Towards Inclusive Fairness Evaluation via Eliciting Disagreement Feedback from Non-Expert Stakeholders

Mukund Telukunta, **Venkata Sriram Siddhardh (Sid) Nadendla**

Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, USA.

Papers/P1_Disagreements_GF/BIAS 2023/mst-logo-

ALGORITHMIC FAIRNESS VS. HUMAN PERCEPTION

- Algorithmic fairness notions compare predictions with true outcomes
  - Example: In the criminal justice domain, COMPAS' predicted recidivism rate is compared against the true posterior recidivism rates computed during the next two years.
- Algorithmic fairness scores generally take the form [1]

$$f \triangleq \max_k \left(\max_{m,m'} \; f_{m,k} - f_{m',k} \right),$$

where different notions are defined as

Fairness Notion (f)	Groupwise Rate $f_{m,k}$
Statistical Parity (SP)	$SP_{m,k} = \mathbb{P}(\hat{y} = k \mid x \in \mathcal{X}_m)$
Calibration (C)	$C_{m,k} = \mathbb{P}(y = k \mid \hat{y} = k, x \in \mathcal{X}_m)$
Accuracy Equality (AE)	$AE_{m,k} = \mathbb{P}(\hat{y} = y \mid x \in \mathcal{X}_m)$
Equal Opportunity (EO)	$EO_{m,k} = \mathbb{P}(\hat{y} = k \mid y = k, x \in \mathcal{X}_m)$
Predictive Equality (PE)	$PE_{m,k} = \mathbb{P}(\hat{y} = k \mid y \neq k, x \in \mathcal{X}_m)$
Overall Misclassification Rate (OMR)	$OMR_{m,k} = \mathbb{P}(\hat{y} \neq k \mid y = k, x \in \mathcal{X}_m)$

- Human perception of fairness compares algorithmic predictions against people's outcome predictions [4].
  - True label observed in hindsight, y, is replaced with critic's label $\tilde{y}$
  - Need such an approach for a quick preliminary fairness evaluation.

