Fairness and Explainability: Bridging the Gap Towards Fair Model Explanations

Welcome to my website
for code and more details

Yuying Zhao

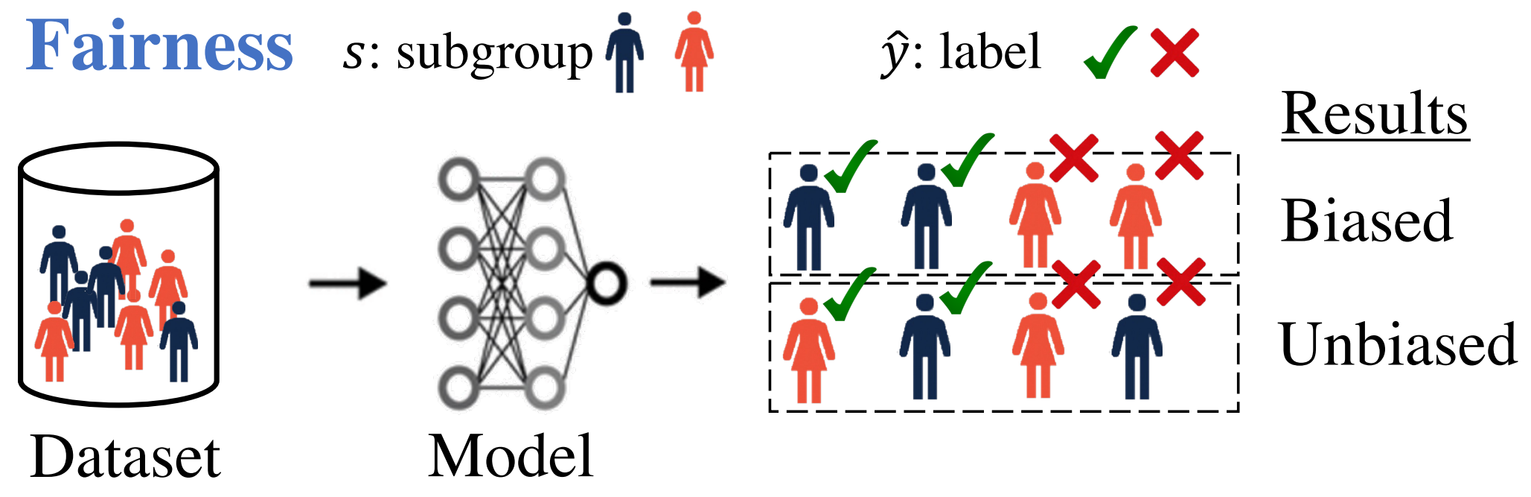
Yu Wang

Tyler Derr

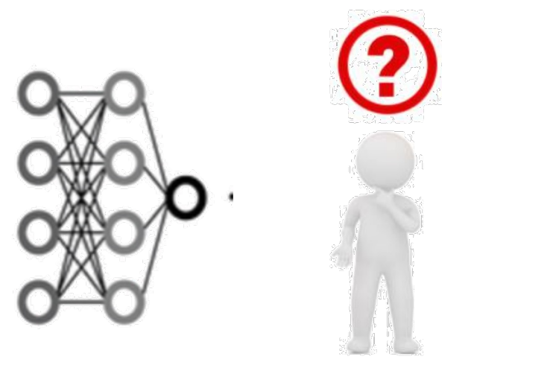
