Superhuman Fairness

Omid Memarrast, Linh Vu, Brian D. Ziebart

UIC COMPUTER SCIENCE

University of Illinois Chicago

Motivation

Defining desired fairness trade-offs precisely is difficult

- **Multiple fairness metrics** [dp, eqodds, eqopp, prp, ...]

A new perspective: Multiple stakeholders

- with **different notions of fairness** and **desired performance-fairness trade-offs**

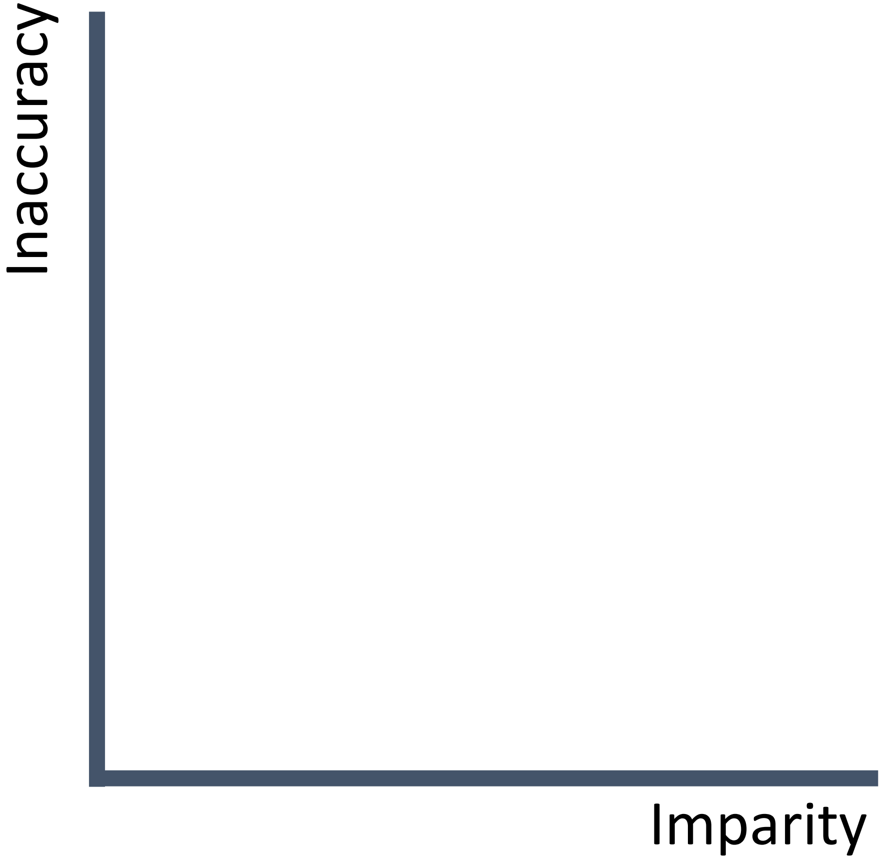
Example

- Admission: [CS department, Civil department, ..]
- Each department: Their own perception of fairness

Solution:

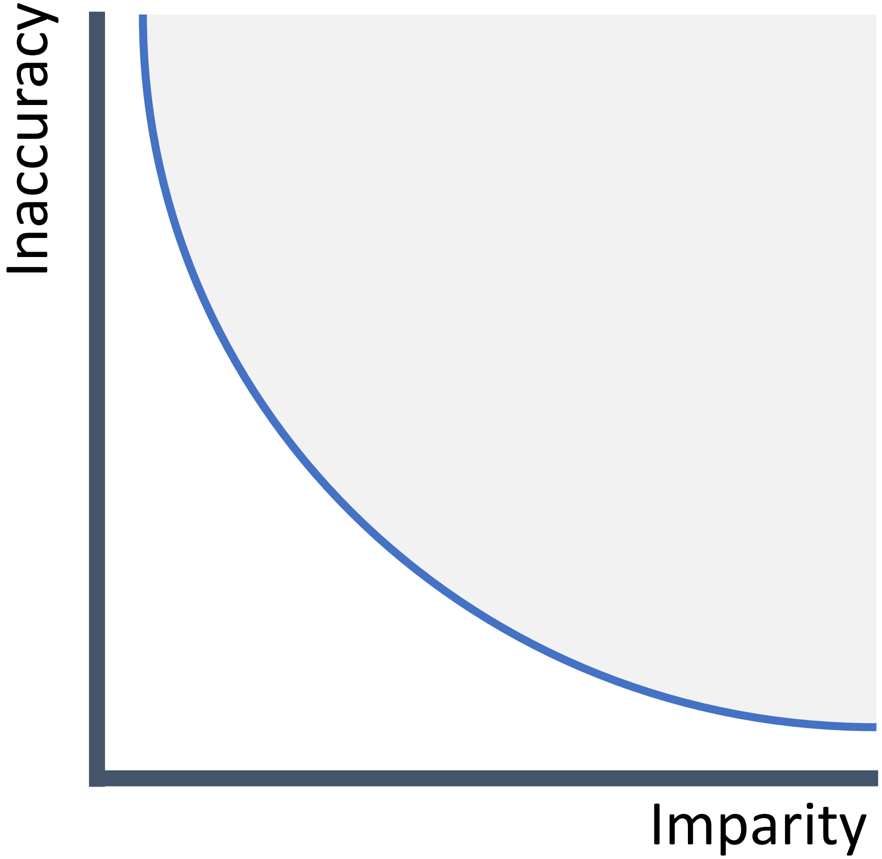
Instead of optimal fairness, **outperform humans** across many metrics

Defining Superhuman Behavior



metrics f$_1$, f$_2$, ...

