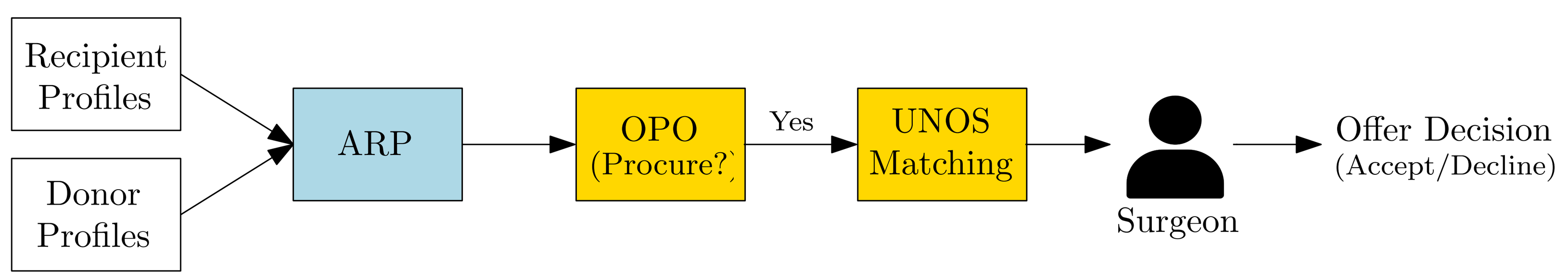
Learning Social Fairness Preferences from Non-Expert Stakeholder Opinions in Kidney Placement

Mukund Telukunta [1] Sukruth Rao [2] Gabriella Stickney [2] Venkata Sriram Siddardh Nadendla [1] Casey Canfield [1]

[1]Missouri University of Science and Technology [2]Michigan State University

Machine Learning in Kidney Placement: Concerns



Acceptance Rate Predictor (ARP) supports organ procurement teams via predicting the probability that a deceased donor kidney gets accepted [1].

- Trained using past kidney placement decisions
- `Race` and `Age` in Kidney Donor Profile Index (KDPI) and Estimated Glomerular Filtration Rate (eGFR) scores.

ARP inherits social biases from past kidney placement decisions!

Group Fairness Tradeoffs and Fairness Preferences

Group Fairness [2]: Compare ARP's statistical performance (function of predicted offer acceptance rate $\hat{y}$ and patient survival outcome y) across two social groups $\mathcal{X}_m, \mathcal{X}_{m'}$, i.e. compute $f \triangleq \max_{m,m'} f_m - f_{m'}$, where

Fairness Notion (f)	Groupwise Rate f_m
Statistical Parity (SP)	$SP = \mathbb{P}(\hat{y} = 1 \mid x \in \mathcal{X}_m)$
Calibration (C)	$C = \mathbb{P}(y = 1 \mid \hat{y} = 1, x \in \mathcal{X}_m)$
Accuracy Equality (AE)	$AE = \mathbb{P}(\hat{y} = y \mid x \in \mathcal{X}_m)$
Equal Opportunity (EO)	$EO = \mathbb{P}(\hat{y} = 1 \mid y = 1, x \in \mathcal{X}_m)$
Predictive Equality (PE)	$PE = \mathbb{P}(\hat{y} = 1 \mid y = 0, x \in \mathcal{X}_m)$
Overall Misclassification Rate (OMR)	$OMR = \mathbb{P}(\hat{y} = 0 \mid y = 1, x \in \mathcal{X}_m)$

Challenges in evaluating ARP's fairness:

1. Group fairness notions exhibit fundamental trade-offs [3].
   - Which notion of fairness does evaluators prefer?
2. Fairness evaluations only by surgeons who forecast patient outcomes.
   - What about fairness opinions of non-expert stakeholders (e.g. patients, donors)?

Survey Design

Prolific survey deployed on in Dec 2023: Recruited 85 participants.

- Kidney matching data from OPTN's Standard Transplant Analysis and Research (STAR) datasets.
- 10 data tuples (donor, 10 matched recipients, surgeon's decisions y, ARP outputs $\hat{y}$) per participant.
- We ask: On a scale of 1-7, rate the fairness of the ARP outputs. Here 1 indicates completely unfair and 7 indicates completely fair.