Network and Data Science Lab

VANDERBILT UNIVERSITY

Fairness and Explainability



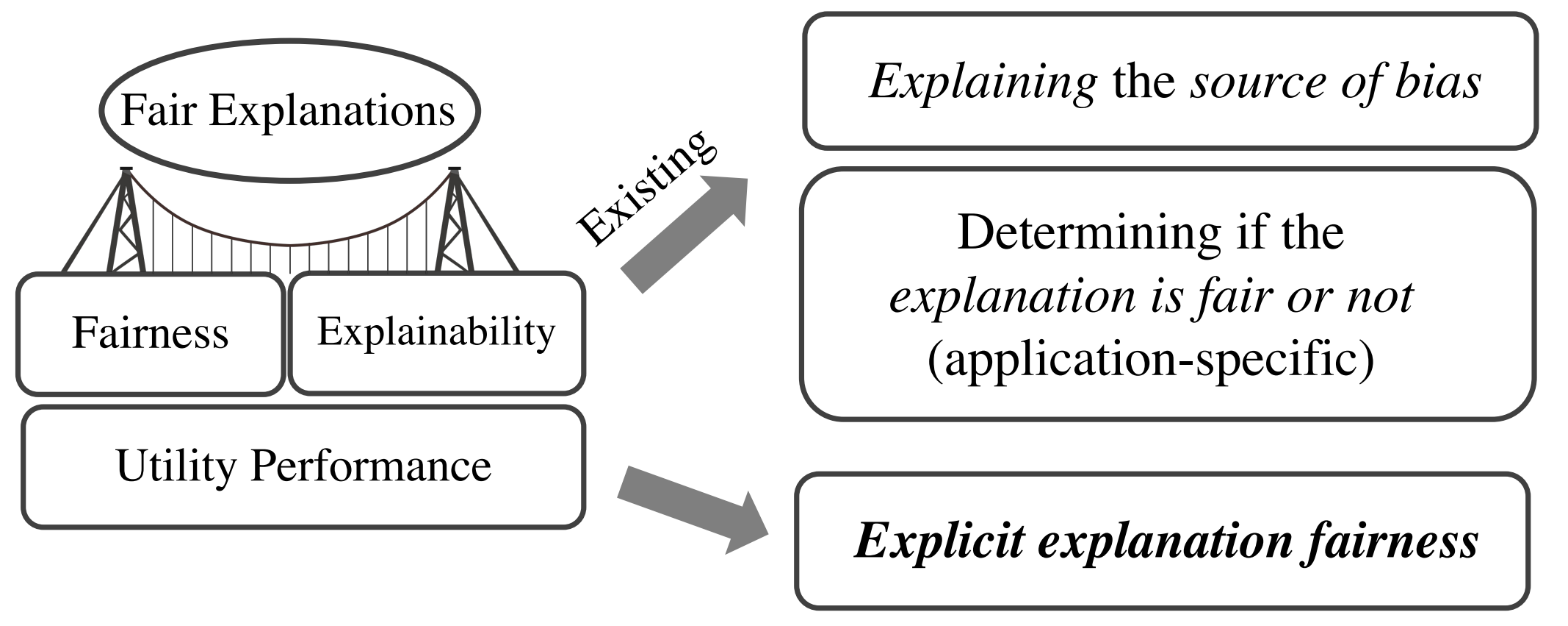
Explainability



Why should we trust the model?

Why did the model make a certain decision?