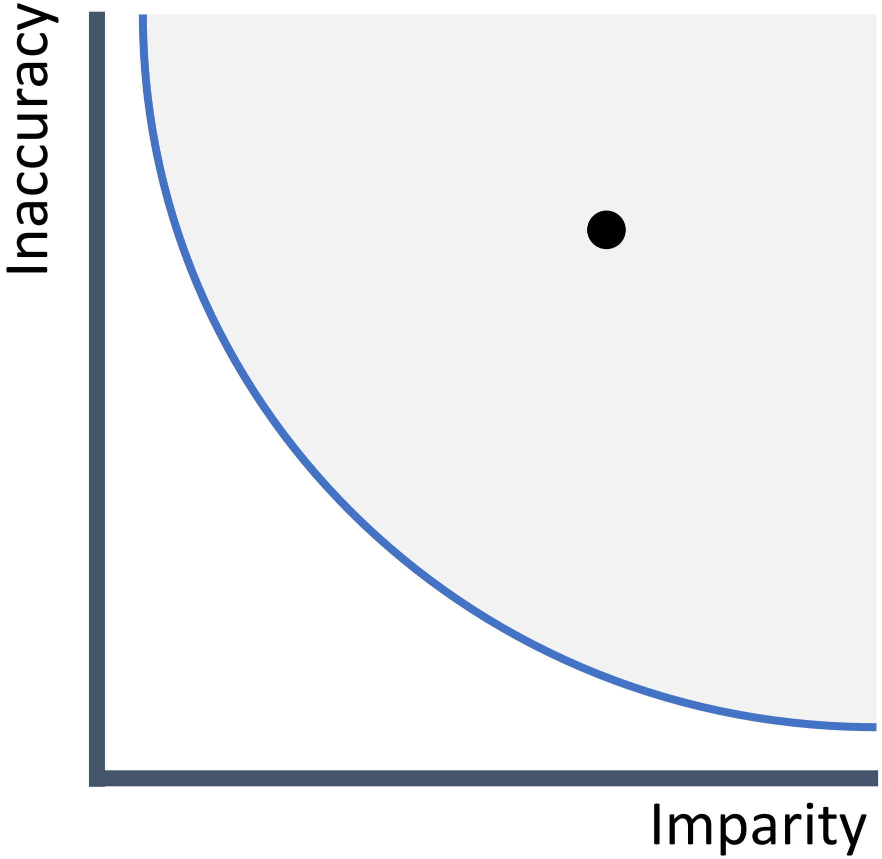
Defining Superhuman Behavior



metrics f$_1$, f$_2$, …

Pareto frontier

Defining Superhuman Behavior

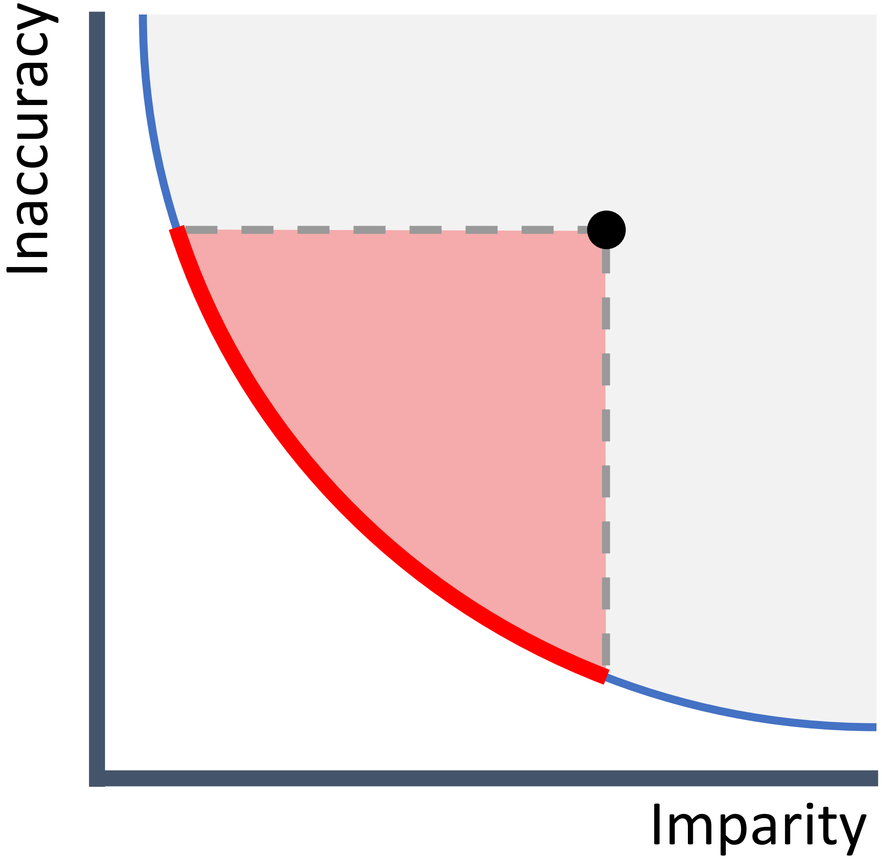


metrics $f_1, f_2, \ldots$

human demonstrations

Pareto frontier

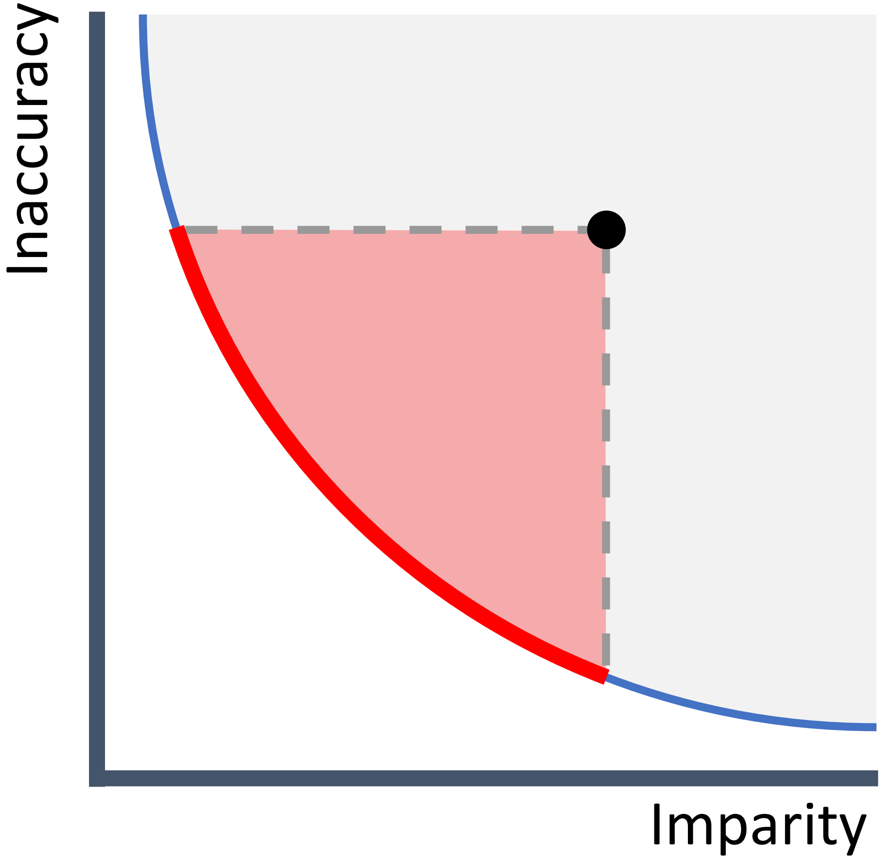
Defining Superhuman Behavior



