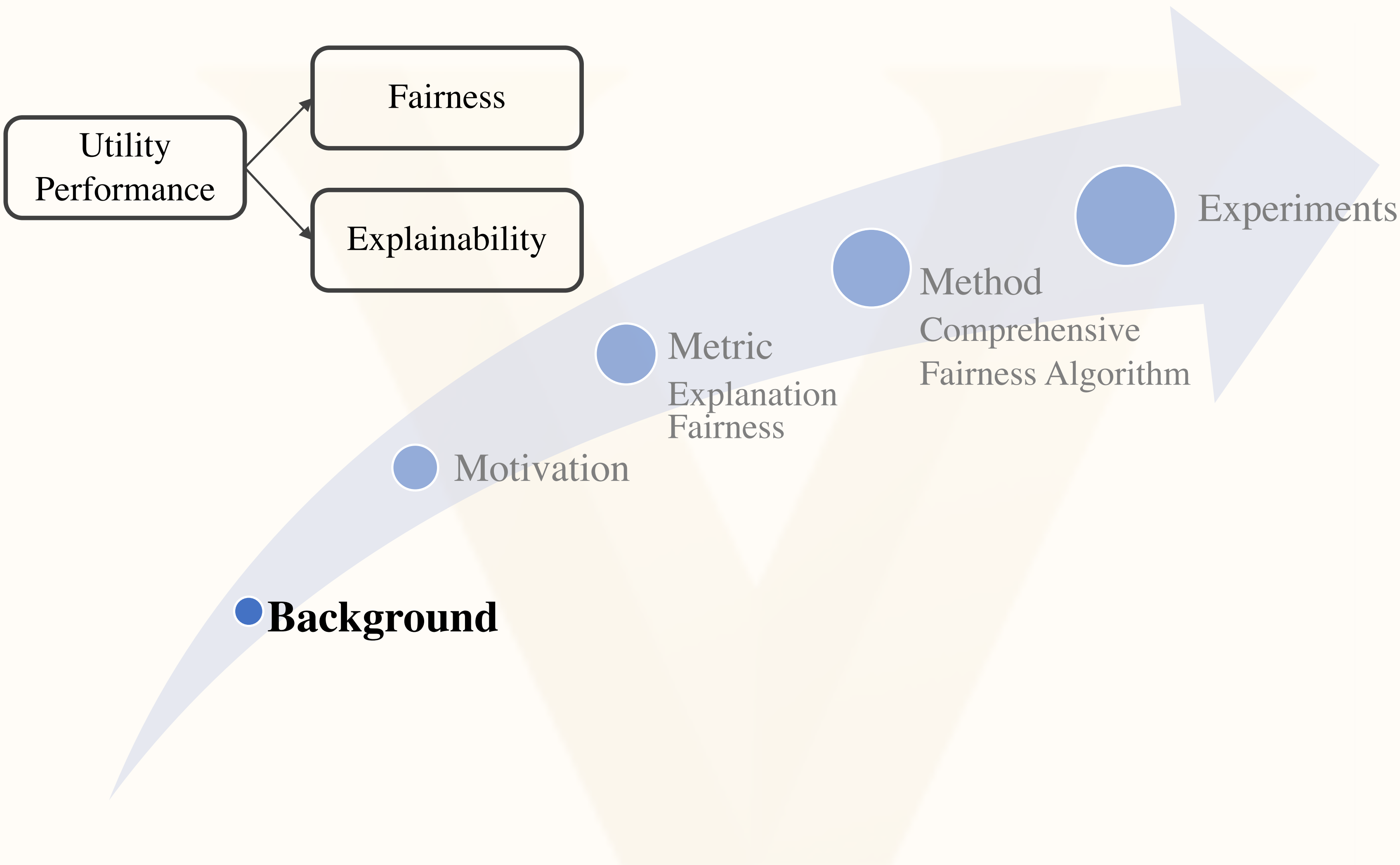
Education

Business perspective:
- Trust before deployment
- Find justification

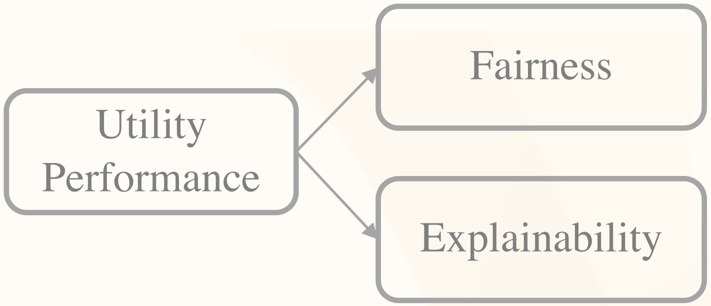
Model perspective:
- Debug model (mis)predictions
- Improve/verify ML models

Regulatory perspective:
- GDPR: Article 22 empowers individuals with the right to demand an explanation of how an automated system made a decision that affects them.

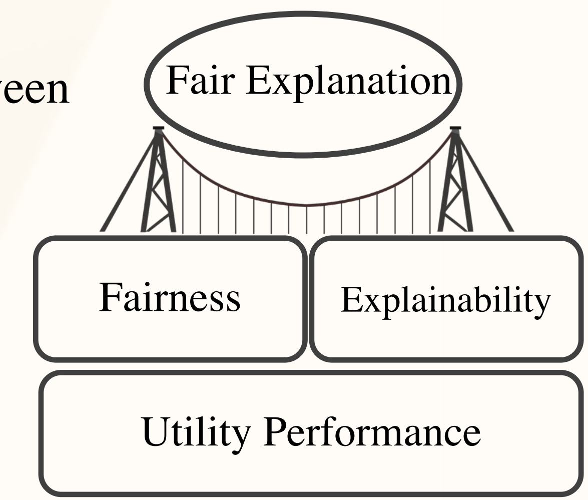
Content


Utility Performance
Fairness
Explainability
Background
Motivation
Metric
Explanation
Fairness
Method
Comprehensive
Fairness Algorithm
Experiments

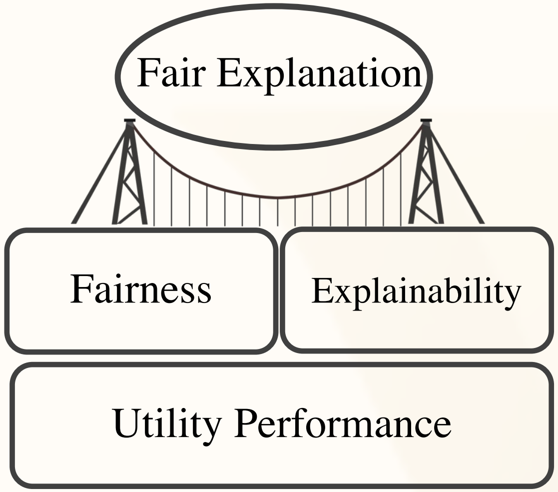
Content



- Background
- **Motivation**
  - Why we bridge the gap between fairness and explainability?
- Metric
  - Explanation Fairness
- Method
  - Comprehensive Fairness Algorithm
- Experiments



Existing Work



Explain the source of bias[1-4]

Whether the explanation is fair[5]

Result

Procedure

A general framework:
- Measure explanation fairness
- Improve explanation fairness

- Explanation quality

[1] Lundberg, S. M. "Explaining Quantitative Measures of Fairness." Fair & Responsible AI Workshop. 2020.
[2] Begley, Tom, et al. "Explainability for fair machine learning". arXiv. 2020.
[3] Chiappa, S. "Path-specific counterfactual fairness. " AAAI. 2019.
[4] Pan, Weishen, et al. "Explaining algorithmic fairness through fairness-aware causal path decomposition". KDD. 2021.
[5] Fu, Zuohui, et al. "Fairness-aware explainable recommendation over knowledge graphs." SIGIR. 2020.