Fairness and Explainability

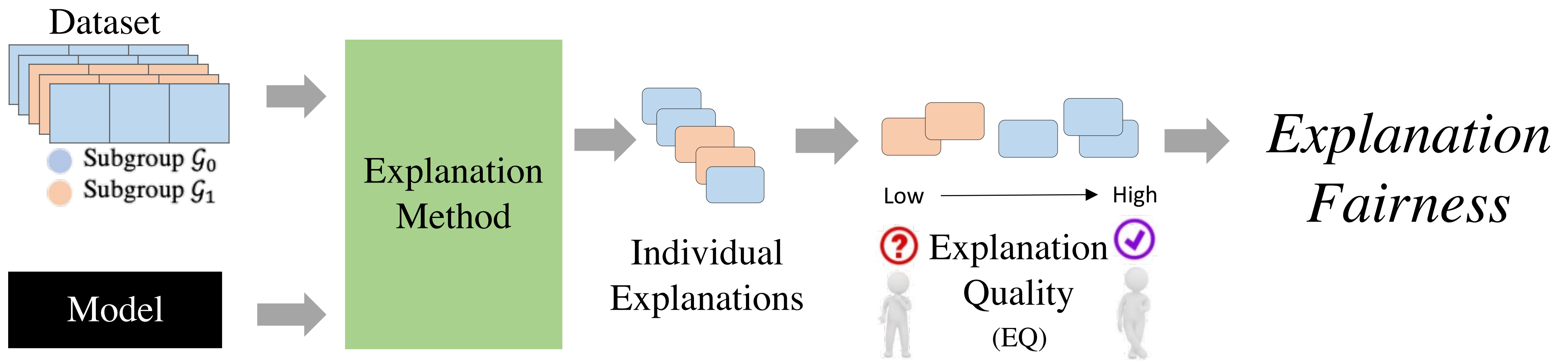


Explaining the *source of bias*

Determining if the *explanation is fair or not* (application-specific)

Explicit explanation fairness

Novel Procedure-Oriented Fairness Perspective



(1) Ratio-based Fairness Δ_{REF}

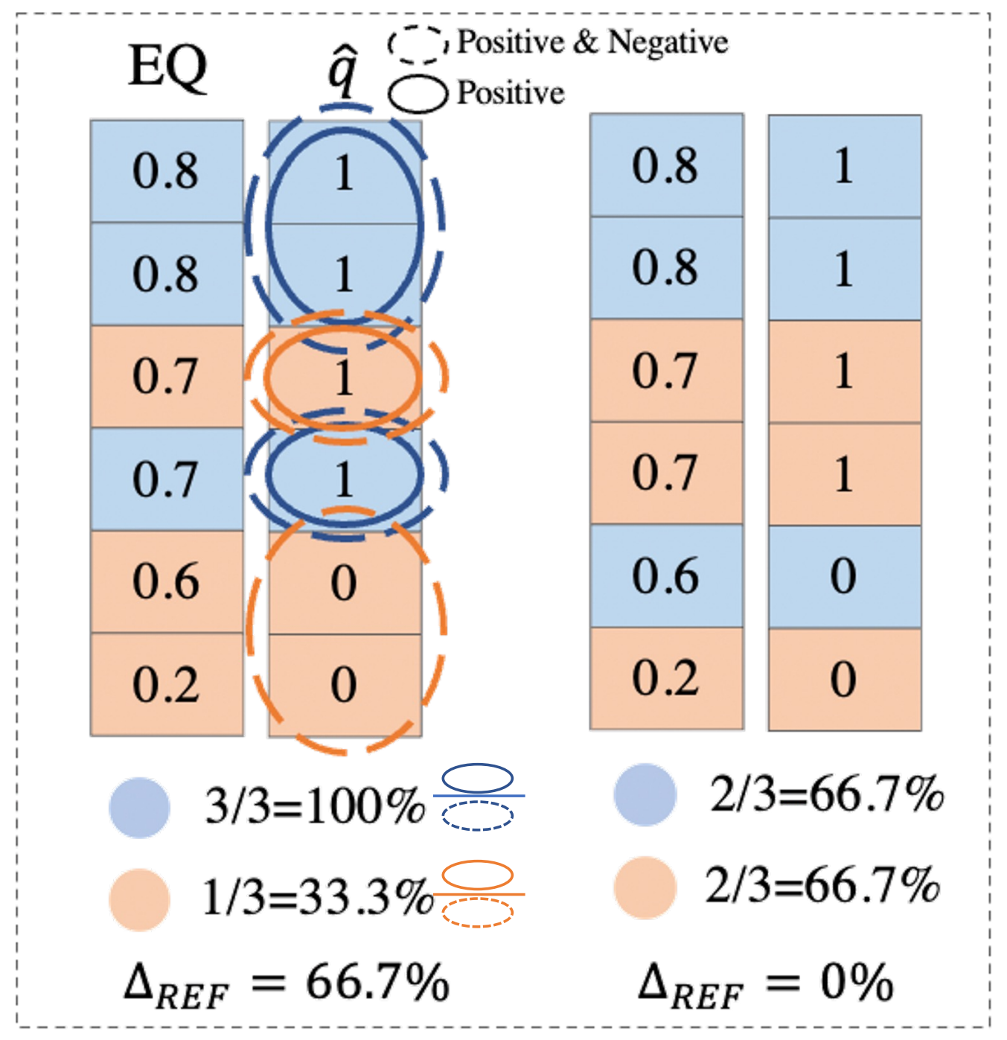
$$\Delta_{SP} = |P(\hat{y}=1|s=0) - P(\hat{y}=1|s=1)|$$