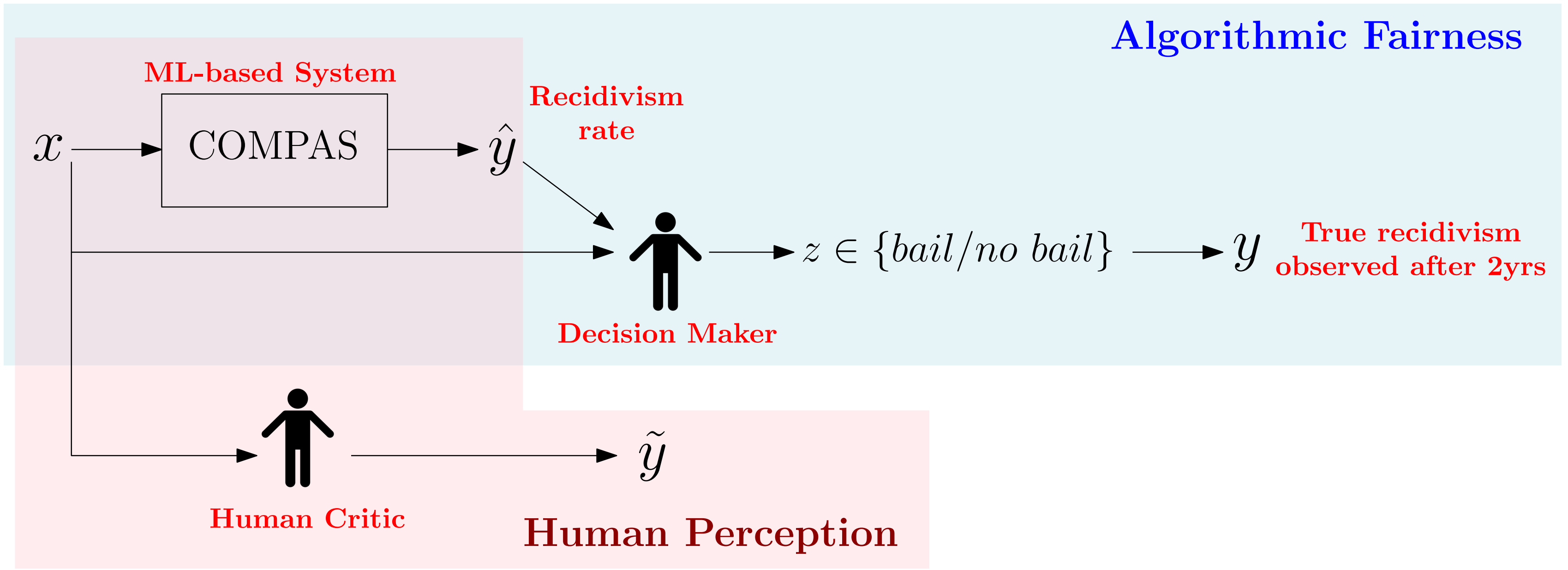


Figure 1: Algorithmic Fairness vs. Human Perception of Fairness in COMPAS

MOTIVATION: OPINIONS FROM DIVERSE STAKEHOLDERS

- Most practical application domains involve diverse stakeholders with varied technical expertise.
  - **Criminal Justice:** Judges, lawyers, prisoners and their family members, other people...
  - **Kidney Transplantation:** Organ Procurement Organizations, Transplant Centers, Surgeons, Recipients, Donors, Donor/Recipient family members, Transport Personnel...
- Some stakeholders lack technical expertise
  - Currently, their opinions are neglected!
  - Can only obtain lower-dimensional feedback (e.g. disagreements) at most!

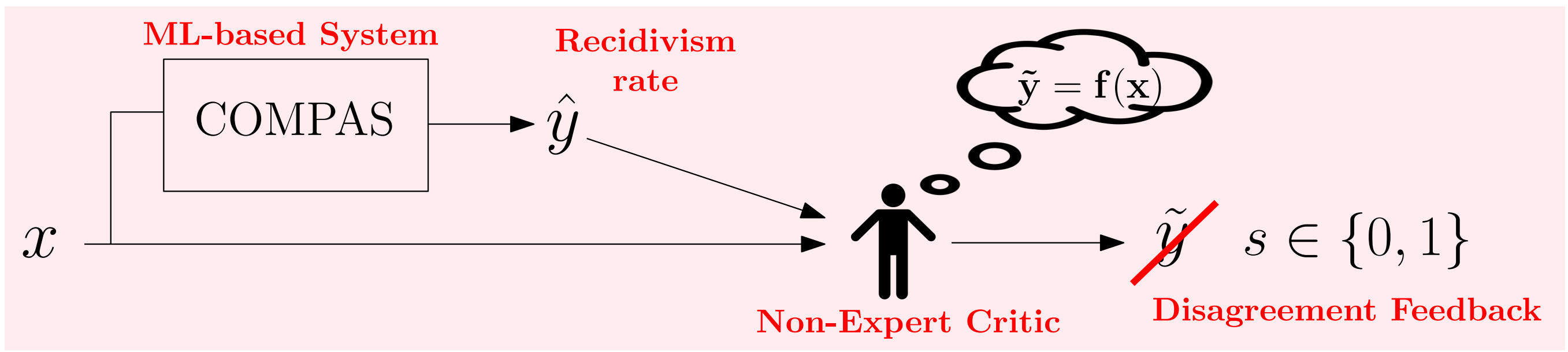
Can we estimate fairness notions using disagreement feedback from non-expert stakeholders? [2]

NON-EXPERT DISAGREEMENT MODEL

Given an input profile $x \in \mathcal{X}$ and outcome label $\hat{y} = g(x)$ from an ML-based classifier $g : \mathcal{X} \to \mathcal{Y}$, the non-expert disagreement model is given by

$$s \;=\; \begin{cases} 1, & \text{if } \tilde{y} \neq \hat{y}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where, $\tilde{y}$ is the unknown non-expert's intrinsic label.