A **policy** is **superhuman** if it has smaller **metrics** f_1, f_2, … for all **human demonstrations**

Pareto frontier

Defining Superhuman Behavior

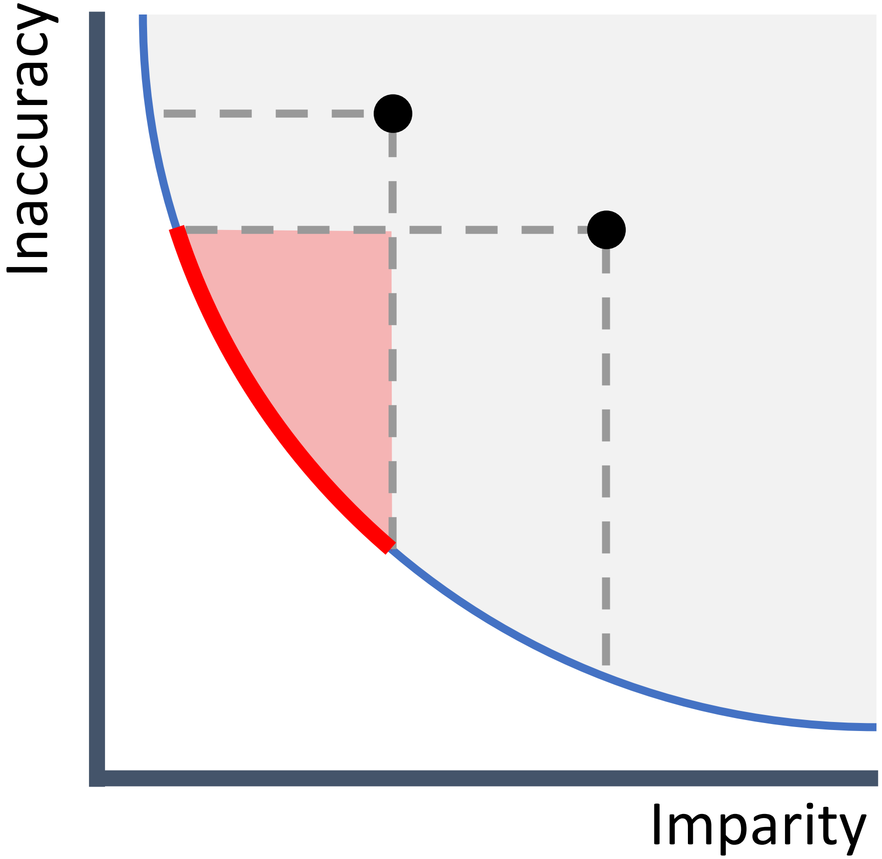


A **policy** is **superhuman** if it has smaller **metrics** f_1, f_2, … for all **human demonstrations**