Same opportunity of having positive prediction

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

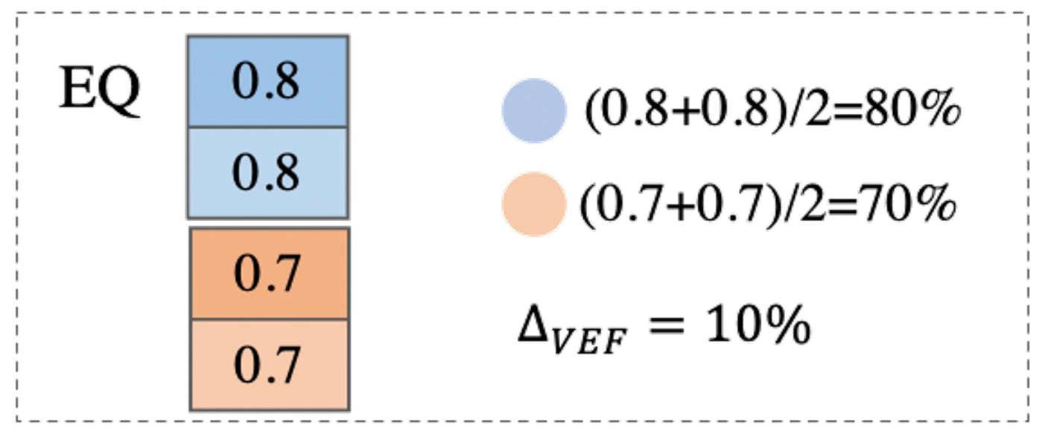
Same opportunity of having high-quality explanations