Race		Age		Gender	
White	60%	18-25	8%	Male	49%
Black	19%	25-40	57%	Female	49%
Asian	12%	40-60	29%	Non-binary	2%
Hispanic	3.4%	>60	6%		
Other	5.6%				

Table 1. Participant Demographics

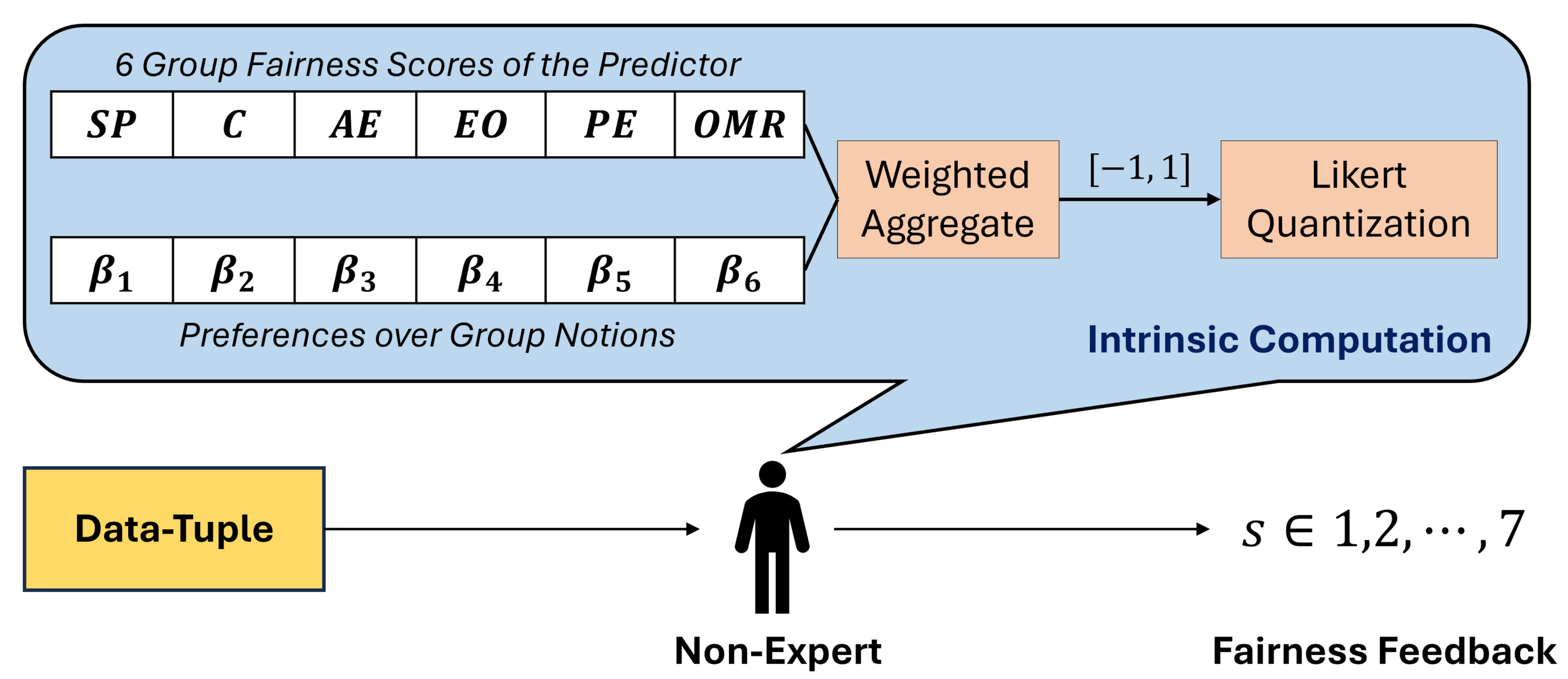
Fairness Feedback Model

Assumption: Participants exhibit an unknown **weighted preference** over L group fairness notions.

1. Participant's fairness preferences (weights): $\beta = \{\beta_1, \cdots, \beta_L\}$
2. Participant's **Intrinsic Weighted Fairness Evaluation:**

$$\psi = \text{Preferences} \odot \text{Fairness Scores} \in [-1, 1]$$

   - If ψ is -1 or 1, the predictor is deemed **unfair**.
   - If ψ is closer to 0, the predictor is **fair**.
3. Participant receives utility u following Logit-Normal distribution with parameters μ and σ.
4. Estimated fairness evaluation $\tilde{s}$: modeled as Mixed-Logit probability [4].



Social Aggregation of Fairness Feedback

Given N non-expert participants each receiving M data-tuples, the social preference weight $\boldsymbol{\beta}^*$ is computed by minimizing the feedback regret

$$\mathcal{L}_F(\boldsymbol{\beta}) \triangleq \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{N} \sum_{n=1}^{N} \| s_{n,m} - \tilde{s}_m^*(\boldsymbol{\beta}) \|_2^2 \right), \tag{1}$$

Projected Gradient Descent: $\boldsymbol{\beta}^{(e+1)} \leftarrow \mathbb{P}\left[\boldsymbol{\beta}^{(e)} - \delta \cdot \nabla \mathcal{L}_F(\boldsymbol{\beta}^{(e)})\right]$