Motivation: Fairness and Explainability

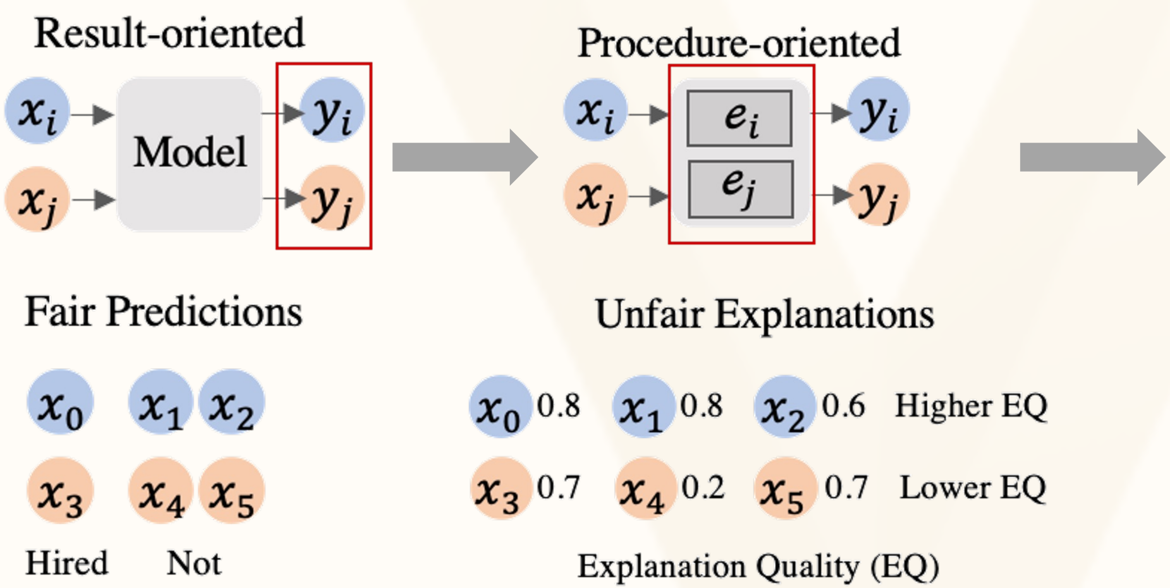
Groups divided by sensitive feature

$\hat{y}$: predictions
y: ground truth
s: sensitive features

Motivation:
most fairness metrics: result-oriented
hide the potential bias during the procedure

Statistical Parity:

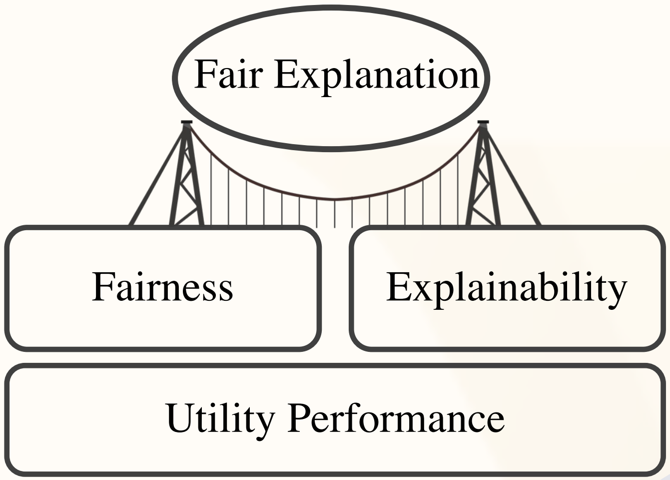
$$\Delta_{SP} = |P(\hat{y} = 1|s = 0) - P(\hat{y} = 1|s = 1)\,|$$



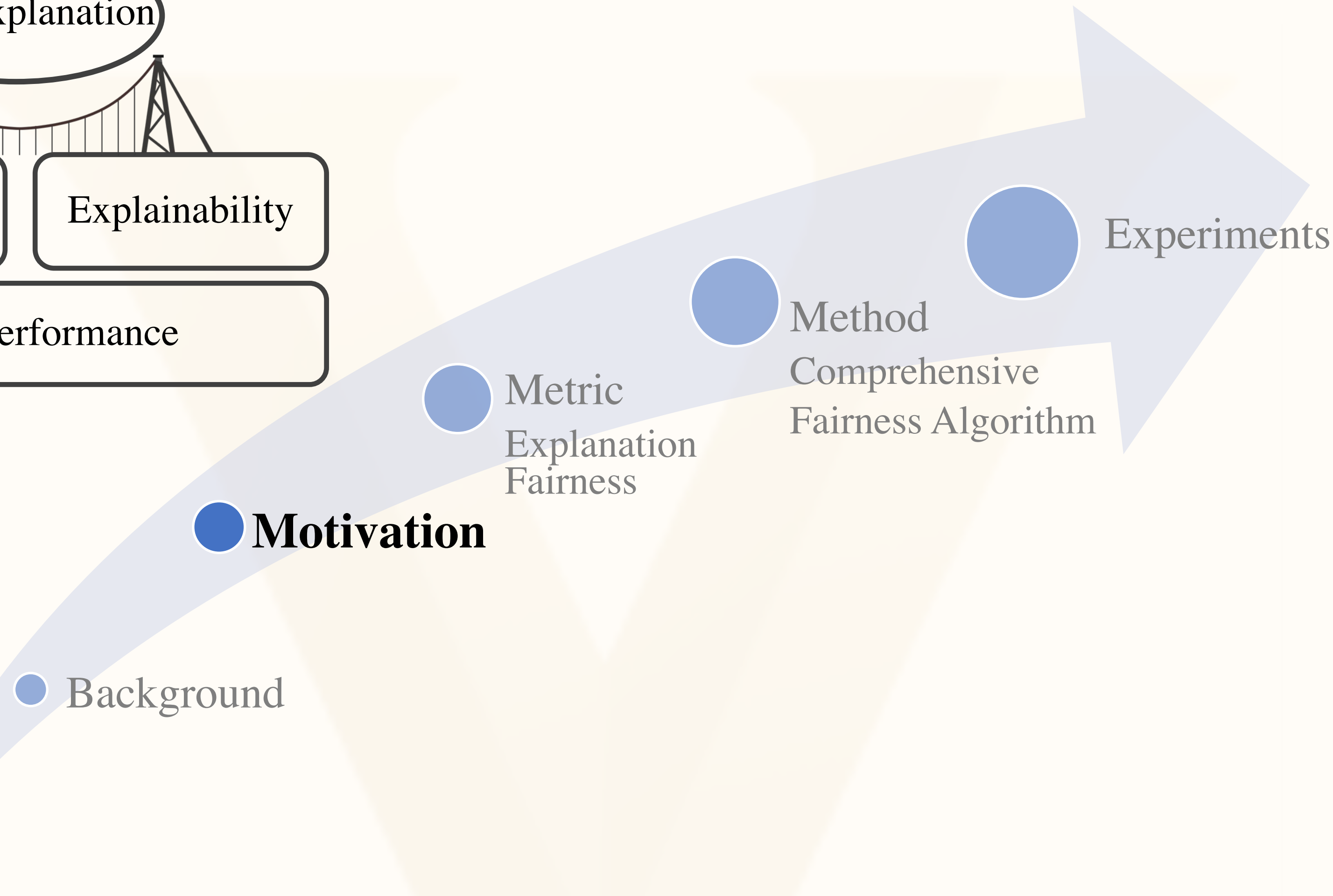
Unfairness:
Better explanation for one group than the other

Example: Job hiring
Well-explained vs
Ambiguous explanation

Content

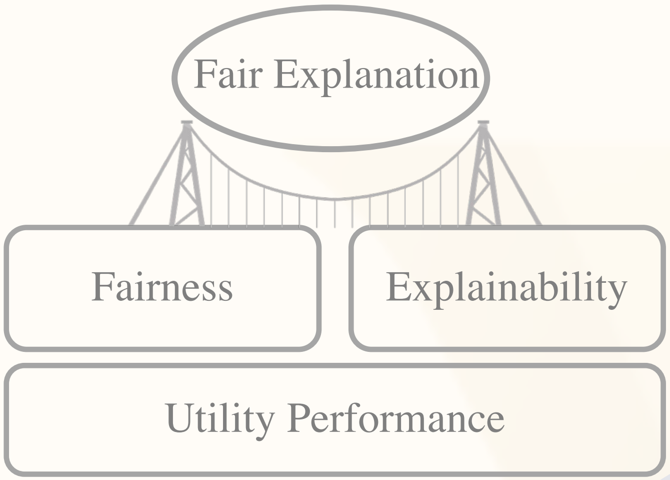


Fair Explanation
Fairness
Explainability
Utility Performance



Background
Motivation
Metric
Explanation
Fairness
Method
Comprehensive
Fairness Algorithm
Experiments