$\hat{y}$: prediction $\hat{q}$: explanation quality s: sensitive labels



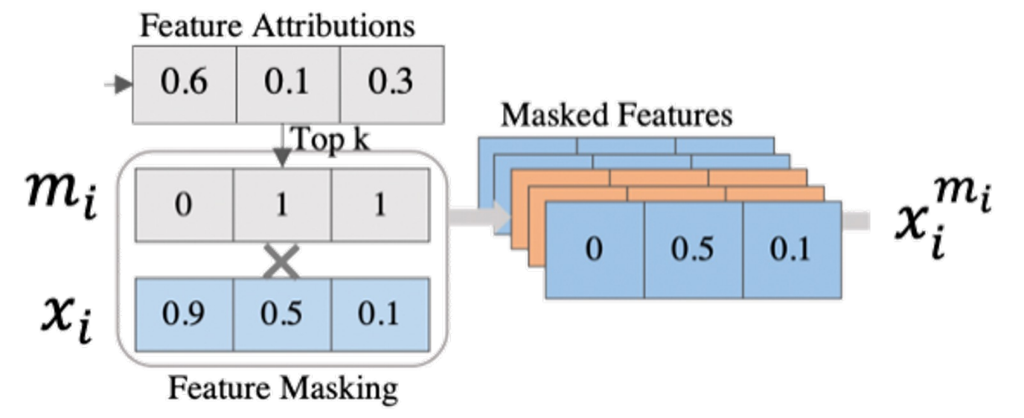
(2) Value-based Fairness Δ_{VEF}

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

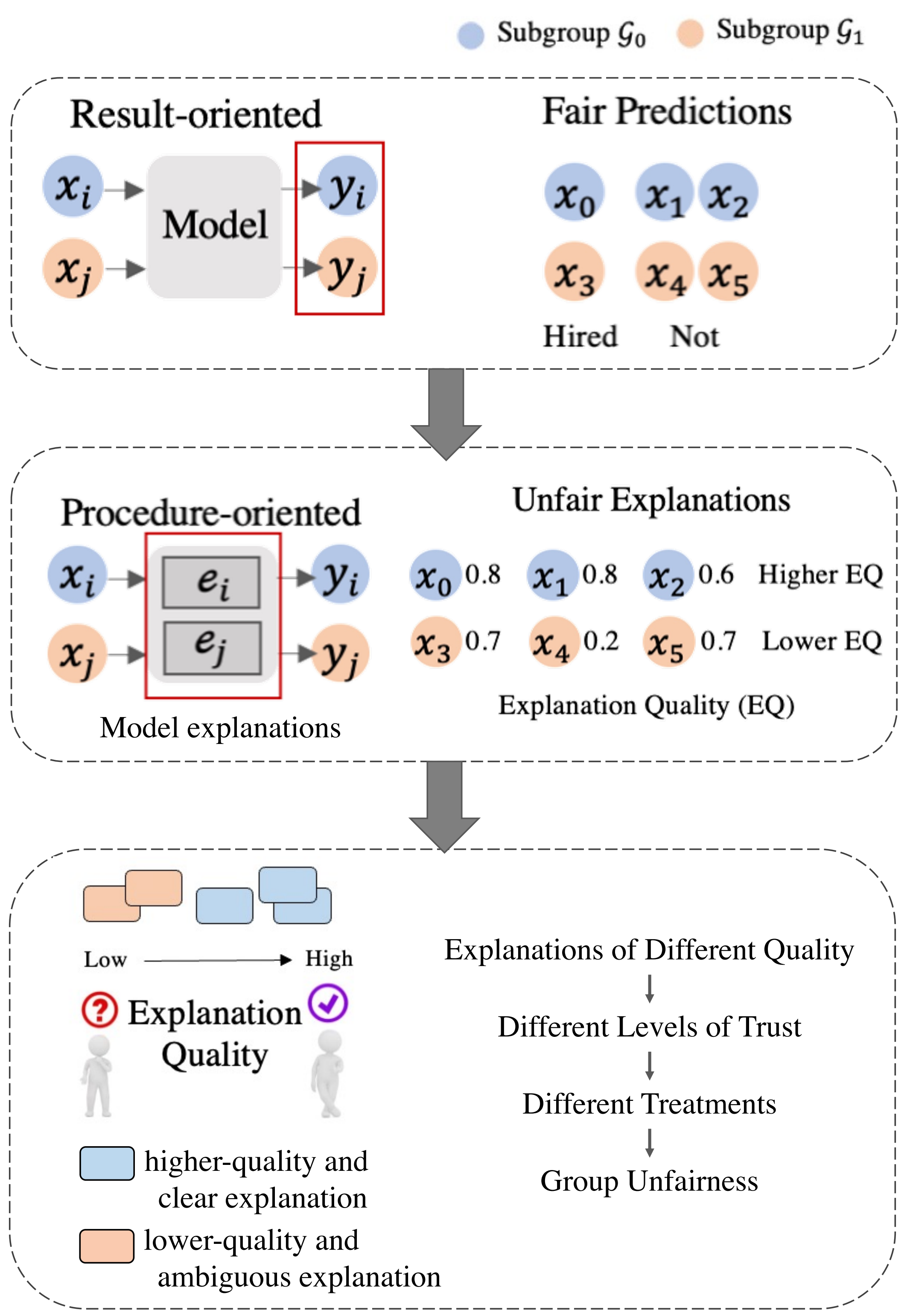


EQ: Fidelity of Model