Guarantees lower cost than demonstration costs for family of additive trade-offs

Pareto frontier

Defining Superhuman Behavior

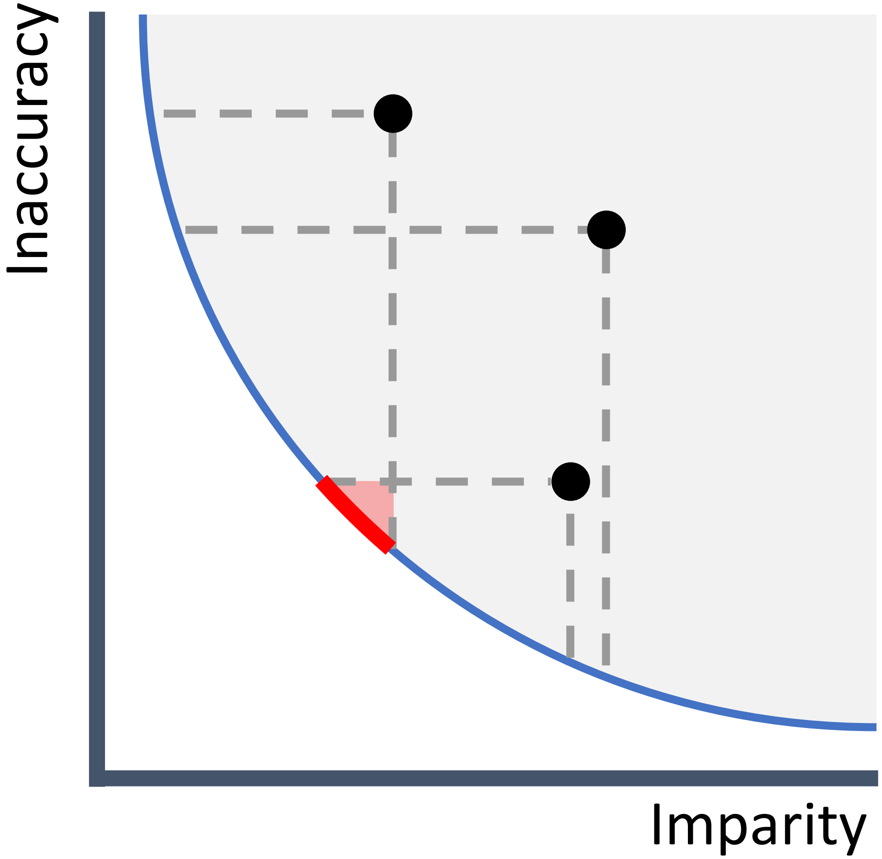


A **policy** is **superhuman** if it has smaller **metrics** f_1, f_2, … for all **human demonstrations**

Guarantees lower cost than demonstration costs for family of additive trade-offs

Set of **superhuman policies** on the **Pareto frontier** shrinks as demonstrations grow

Defining Superhuman Behavior

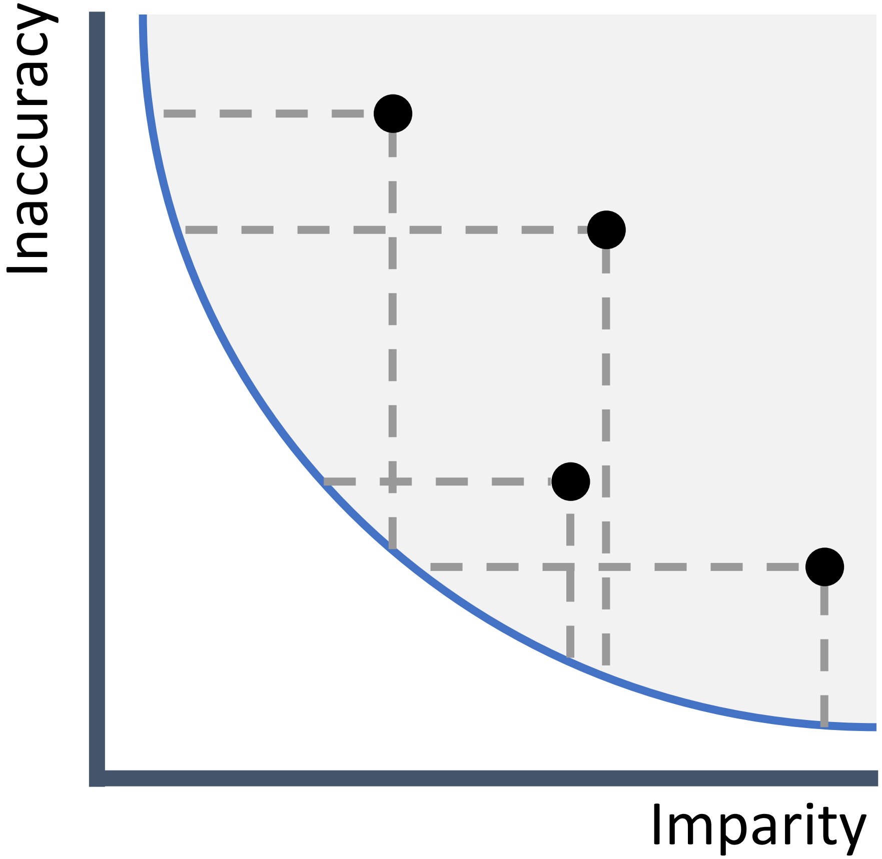


A **policy** is **superhuman** if it has smaller **metrics** f_1, f_2, … for all **human demonstrations**