Content



Background

Motivation

Metric

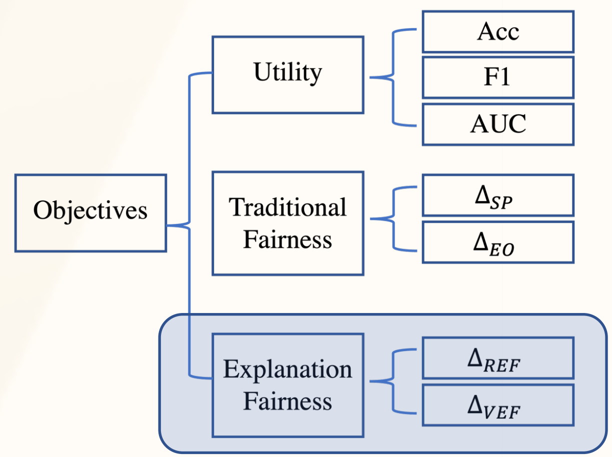
Explanation Fairness

Method

Comprehensive Fairness Algorithm

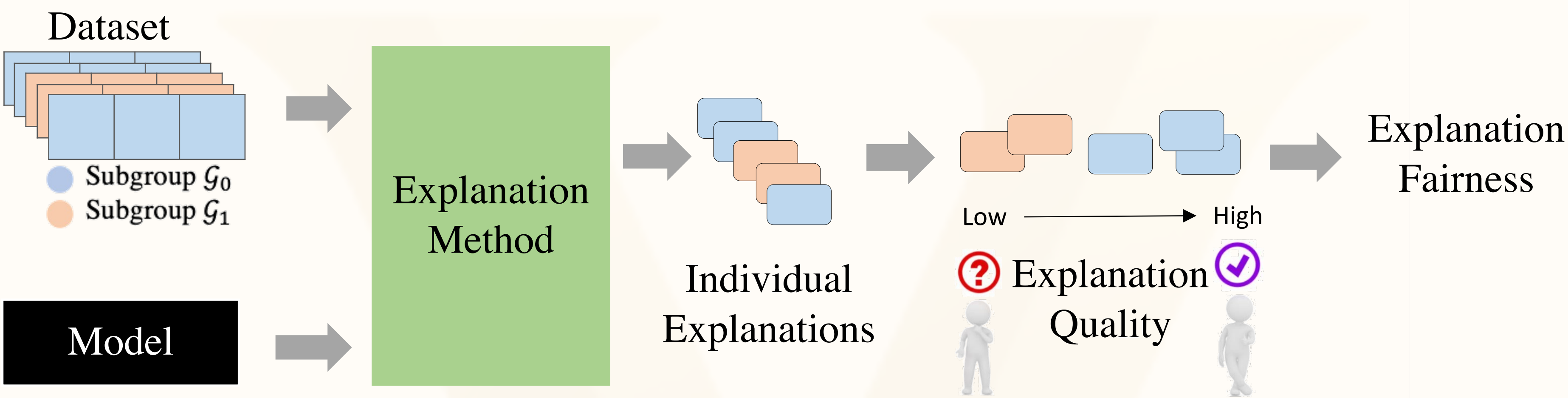
Experiments

- How to quantify explanation fairness?



Metric: High-level Idea of Explanation Fairness

Compare explanation quality from two subgroups

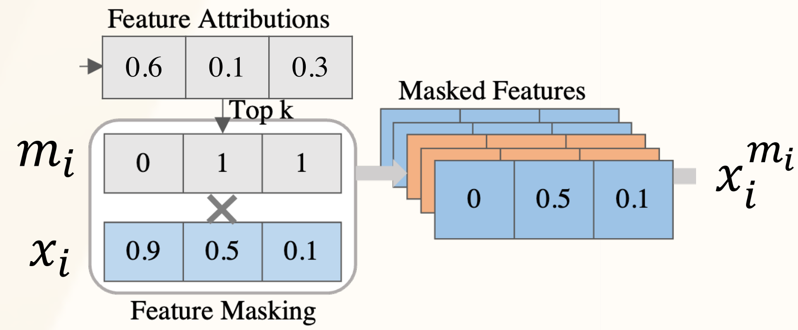


What makes a "good" explanation?

Fidelity:

$$\Delta_{F_i} = P(\hat{y}_i = y_i | x = x_i) - P\left(\hat{y}_i = y_i \middle| x = x_i^{m_i}\right)$$

How well does the explanation approximate the prediction of the black-box model?



Metric: Quantification of Explanation Fairness

Given explanation quality (EQ),
how to quantify explanation fairness?

(1) Ratio-based Fairness Δ_{REF}

$$\Delta_{SP} = |P(\boxed{\hat{y}=1} \mid s=0) - P(\boxed{\hat{y}=1} \mid s=1)|$$

Same opportunity of having positive prediction

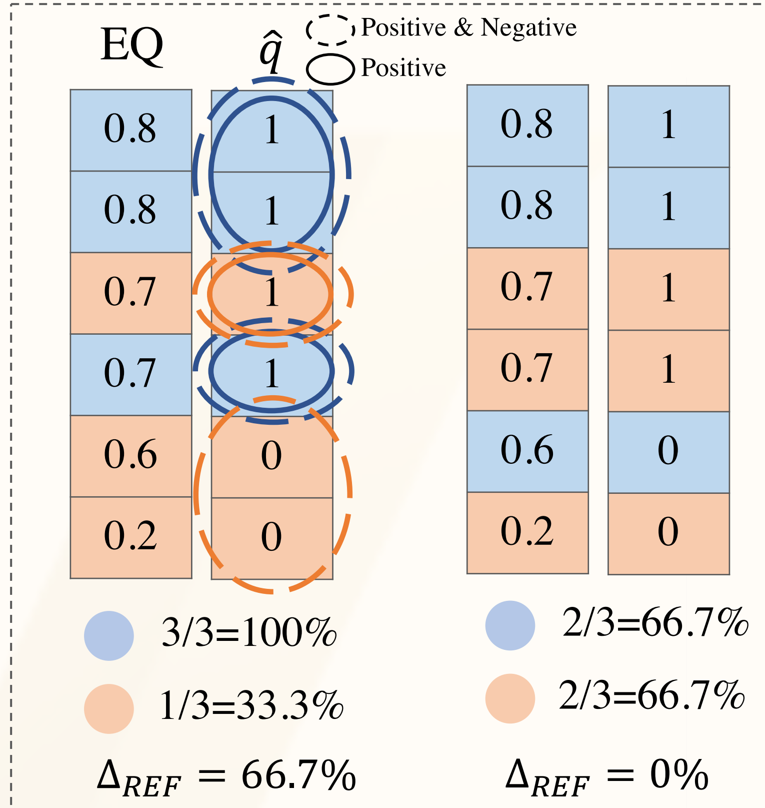
$$\Delta_{\mathrm{REF}} = |P(\boxed{\hat{q}=1}|s=0) - P(\boxed{\hat{q}=1}|s=1)|$$

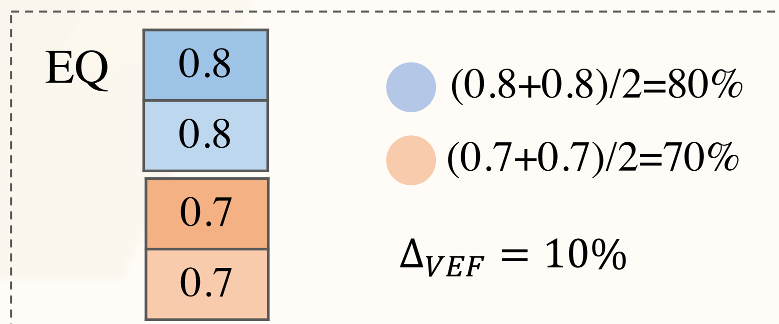
Same opportunity of having high-quality explanations