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

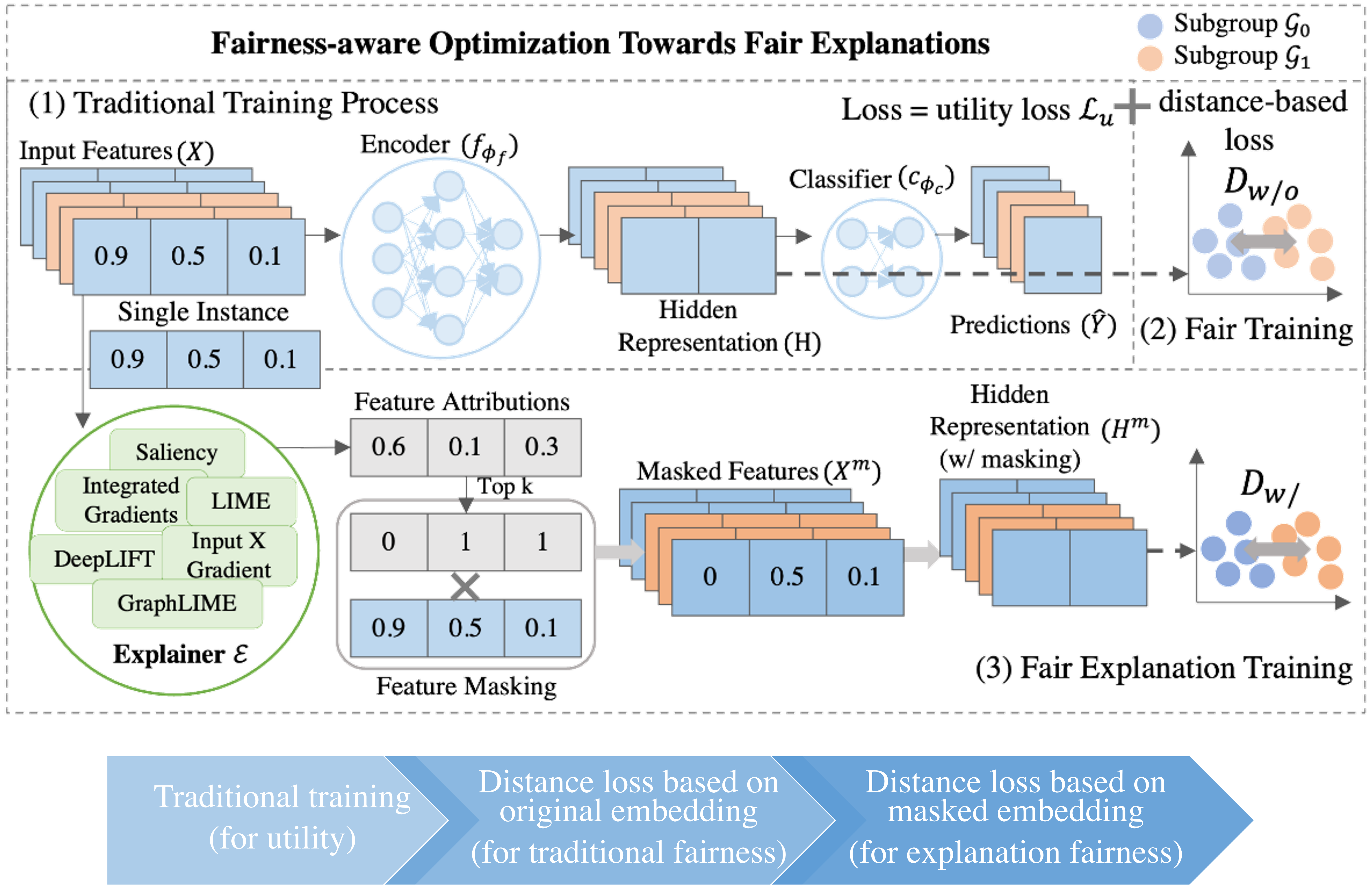


Motivation



Explanations of Different Quality
↓
Different Levels of Trust
↓
Different Treatments
↓
Group Unfairness

Comprehensive Fairness Algorithm (CFA)