Guarantees lower cost than demonstration costs for family of additive trade-offs

Set of **superhuman policies** on the **Pareto frontier** shrinks as demonstrations grow

Defining Superhuman Behavior



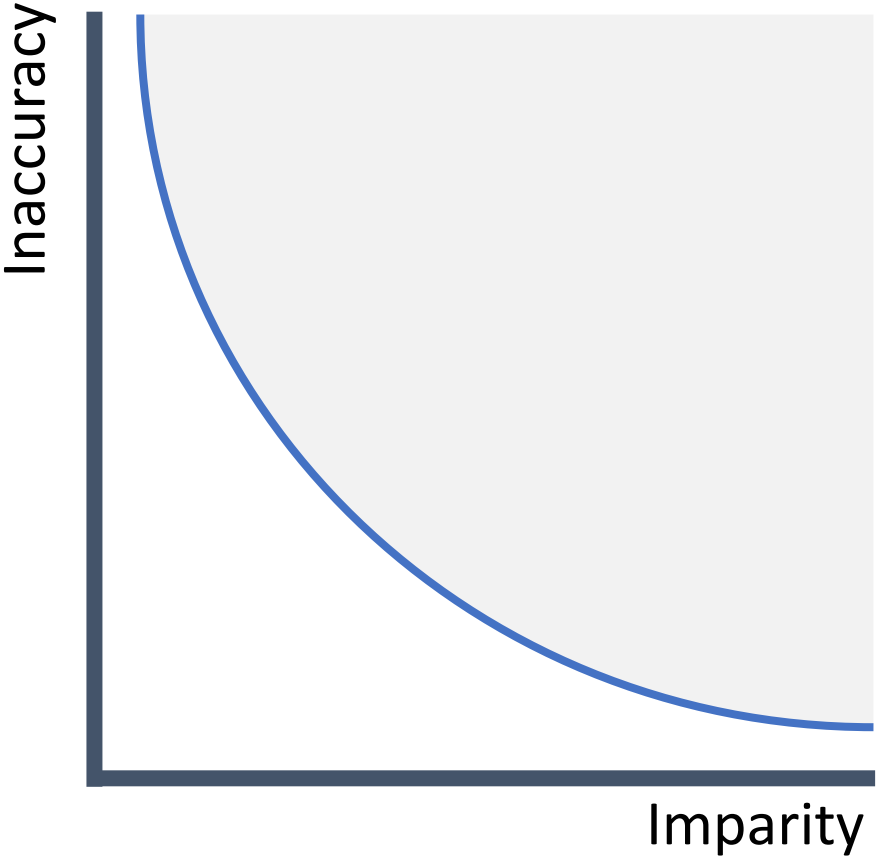
A **policy** is **superhuman** if it has smaller **metrics** f_1, f_2, … for all **human demonstrations**

Guarantees lower cost than demonstration costs for family of additive trade-offs

Set of **superhuman policies** on the **Pareto frontier** shrinks as demonstrations grow

Can become empty!