Hence, the disagreement rate with respect to the group $\mathcal{X}_m$ is defined as

$$DR_m \;=\; \mathbb{P}(s = 1 \mid x \in \mathcal{X}_m) \;=\; \mathbb{P}(\tilde{y} \neq \hat{y} \mid x \in \mathcal{X}_m) \tag{2}$$

Furthermore, for a given outcome label $k \in \mathcal{Y}$ be denoted as

$$DR_{m,k} \;=\; \mathbb{P}(s = 1 \mid \hat{y} = k, x \in \mathcal{X}_m) \;=\; \mathbb{P}(\tilde{y} \neq k \mid \hat{y} = k, x \in \mathcal{X}_m) \tag{3}$$

DEFINITE NOTIONS

Group fairness notions that can be *exactly* computed from disagreement rates.

Proposition 1: *Calibration* of the ML-based system is given as

$$CA = \max_k \left(\min_{m,m'} \; DR_{m,k} - DR_{m',k} \right). \tag{4}$$

Proposition 2: *Accuracy Equality* of the ML-based system is given as

$$AE = \max_k \left(\min_{m,m'} \sum_{k \in \mathcal{Y}} DR_{m,k} \cdot SP_{m,k} - \sum_{k \in \mathcal{Y}} DR_{m',k} \cdot SP_{m',k} \right). \tag{5}$$

INDEFINITE NOTIONS

Group fairness notions that can be *estimated* from disagreement rates.

Proposition 3: *Equal Opportunity* of the system can be estimated as

$$\hat{EO} = \frac{1}{2}\left[\max_k \left(\phi(m,k) - 1\right) + \max_k \left(1 - \phi(m',k)\right) \right], \tag{6}$$

where $\phi(m,k) = \max_m \dfrac{(1 - DR_{m,k}) \cdot SP_{m,k}}{(1 - DR_{m,k}) \cdot SP_{m,k} + \sum_{l \neq k} SP_{m,l}}$.

Proposition 4: *Predictive Equality* of the system can be estimated as

$$\hat{PE} = \frac{1}{2}\left[\max_k \left(\mu(m,k) - 1\right) + \max_k \left(1 - \mu(m',k)\right) \right], \tag{7}$$

where $\mu(m,k) = \max_m \dfrac{DR_{m,k} \cdot SP_{m,k}}{DR_{m,k} \cdot SP_{m,k} + \sum_{l \neq k} SP_{m,l}}$.

Proposition 5: *Overall misclassification rate* of the system is given as

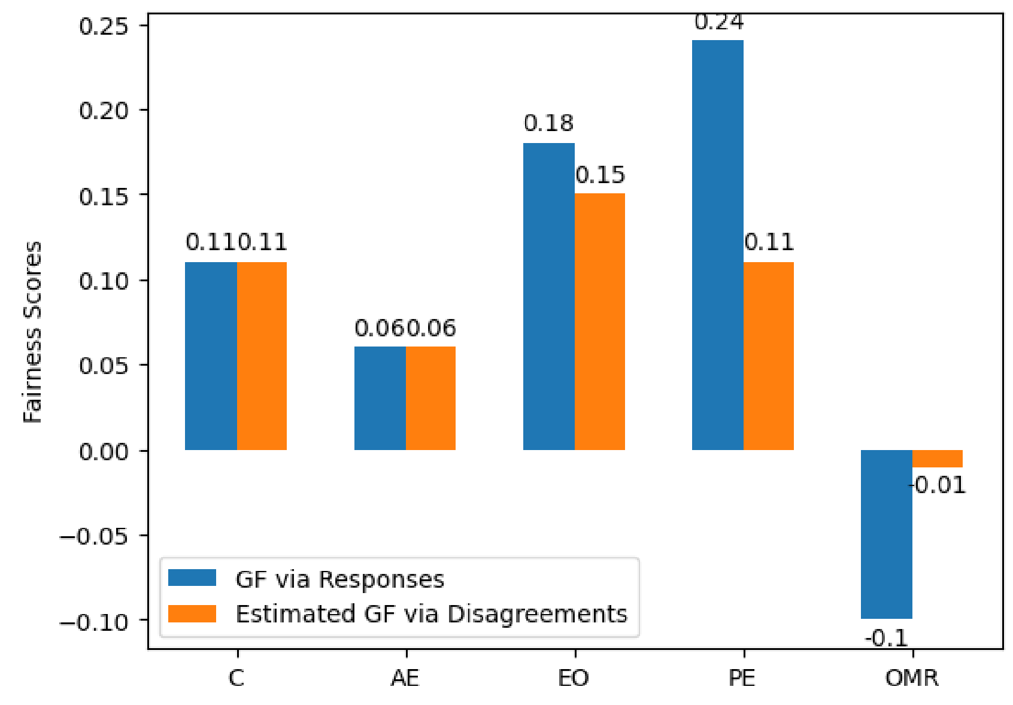
$$\hat{OMR} = \frac{1}{2}\left[\max_k \left(\omega(m,k) - 1\right) + \max_k \left(1 - \omega(m',k)\right) \right], \tag{8}$$