Fairness-aware Optimization Towards Fair Explanations

(1) Traditional Training Process

Loss = utility loss $\mathcal{L}_u$ + distance-based loss

(2) Fair Training

(3) Fair Explanation Training

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

Experiments

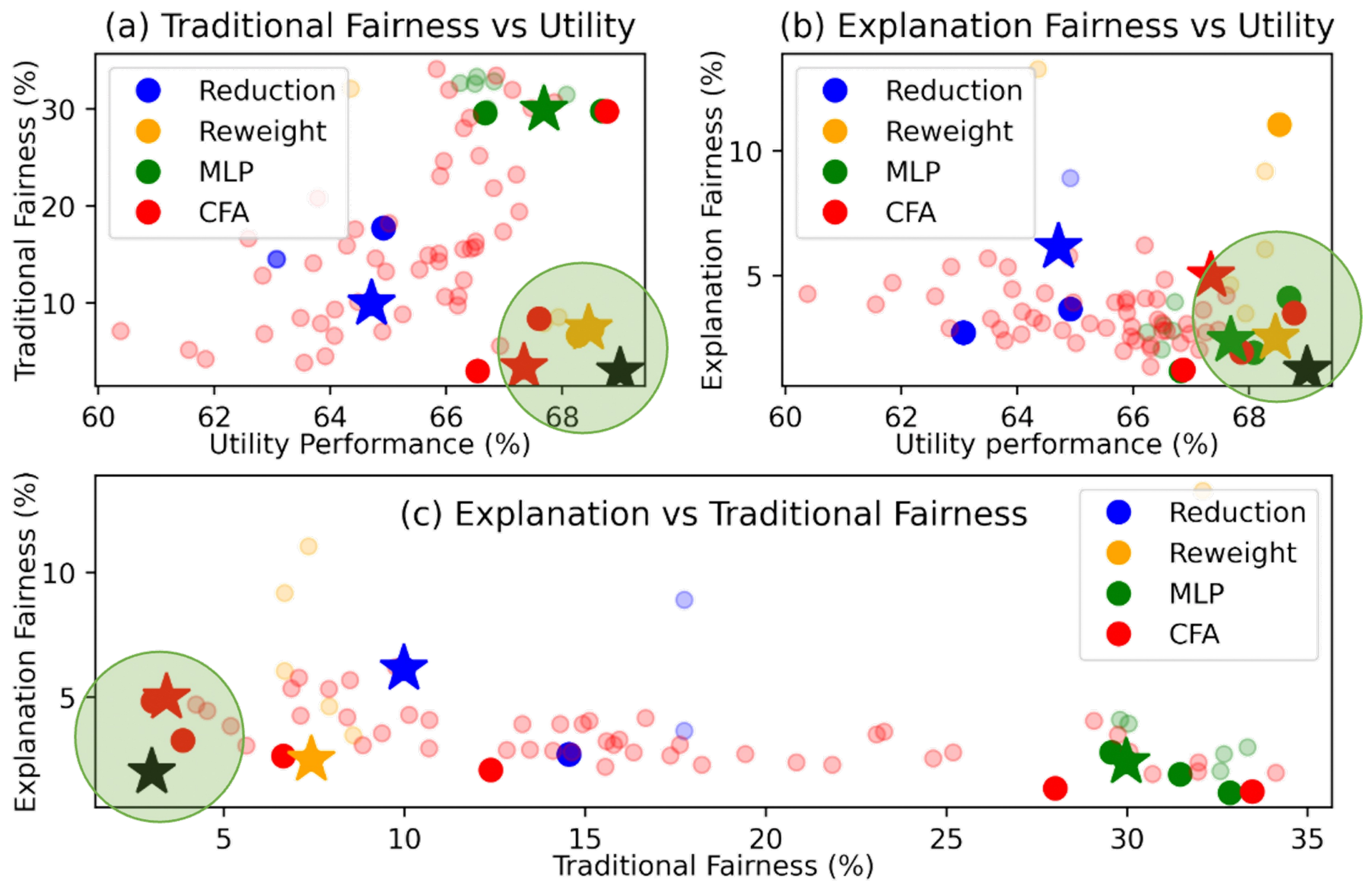
RQ1: Bias Mitigation

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**
	Δ_{SP} ↓	6.07 ± 2.18	2.04 ± 0.00	4.27 ± 0.00	**1.16 ± 0.49**
	Δ_{EO} ↓	3.19 ± 0.73	4.66 ± 0.00	3.37 ± 0.00	**1.14 ± 0.39**
	Δ_{REF} ↓	4.45 ± 2.96	**0.53 ± 0.00**	1.34 ± 0.91	1.98 ± 1.23
	Δ_{VEF} ↓	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**
	Δ_{SP} ↓	2.08 ± 0.75	1.93 ± 0.00	1.93 ± 0.00	**1.00 ± 0.72**
	Δ_{EO} ↓	32.35 ± 7.07	**20.59 ± 0.00**	**20.59 ± 0.00**	27.65 ± 5.44
	Δ_{REF} ↓	8.68 ± 3.18	**1.37 ± 0.00**	8.68 ± 0.00	4.66 ± 3.76
	Δ_{VEF} ↓	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
🟨 Overall Score