Superhuman Percentile & Subdominance



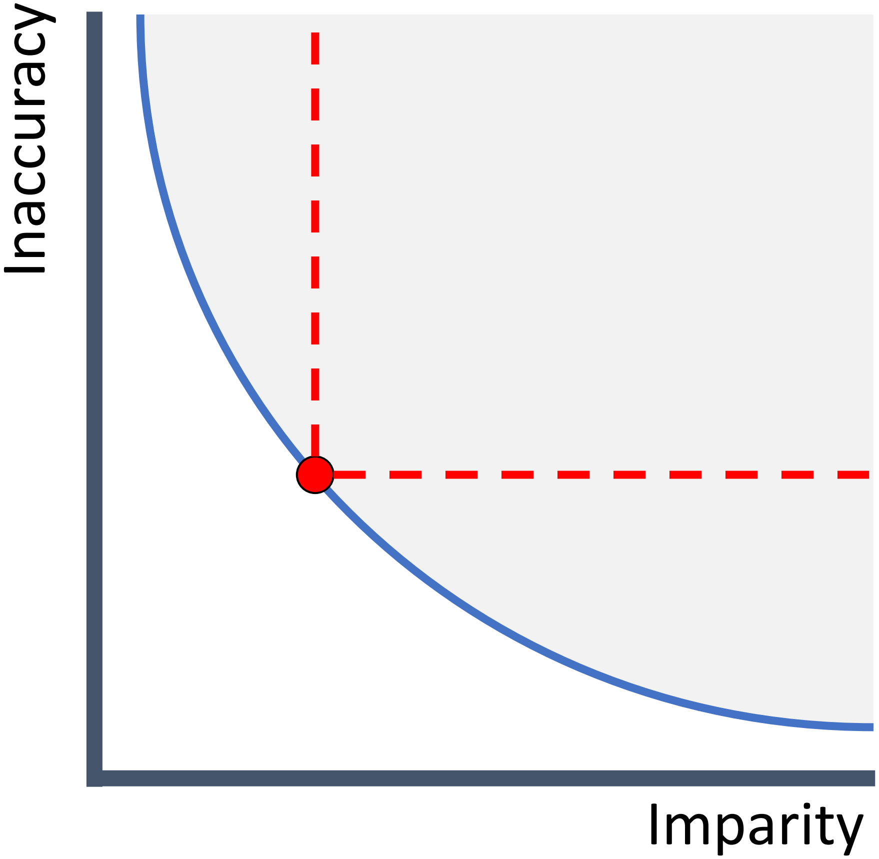
Pareto frontier

metrics f_1, f_2, … t

Superhuman Percentile & Subdominance

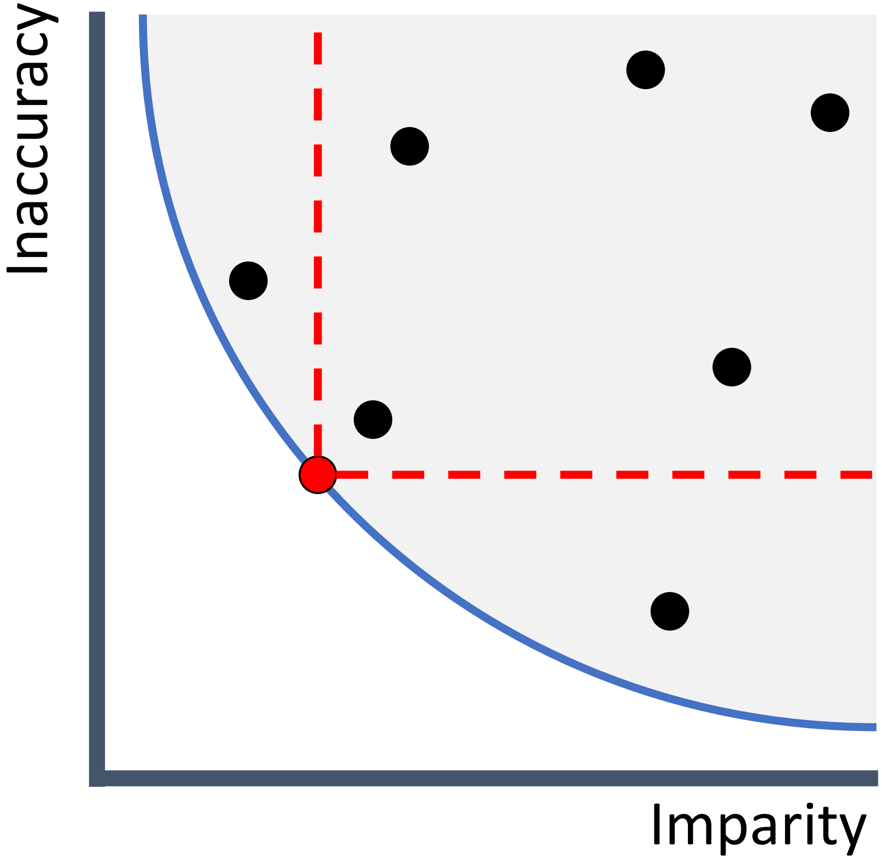
policy

metrics f_1, f_2, ...



