A Human-in-the-loop Framework to Construct Context-aware Mathematical Notions of Outcome Fairness

Mohammad Yaghini (University of Toronto), Andreas Krause (ETH Zürich) and Hoda Heidari (Carnegie Mellon University)

Motivation

Fairness is relative, complex and context-dependent
- No single off-the-shelf definition can capture it

How to define a context-aware notion of fairness?
- Bringing back human judgement into the decision-making loop

Two-pronged solution:
1. Eliciting judgement by pairwise comparison
2. Aggregating judgment through a social choice mechanism

Equality of Opportunity (EOP)

Distinguish morally-justifiable (**desert**) attributes from **circumstantial** ones.
A utility distribution CDF F under policy ϕ satisfies EOP if for all circumstances c, c' and all desert levels d

$$F^{\phi}(.\mid c, d) = F^{\phi}(.\mid c', d)$$

EOP Parameter Estimation

Circumstance. Estimate $\mathbf{c} := \mathbf{z}_p \in \mathbb{R}^k$ for participant p

> To what extent do you agree with the following statement?
> It is ethically acceptable for the attribute [...] to impact the decision a defendant receives.

Desert. Estimate $\mathrm{d_p} = \delta_\mathrm{p} \cdot [\mathrm{x}, \mathrm{y}]$, where $\mathrm{d_p}$ is not directly observable. Assuming there exists a linear function $\mathrm{D_p}: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ such that $\mathrm{d_p} = \mathrm{D_p}(\mathrm{x}, \mathrm{y})) = \delta_\mathrm{p} \cdot [\mathrm{x}, \mathrm{y}]$, we wish to find δ_p

Desert Queries. Q pairwise questions about two scenarios $t_1^q = [\mathbf{x}_1^q, y_1^q]$ and $t_2^q = [\mathbf{x}_2^q, y_2^q]$

> From an ethical standpoint, between the two decision subjects, who do you believe *deserves* a more lenient decision?

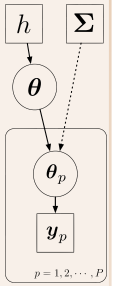
Utility. Similar to desert, estimate $u_p = v_p \cdot [\mathbf{x}, y, \hat{y}]$ with Q pairwise questions about two scenarios $t_1^q = [\mathbf{x}_1^q, y_1^q, \hat{y}_1^q]$ and $t_2^q = [\mathbf{x}_2^q, y_2^q, \hat{y}_2^q]$:

> …, who do you think *will benefit more* from their algorithmic decision?

MLE. Find δ_p and $\mathrm{v_p}$ that maximizes the likelihood of observed desert/utility differences

Preference Aggregation (Social Choice)

Borda Count.
- Feature is circumstantial if most participants agree
- δ_p (and v_p) are averaged



Hierarchical Bayesian Model. Joint Parameter Estimation of
$\theta :=$ society's preference vector
$\boldsymbol{\theta}_p \sim \mathcal{N}(\boldsymbol{\theta}, \Sigma)$ is participant p's

$$\underset{\theta_p, \boldsymbol{\theta}}{\operatorname{argmin}} - \sum_p \sum_q log\, \Phi\left(a^q \boldsymbol{\theta}_p \cdot [\mathbf{x}_1^{p,q} - \mathbf{x}_2^{p,q}, y_1^{p,q} - y_2^{p,q}]\right)$$

$$\text{s.t.} \parallel \boldsymbol{\theta}_p - \boldsymbol{\theta} \parallel_2 \leq \lambda, \|\boldsymbol{\theta}_p\|_2 \leq 1, \parallel \boldsymbol{\theta} \parallel_2 \leq 1$$

Experimental Setup

From an ethical standpoint, between the two following decision subjects, **who do you think will benefit more from their algorithmic decision?**

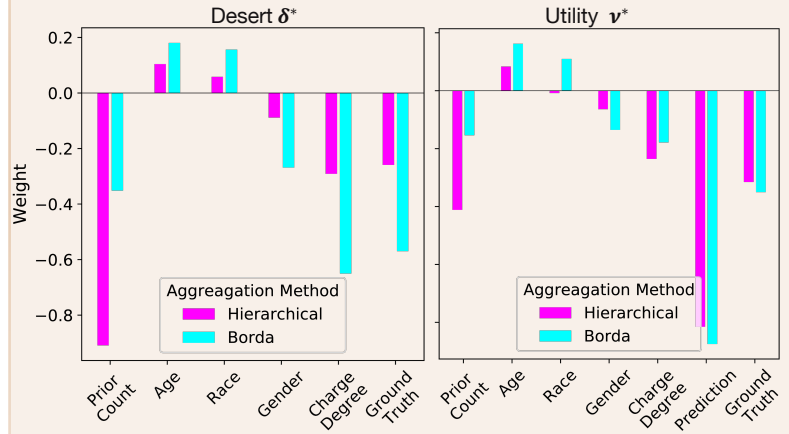
ATTRIBUTE	SUBJECT #1	SUBJECT #2
Age Category	Older than 25	Older than 25
Race	Non-white	White
Gender	Male	Male
Charge Degree	Misdemeanor	Felony
Prior Counts	4	4
Algorithmic Decision	Low risk to reoffend	Low risk to reoffend
Actual Outcome	Will not reoffend	Will not reoffend

Note: The decision subject differences are marked in **blue**. If you are unsure about the meaning of any attribute, hold the cursor on it to see a definition.

Proof-of-concept study ran on AMT
- 99 participants
- Conversational Interface

Study Results (on AMT with 99 participants)

- Mixed opinion on age. Some thought younger people are less in control of their leniency



- EOP performs better in terms of equalizing the utility distributions (instead of pre-defined metrics) across groups
- EOP improves the utility for the whole populations; regardless of desert group