$\hat{y}$: *prediction* $\quad \hat{q}$: *explanation quality*

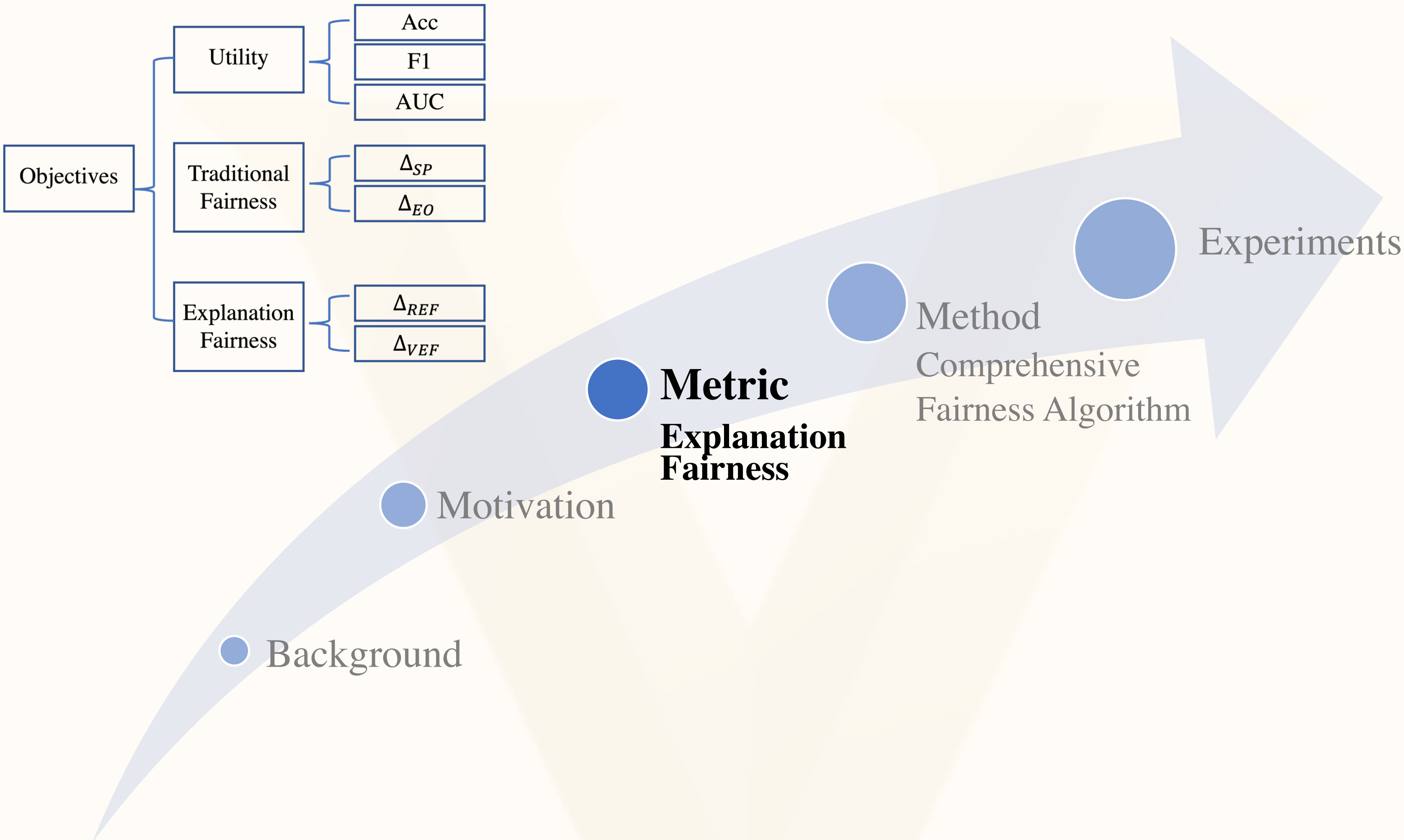
(2) Value-based Fairness Δ_{VEF}

$$\Delta_{\mathrm{VEF}} = \left| \frac{1}{|\mathcal{G}_0^K|} \sum_{i \in \mathcal{G}_0^K} \mathrm{EQ}_i - \frac{1}{|\mathcal{G}_1^K|} \sum_{i \in \mathcal{G}_1^K} \mathrm{EQ}_i \right|$$



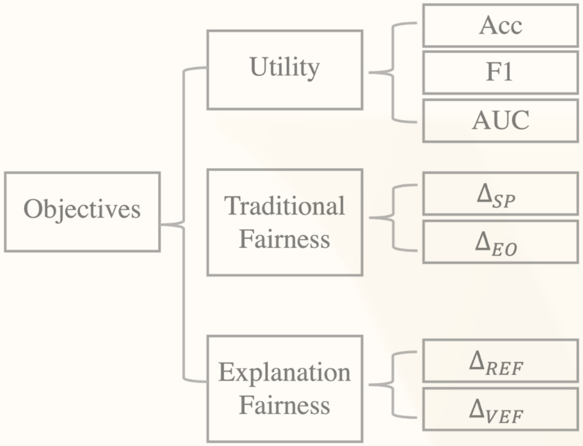


Content



Objectives
Utility
Acc
F1
AUC
Traditional Fairness
Δ_{SP}
Δ_{EO}
Explanation Fairness
Δ_{REF}
Δ_{VEF}
Background
Motivation
Metric
Explanation Fairness
Method
Comprehensive Fairness Algorithm
Experiments

Content



Background

Motivation

Metric
Explanation Fairness

Method
Comprehensive Fairness Algorithm

Experiments

- How to optimize multiple goals simultaneously?

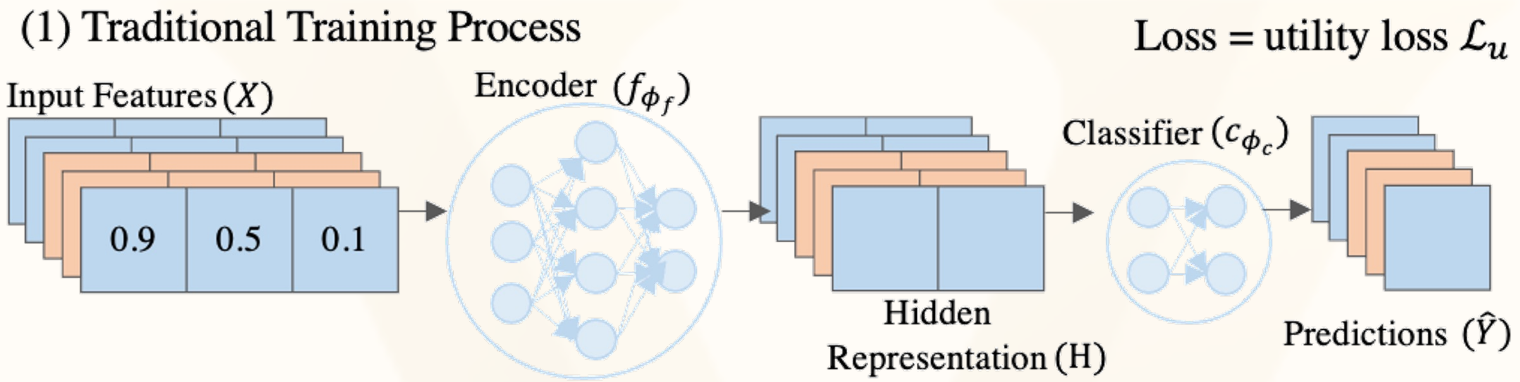
Traditional training
(for utility)

Distance loss based on original embedding
(for traditional fairness)

Distance loss based on masked embedding
(for explanation fairness)

Comprehensive Fairness Algorithm (CFA)