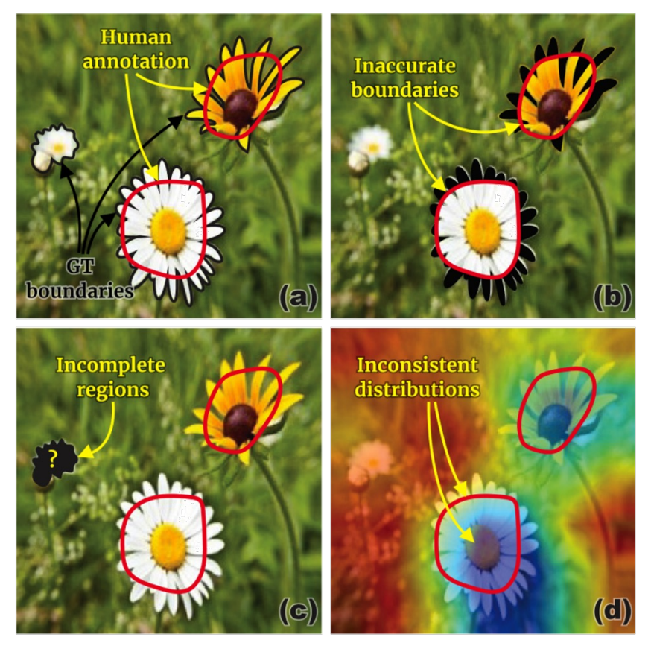
Overall score: $\frac{\mathrm{AUC+F1+ACC}}{3} - \frac{\Delta_{SP}+\Delta_{EO}}{2} - \frac{\Delta_{VEF}+\Delta_{REF}}{2}$ (model selection)

RQ2: Multi-objective Tradeoff

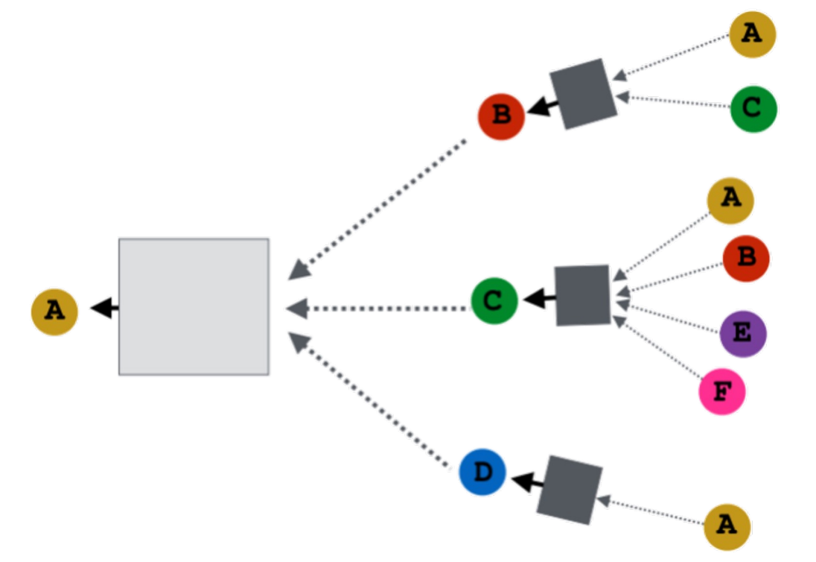


Future Directions

Extending CFA
- Data types
- EQ metrics



Fair and Explainable Graph Neural Networks



Fair Explanation for Inherently Explainable Models (e.g., Decision Tree)