Pareto frontier

Superhuman Percentile & Subdominance



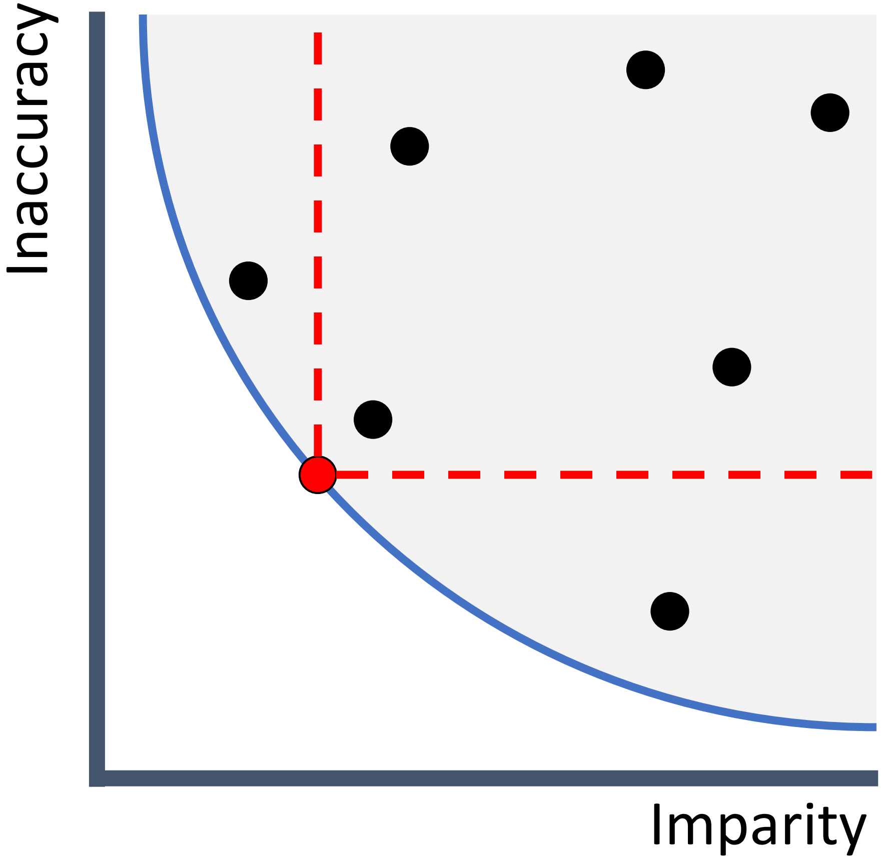
policy

metrics f$_1$, f$_2$, …

human demonstrations

Pareto frontier

Superhuman Percentile & Subdominance

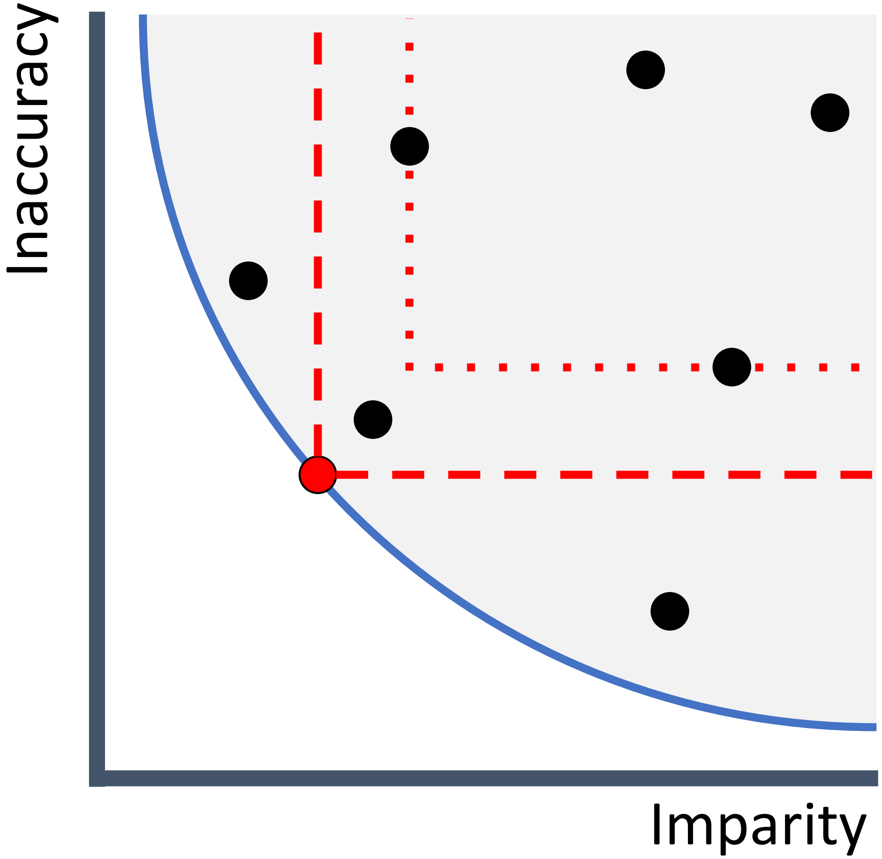


A **policy** is **γ-superhuman** if it has smaller **metrics** f_1, f_2, … than γ% of **human demonstrations**

Pareto frontier

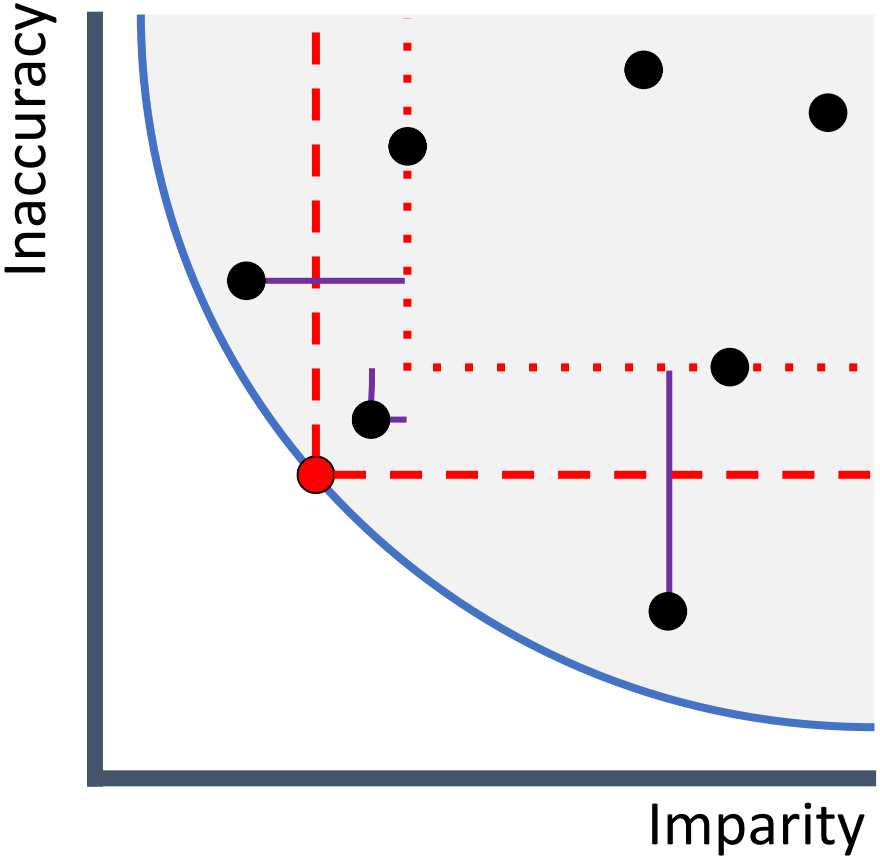
Superhuman Percentile & Subdominance

A **policy** is **γ-superhuman** if it has smaller **metrics** f_1, f_2, … than γ% of **human demonstrations**