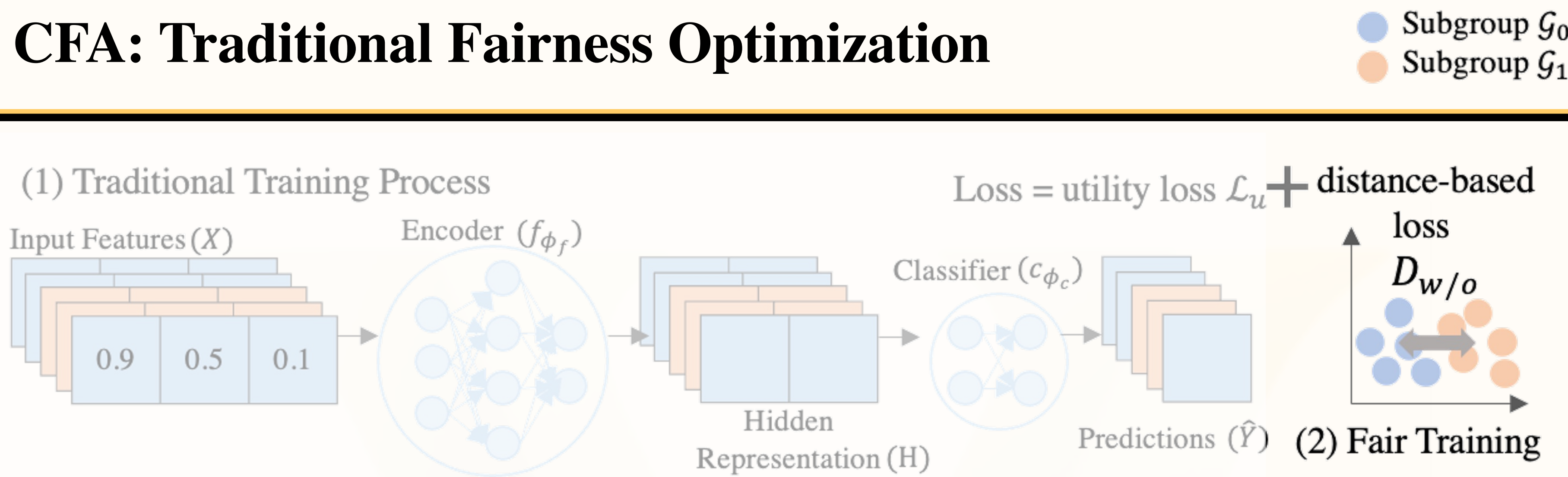
Subgroup $\mathcal{G}_0$
Subgroup $\mathcal{G}_1$



Utility loss: entropy loss (for binary classification)

$$\mathcal{L}_u = -\sum_{i=1}^{|\mathbf{Y}|}(y_i log(p) + (1 - y_i) log(1 - p))$$

CFA: Traditional Fairness Optimization



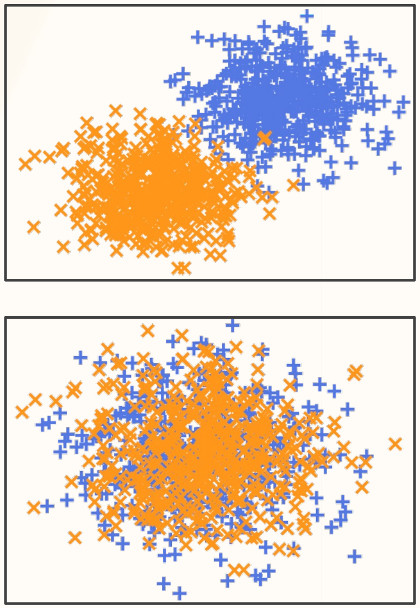
$$D_{w/o} = \mathcal{D}(\mathbf{H}_{\mathcal{G}_0}, \mathbf{H}_{\mathcal{G}_1})$$

w/o: without masking
based on the original feature

$$\Delta_{SP} = |P(\hat{y} = 1|s = 0) - P(\hat{y} = 1|s = 1)|$$
$$\Delta_{EO} = |P(\hat{y} = 1|y = 1, s = 0) - P(\hat{y} = 1|y = 1, s = 1)|$$

The predictions should be irrelevant to sensitive features

Requirements to the hidden representation
(1) Encode sufficient information for prediction
(2) Hide information related to sensitive features



[1]

[1] Dong, Yushun, et al. "Edits: Modeling and mitigating data bias for graph neural networks." WWW. 2022

CFA: Explanation Fairness Optimization

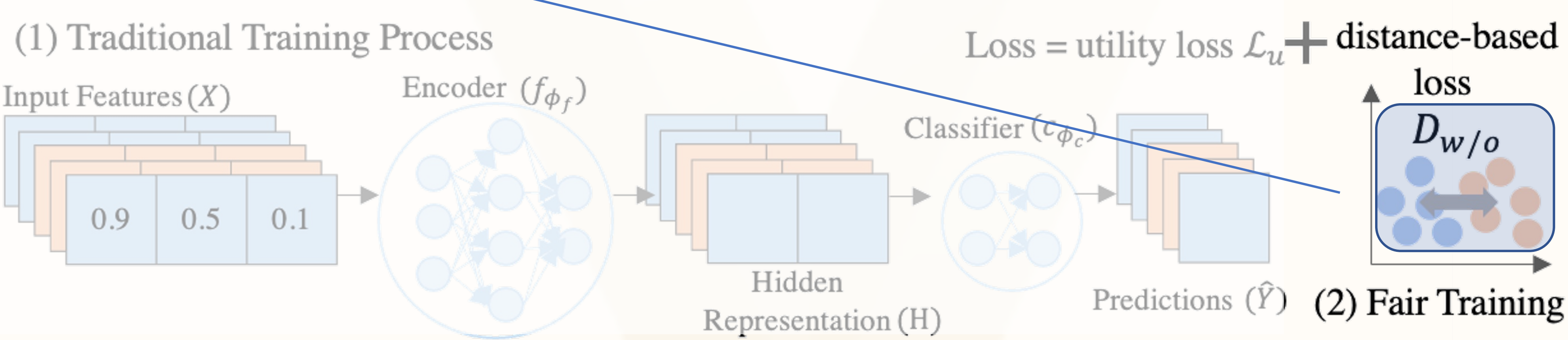
Subgroup $\mathcal{G}_0$
Subgroup $\mathcal{G}_1$

Explanation Fairness → Explanation Quality → Fidelity →

Fair explanation quality (measured by fidelity)

Fidelity: $\boxed{P(\hat{y}_i = y_i | x = x_i)} - \boxed{P(\hat{y}_i = y_i | x = x_i^{m_i})}$

Original feature — Masked feature



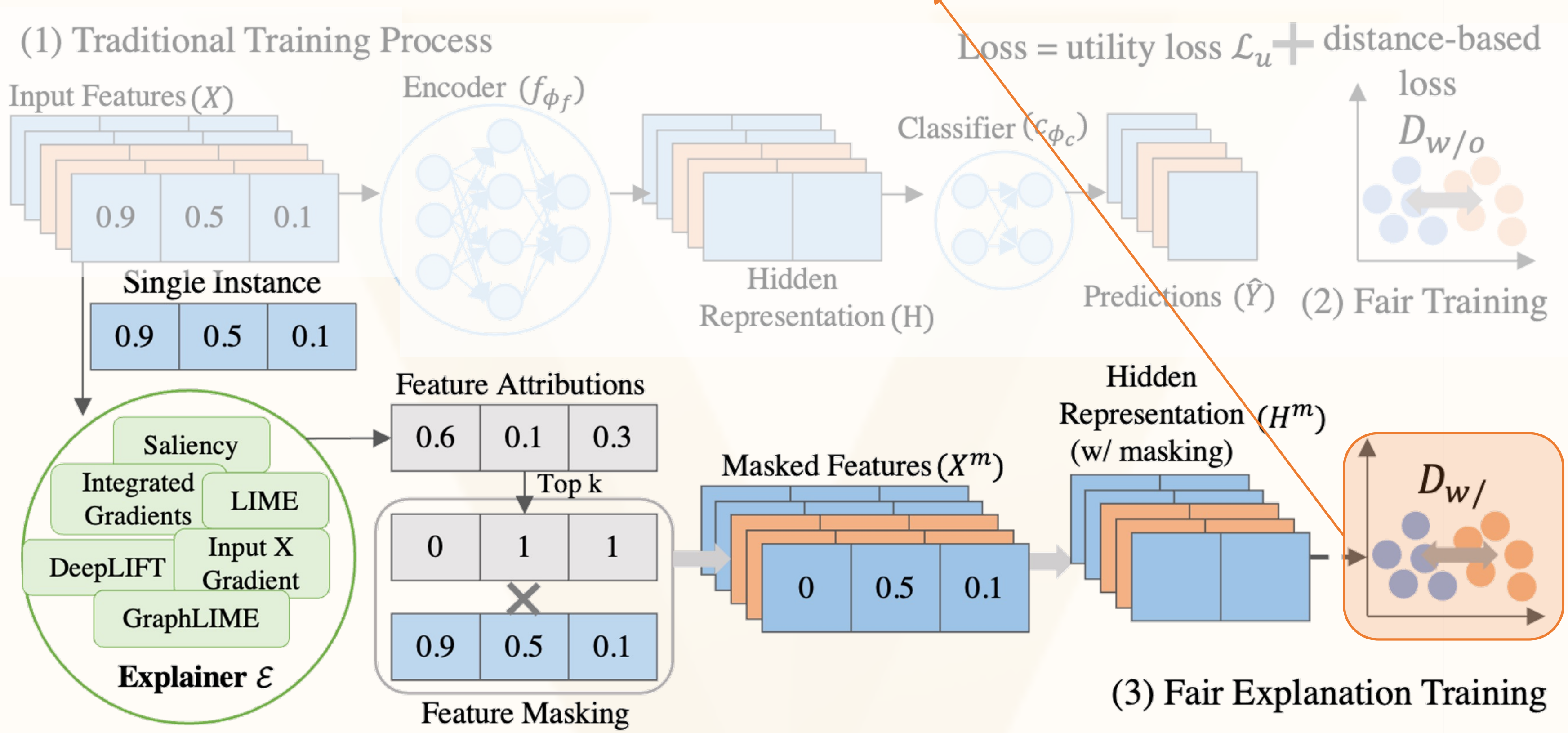
CFA: Explanation Fairness Optimization

Subgroup $\mathcal{G}_0$
Subgroup $\mathcal{G}_1$

Fidelity:

$$\boxed{P(\hat{y}_i = y_i | x = x_i)} - \boxed{P(\hat{y}_i = y_i | x = x_i^{m_i})}$$