Computation of Loss Gradient

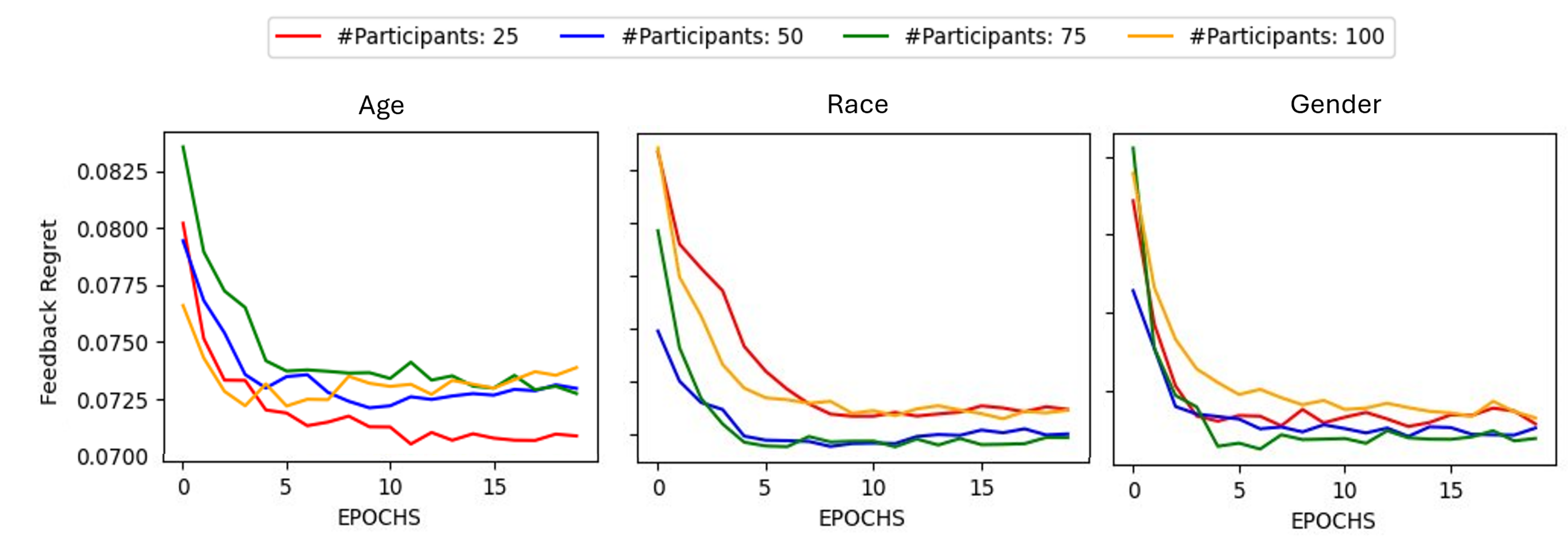
Dependency chain of variables: $\mathcal{L}_F \leftarrow \tilde{s}^* \leftarrow \boldsymbol{u} \leftarrow \boldsymbol{\psi} \leftarrow \boldsymbol{\beta}$

$$\nabla_{\boldsymbol{\beta}} \mathcal{L}_F = \underbrace{(\nabla_{\tilde{s}^*} \mathcal{L}_F)^T}_{\text{Regret Gradient}} \cdot \underbrace{(\nabla_{\boldsymbol{u}} \tilde{s}^*)^T}_{\text{Social Feedback Gradient}} \cdot \underbrace{(\nabla_{\boldsymbol{\psi}} u)^T}_{\text{Utility Gradient}} \cdot \underbrace{\nabla_{\boldsymbol{\beta}} \boldsymbol{\psi}}_{\text{Fairness Evaluation Gradient}}$$

- **Regret Gradient** (Known)
- **Social Feedback Gradient** (Known)
- **Utility Gradient** — Depends on:
  - Likert Quantization
  - log-Normal Distri.
  - (Closed form expression provided)
- **Fairness Evaluation Gradient** (Known)

Results

Simulation Experiments: 15 data-tuples to $N = 25, 50, 75, 100$ simulated non-experts ⇒ Feedback regret converges within 5 epochs.



Survey Experiment: Accuracy Equality ⇒ Crowd's most preferred notion.

- Biases only matter if surgeon rejects the offer
- Some preference to demographic parity

Sensitive Attribute	Social Fairness Preference					
	SP	C	AE	EO	PE	OMR
Age	0.15	0	**0.45**	0.007	0.37	0.01
Gender	0.19	0.02	**0.48**	0	0.24	0.06
Race	0.28	0.10	**0.38**	0	0.19	0.03

References

[1] L. Ashiku, R. Threlkeld, C. Canfield, and C. Dagli, "Identifying AI Opportunities in Donor Kidney Acceptance: Incremental Hierarchical Systems Engineering Approach," in 2022 IEEE International Systems Conference (SysCon), pp. 1–8, IEEE, 2022.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1–35, 2021.

[3] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," Innovations in Theoretical Computer Science (ITCS) Conference, 2017.

[4] D. McFadden et al., "Conditional Logit Analysis of Qualitative Choice Behavior," Frontiers in Econometrics, pp. 105–142, 1973.

Acknowledgements: NSF grants #2222801, #2150210