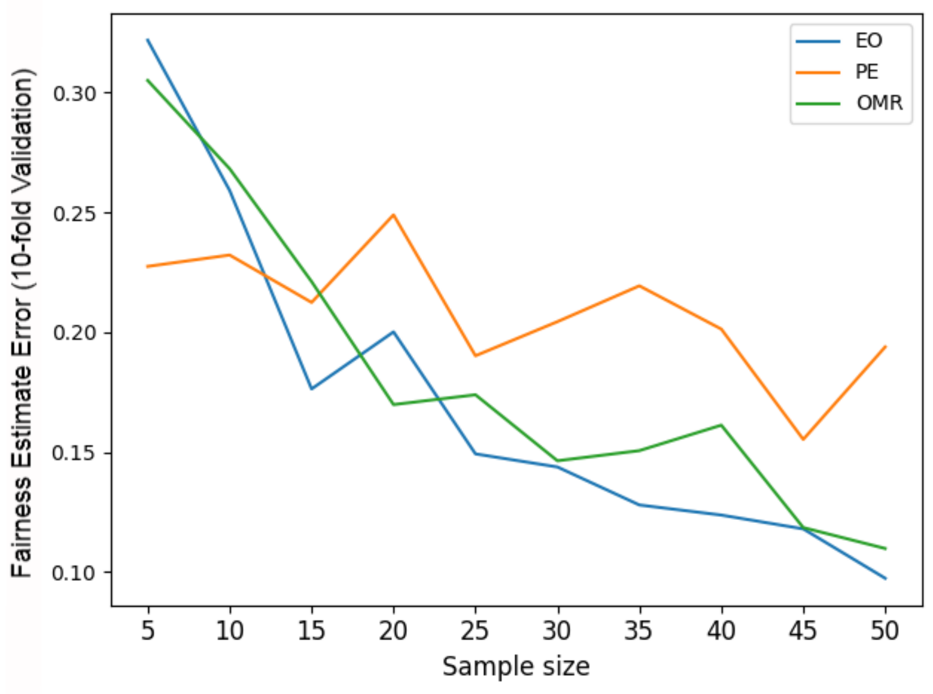
where $\omega(m,k) = \max_m \dfrac{\sum_{l \neq k} SP_{m,l}}{(1 - DR_{m,k}) \cdot SP_{m,k} + \sum_{l \neq k} SP_{m,l}}$.

VALIDATION USING A REAL DATASET

Dataset: Real human feedback curated by Dressel and Farid [3].

- 1000 defendant descriptions from COMPAS dataset
- 400 critics responded *yes* or *no* to "Will this person recidivate in 2 years?".
- Critics' responses are aggregated based on majority rule.
- Critic disagreements: $s = critic\_feedback \oplus compas\_label$.





CONCLUSION AND FUTURE WORK

- Proposed a novel and inclusive disagreement-based feedback model for non-expert stakeholders.
- Fairness Estimation: (i) Definite notions can be precisely quantified from disagreement rates, (ii) Indefinite notions can be estimated from bounds.
- In the future, we will apply the proposed feedback model to kidney placement to collect patient and donor opinions.

REFERENCES

[1] W. Alghamdi and et al. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, 2022.

[2] A. Chouldechova and M. G'Sell. Fairer and More Accurate, but for Whom? *arXiv preprint arXiv:1707.00046*, 2017.

[3] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

[4] M. Yaghini, A. Krause, and H. Heidari. A human-in-the-loop framework to construct context-aware mathematical notions of outcome fairness. In *Proceedings of AIES 2021*, pages 1023–1033, 2021.

https://sid-nadendla.github.io September 22, 2023 nadendla@mst.edu