Masked feature



w/: with masking
based on the masked feature

$$\mathcal{L}_u = -\sum_{i=1}^{|\mathbf{Y}|}(y_i log(p) + (1-y_i) log(1-p))$$

$$D_{w/o} = \mathcal{D}(\mathbf{H}_{\mathcal{G}_0}, \mathbf{H}_{\mathcal{G}_1})$$

$$D_{w/} = \mathcal{D}(\mathbf{H^m}_{\mathcal{G}_0}, \mathbf{H^m}_{\mathcal{G}_1})$$

$$\mathcal{L}_{exp} = D_{w/o} + D_{w/}$$

Traditional fairness

Explanation fairness

Content

Traditional training
(for utility)

Distance loss based on original embedding
(for traditional fairness)

Distance loss based on masked embedding
(for explanation fairness)

Background

Motivation

Metric
Explanation Fairness

Method
Comprehensive Fairness Algorithm

Experiments

Content

Traditional training
(for utility)

Distance loss based on original embedding
(for traditional fairness)

Distance loss based on masked embedding
(for explanation fairness)

Background

Motivation

Metric
Explanation Fairness

Method
Comprehensive Fairness Algorithm

Experiments

RQ1: Bias Mitigation
How well can CFA mitigate the bias?

RQ2: Tradeoff
How well can CFA balance different categories of objectives?

Experimental Setting

Dataset

Dataset	German	Recidivism	Math	Por
# Nodes	1,000	18,876	649	649
# Features	27	18	33	33
Sens.	Gender	Race	Gender	Gender
Label	Credit Risk	Recidivism	Grade	Grade

Evaluation metrics

- Utility(↑): accuracy, F1, AUC
- Traditional fairness (result-oriented, ↓): Δ_{SP} and Δ_{EO}
- Explanation fairness (procedure-oriented, ↓): Δ_{VEF} and Δ_{REF}
- Overall score: $\frac{\text{AUC+F1+ACC}}{3} - \frac{\Delta_{SP}+\Delta_{EO}}{2} - \frac{\Delta_{VEF}+\Delta_{REF}}{2}$ (model selection)

Baselines

(1) Reweight[1] : [reweighing-based] reweight the training loss

(2) Reduction[2] : [constraint-based] optimization under fairness constraints

$$\Delta_{SP} = |P(\hat{y}=1|s=0) - P(\hat{y}=1|s=1)|$$

$$\Delta_{EO} = |P(\hat{y}=1|y=1,s=0) - P(\hat{y}=1|y=1,s=1)|$$

$$\Delta_{\text{REF}} = |P(\hat{q}=1|s=0) - P(\hat{q}=1|s=1)|$$

$$\Delta_{\text{VEF}} = \left|\frac{1}{|\mathcal{G}_0^K|}\sum_{i\in\mathcal{G}_0^K}\text{EQ}_i - \frac{1}{|\mathcal{G}_1^K|}\sum_{i\in\mathcal{G}_1^K}\text{EQ}_i\right|$$

[1] Jiang, Heinrich, et al. "Identifying and correcting label bias in machine learning." AISTATS, 2020.
[2] Agarwal, Alekh, et al. "A reductions approach to fair classification." ICML, 2018.

RQ1: Bias Mitigation

Takes up largest proportion of bold/underline

Dataset	Metric	MLP	Reduction	Reweight	CFA
Recidivism	AUC↑	86.12 ± 1.91	81.17 ± 0.00	**89.24 ± 0.00**	$\underline{89.02 \pm 0.86}$
	F1↑	76.54 ± 2.52	$\underline{76.69 \pm 0.00}$	72.99 ± 0.00	**81.28 ± 1.35**
	Acc↑	83.48 ± 1.53	$\underline{84.66 \pm 0.00}$	83.70 ± 0.00	**87.17 ± 0.84**
	$\Delta_{SP}\downarrow$	6.07 ± 2.18	$\underline{2.04 \pm 0.00}$	4.27 ± 0.00	**1.16 ± 0.49**
	$\Delta_{EO}\downarrow$	$\underline{3.19 \pm 0.73}$	4.66 ± 0.00	3.37 ± 0.00	**1.14 ± 0.39**
	$\Delta_{REF}\downarrow$	4.45 ± 2.96	**0.53 ± 0.00**	$\underline{1.34 \pm 0.91}$	1.98 ± 1.23
	$\Delta_{VEF}\downarrow$	$\underline{2.1 \pm 1.38}$	**2.06 ± 0.00**	3.22 ± 0.00	2.70 ± 0.78
	Score↑	74.15 ± 2.03	$\underline{76.19 \pm 0.00}$	$\underline{75.88 \pm 0.00}$	**82.33 ± 0.62**
Por	AUC↑	$\underline{90.86 \pm 0.35}$	67.64 ± 0.00	89.07 ± 0.00	**91.30 ± 0.55**
	F1↑	$\underline{58.41 \pm 4.10}$	51.43 ± 0.00	51.43 ± 0.00	**60.55 ± 4.73**
	Acc↑	$\underline{89.57 \pm 0.78}$	$\underline{89.57 \pm 0.00}$	$\underline{89.57 \pm 0.00}$	**89.82 ± 1.00**
	$\Delta_{SP}\downarrow$	2.08 ± 0.75	$\underline{1.93 \pm 0.00}$	$\underline{1.93 \pm 0.00}$	**1.00 ± 0.72**
	$\Delta_{EO}\downarrow$	32.35 ± 7.07	**20.59 ± 0.00**	**20.59 ± 0.00**	$\underline{27.65 \pm 5.44}$
	$\Delta_{REF}\downarrow$	8.68 ± 3.18	**1.37 ± 0.00**	8.68 ± 0.00	$\underline{4.66 \pm 3.76}$
	$\Delta_{VEF}\downarrow$	$\underline{4.44 \pm 2.22}$	**0.00 ± 0.00**	7.69 ± 0.00	4.70 ± 3.67
	Score↑	55.83 ± 3.97	$\underline{57.60 \pm 0.00}$	57.25 ± 0.00	**61.55 ± 3.26**

Bold text: best performance
Underline text: second best performance

RQ1: Bias Mitigation

Dataset	Metric	MLP	Reduction	Reweight	CFA
Recidivism	AUC↑	86.12 ± 1.91	81.17 ± 0.00	**89.24 ± 0.00**	89.02 ± 0.86
	F1↑	76.54 ± 2.52	76.69 ± 0.00	72.99 ± 0.00	**81.28 ± 1.35**
	Acc↑	83.48 ± 1.53	84.66 ± 0.00	83.70 ± 0.00	**87.17 ± 0.84**
	$\Delta_{SP}\downarrow$	6.07 ± 2.18	2.04 ± 0.00	4.27 ± 0.00	**1.16 ± 0.49**
	$\Delta_{EO}\downarrow$	3.19 ± 0.73	4.66 ± 0.00	3.37 ± 0.00	**1.14 ± 0.39**
	$\Delta_{REF}\downarrow$	4.45 ± 2.96	**0.53 ± 0.00**	1.34 ± 0.91	1.98 ± 1.23
	$\Delta_{VEF}\downarrow$	2.1 ± 1.38	**2.06 ± 0.00**	3.22 ± 0.00	2.70 ± 0.78
	Score ↑	74.15 ± 2.03	76.19 ± 0.00	75.88 ± 0.00	**82.33 ± 0.62**
Por	AUC↑	90.86 ± 0.35	67.64 ± 0.00	89.07 ± 0.00	**91.30 ± 0.55**
	F1↑	58.41 ± 4.10	51.43 ± 0.00	51.43 ± 0.00	**60.55 ± 4.73**
	Acc↑	89.57 ± 0.78	89.57 ± 0.00	89.57 ± 0.00	**89.82 ± 1.00**
	$\Delta_{SP}\downarrow$	2.08 ± 0.75	1.93 ± 0.00	1.93 ± 0.00	**1.00 ± 0.72**
	$\Delta_{EO}\downarrow$	32.35 ± 7.07	**20.59 ± 0.00**	**20.59 ± 0.00**	27.65 ± 5.44
	$\Delta_{REF}\downarrow$	8.68 ± 3.18	**1.37 ± 0.00**	8.68 ± 0.00	4.66 ± 3.76
	$\Delta_{VEF}\downarrow$	4.44 ± 2.22	**0.00 ± 0.00**	7.69 ± 0.00	4.70 ± 3.67
	Score ↑	55.83 ± 3.97	57.60 ± 0.00	57.25 ± 0.00	**61.55 ± 3.26**

Utility Performance

Traditional Fairness

Explanation Fairness

Comparable or better than baselines

RQ1: Bias Mitigation

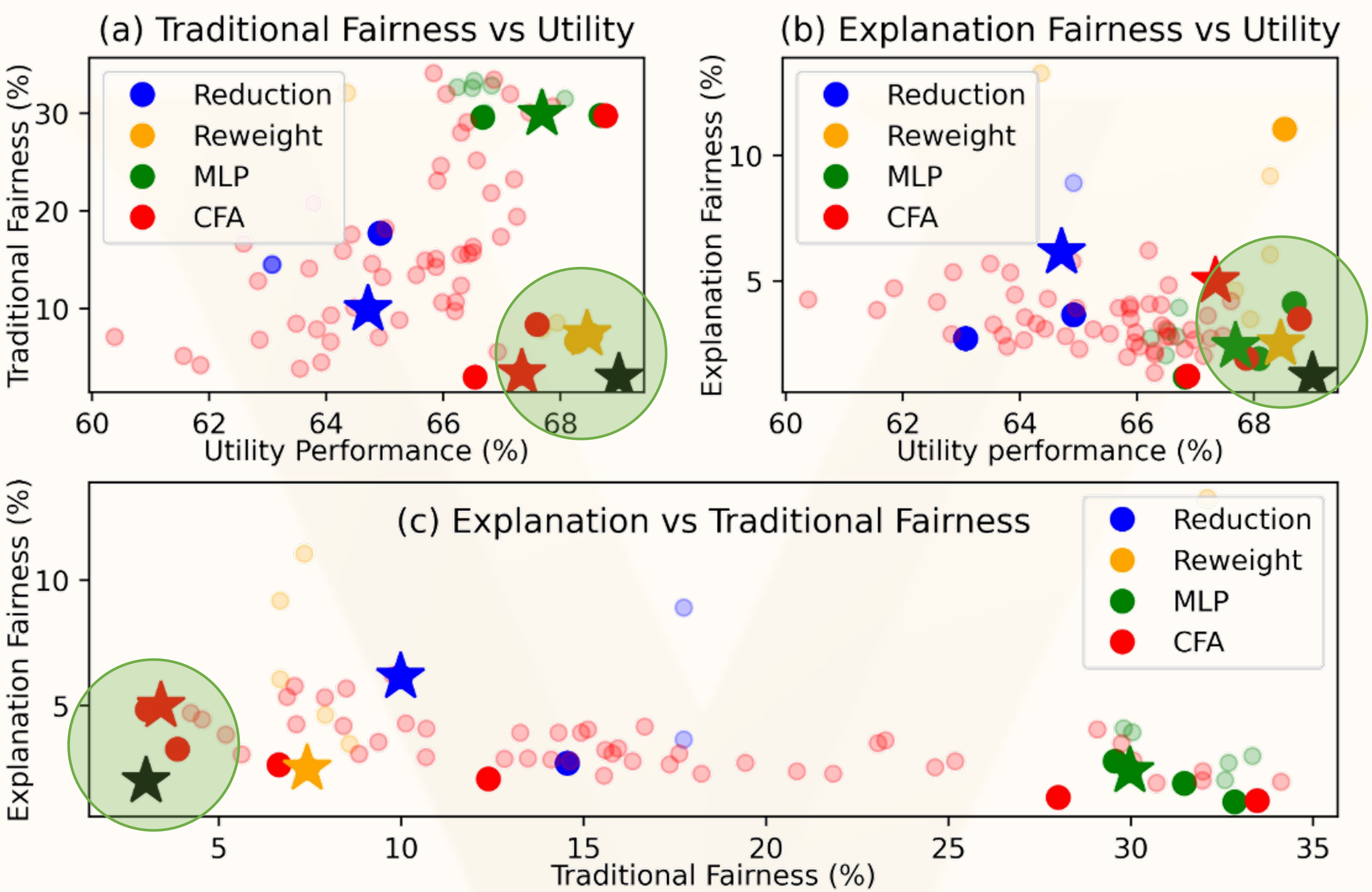
Dataset	Metric	MLP	Reduction	Reweight	CFA
Recidivism	AUC↑	86.12 ± 1.91	81.17 ± 0.00	$\mathbf{89.24 \pm 0.00}$	$\underline{89.02 \pm 0.86}$
	F1↑	76.54 ± 2.52	$\underline{76.69 \pm 0.00}$	72.99 ± 0.00	$\mathbf{81.28 \pm 1.35}$
	Acc↑	83.48 ± 1.53	$\underline{84.66 \pm 0.00}$	83.70 ± 0.00	$\mathbf{87.17 \pm 0.84}$
	$\Delta_{\text{SP}} \downarrow$	6.07 ± 2.18	$\underline{2.04 \pm 0.00}$	4.27 ± 0.00	$\mathbf{1.16 \pm 0.49}$
	$\Delta_{\text{EO}} \downarrow$	$\underline{3.19 \pm 0.73}$	4.66 ± 0.00	3.37 ± 0.00	$\mathbf{1.14 \pm 0.39}$
	$\Delta_{\text{REF}} \downarrow$	4.45 ± 2.96	$\mathbf{0.53 \pm 0.00}$	$\underline{1.34 \pm 0.91}$	1.98 ± 1.23
	$\Delta_{\text{VEF}} \downarrow$	$\underline{2.1 \pm 1.38}$	$\mathbf{2.06 \pm 0.00}$	3.22 ± 0.00	2.70 ± 0.78
	Score ↑	74.15 ± 2.03	$\underline{76.19 \pm 0.00}$	$\underline{75.88 \pm 0.00}$	$\mathbf{82.33 \pm 0.62}$
Por	AUC↑	$\underline{90.86 \pm 0.35}$	67.64 ± 0.00	89.07 ± 0.00	$\mathbf{91.30 \pm 0.55}$
	F1↑	$\underline{58.41 \pm 4.10}$	51.43 ± 0.00	51.43 ± 0.00	$\mathbf{60.55 \pm 4.73}$
	Acc↑	$\underline{89.57 \pm 0.78}$	$\underline{89.57 \pm 0.00}$	$\underline{89.57 \pm 0.00}$	$\mathbf{89.82 \pm 1.00}$
	$\Delta_{\text{SP}} \downarrow$	2.08 ± 0.75	$\underline{1.93 \pm 0.00}$	$\underline{1.93 \pm 0.00}$	$\mathbf{1.00 \pm 0.72}$
	$\Delta_{\text{EO}} \downarrow$	32.35 ± 7.07	$\mathbf{20.59 \pm 0.00}$	$\mathbf{20.59 \pm 0.00}$	$\underline{27.65 \pm 5.44}$
	$\Delta_{\text{REF}} \downarrow$	8.68 ± 3.18	$\mathbf{1.37 \pm 0.00}$	8.68 ± 0.00	$\underline{4.66 \pm 3.76}$
	$\Delta_{\text{VEF}} \downarrow$	$\underline{4.44 \pm 2.22}$	$\mathbf{0.00 \pm 0.00}$	7.69 ± 0.00	4.70 ± 3.67
	Score ↑	55.83 ± 3.97	$\underline{57.60 \pm 0.00}$	57.25 ± 0.00	$\mathbf{61.55 \pm 3.26}$

Overall Score

The highest for all datasets

Overall score: $\frac{\text{AUC}+\text{F1}+\text{ACC}}{3} - \frac{\Delta_{SP}+\Delta_{EO}}{2} - \frac{\Delta_{VEF}+\Delta_{REF}}{2}$ (model selection)

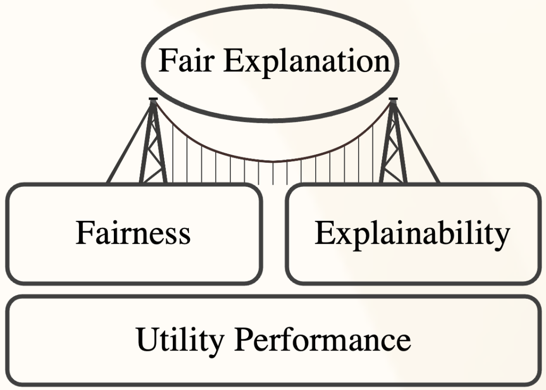
RQ2: Tradeoff

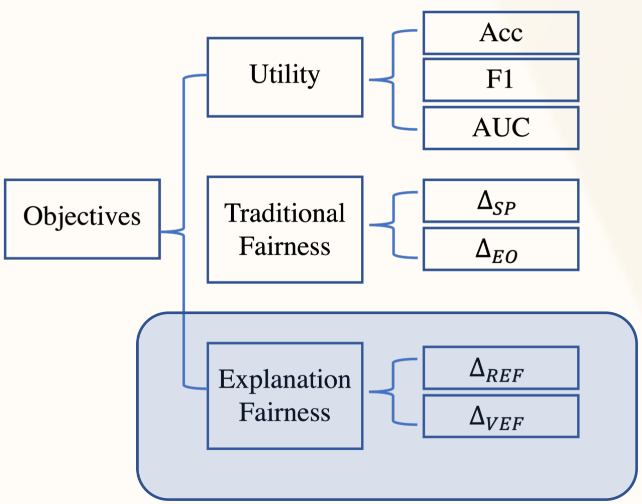


Circle: result of one hyper-parameter (Pareto frontier)
Star: best hyper-parameter setting based on overall score
Black star: the ideal direction of optimal solution

Summary

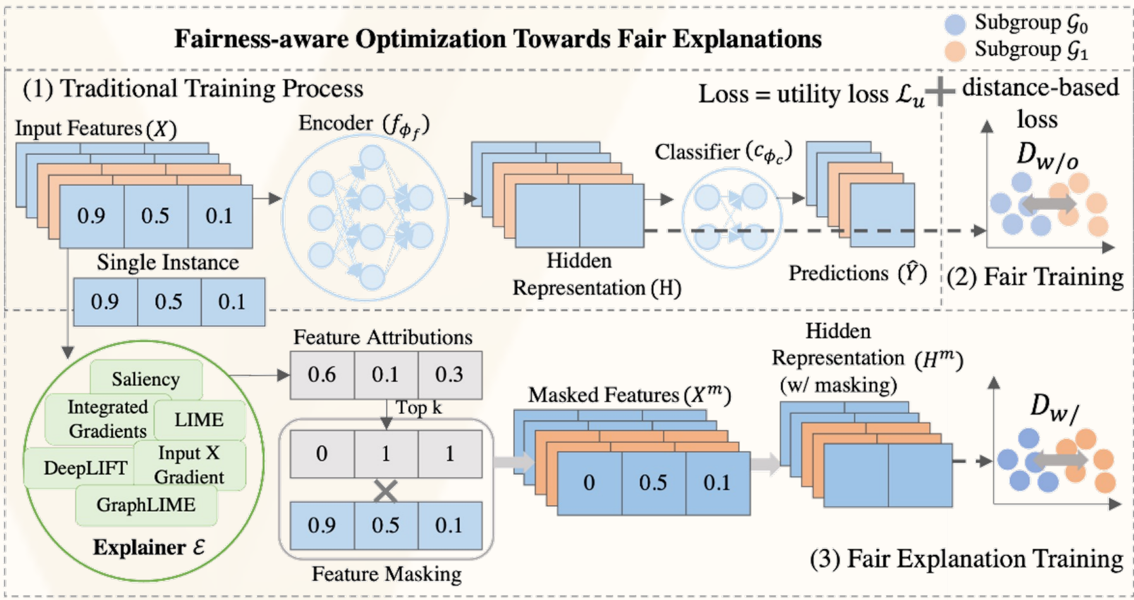
Novel Fairness Perspective





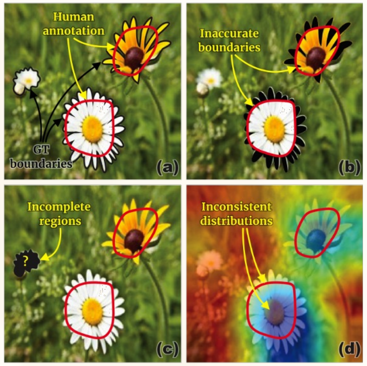
Comprehensive Fairness Algorithm

Traditional training (for utility) → **Distance loss based on original embedding** (for traditional fairness) → **Distance loss based on masked embedding** (for explanation fairness)

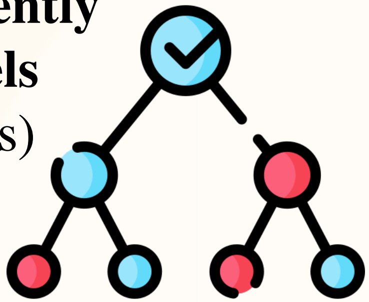


Future Directions

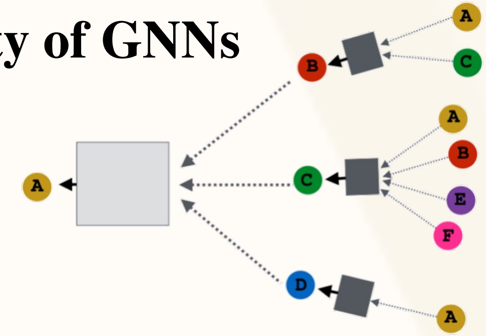
Extending CFA towards fair model explanations in other data types (e.g., images)



Defining novel fair explanation metrics for inherently explainable models (e.g., decision trees)



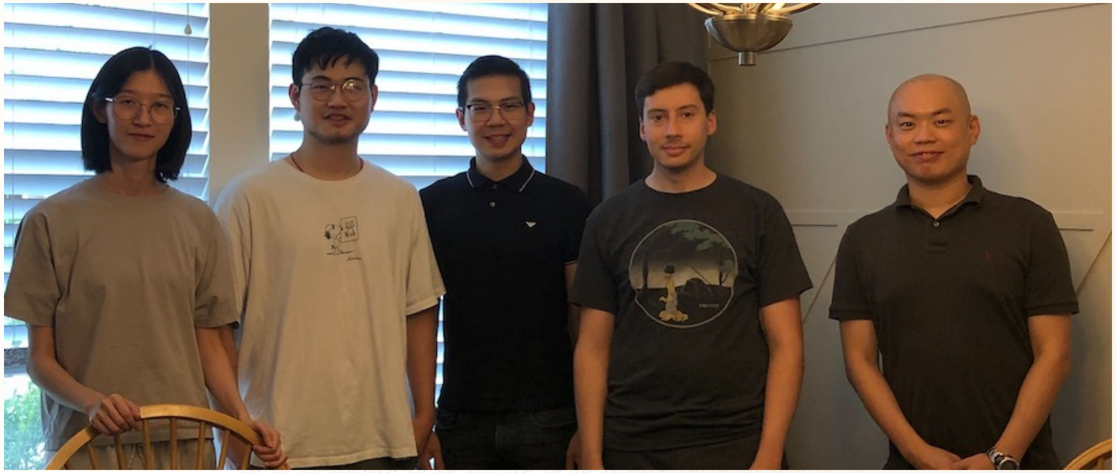
Improved Fairness and Explainability of GNNs



https://yuyingzhao.github.io/

Please see my website for other work

Acknowledgement



Network and
Data Science
Lab

AAAI-23 Student Scholar,
Diversity and Inclusion Scholar,
& Volunteer Program