Pareto frontier

Superhuman Percentile & Subdominance

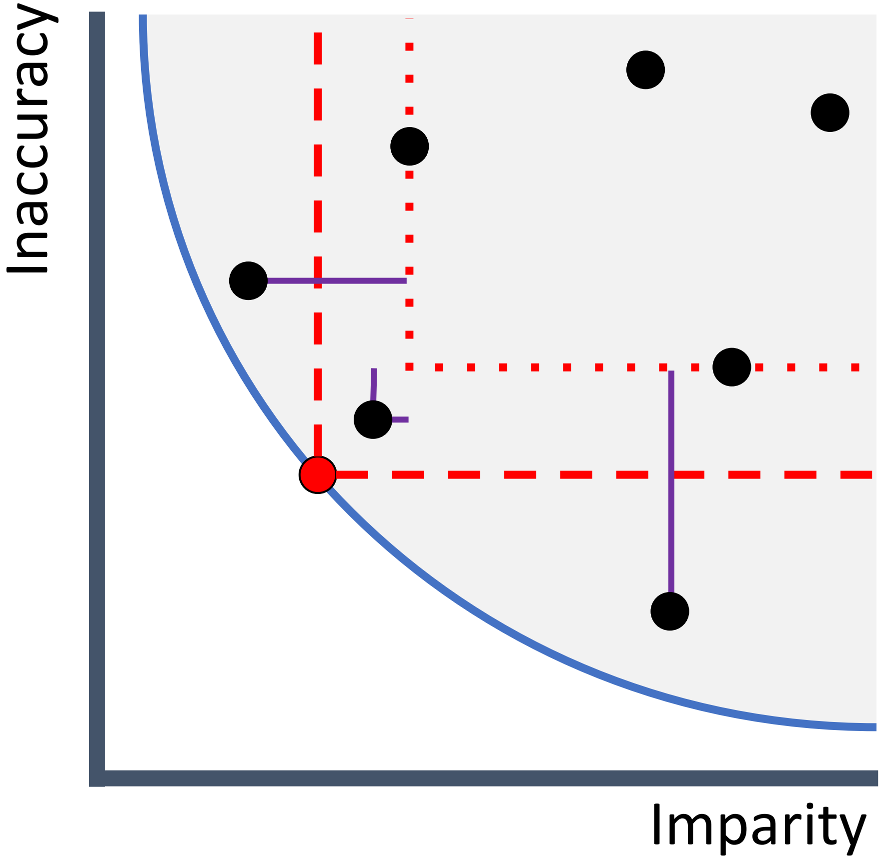


A **policy** is **γ-superhuman** if it has smaller **metrics** f_1, f_2, … than γ% of **human demonstrations**

Subdominance measures how far a policy is from superhuman by some **margins**

Pareto frontier

Superhuman Percentile & Subdominance

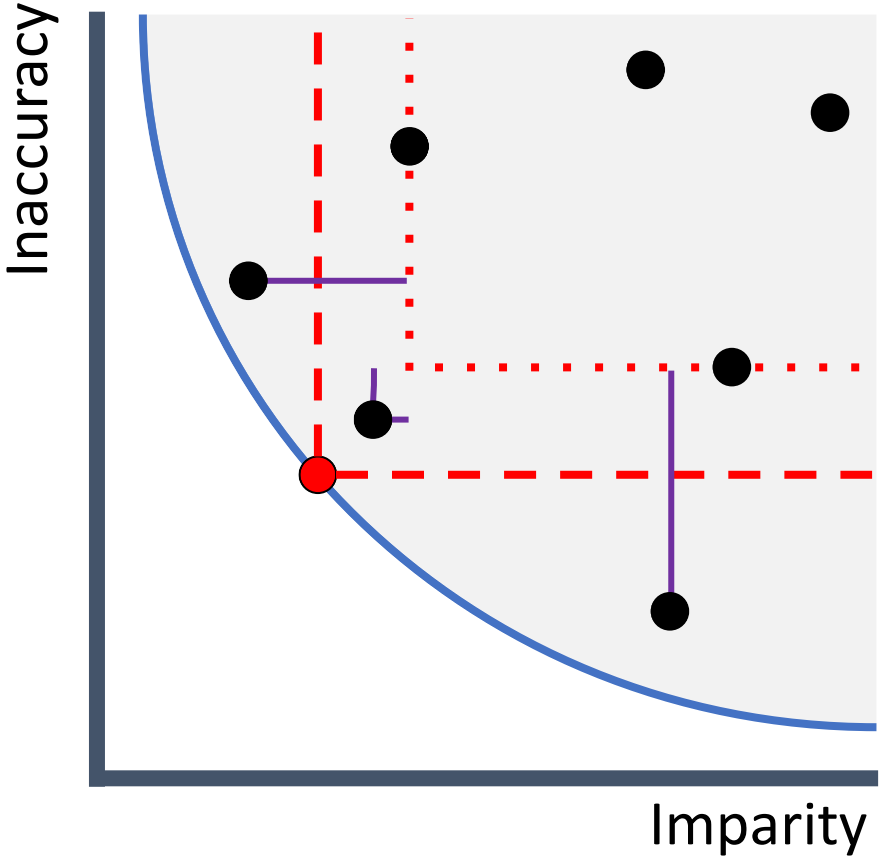


A **policy** is **γ-superhuman** if it has smaller **metrics** f_1, f_2, … than γ% of **human demonstrations**

Subdominance measures how far a policy is from superhuman by some **margins**

Minimally subdominant policy tends to reside close to the **Pareto frontier**

Superhuman Percentile & Subdominance



A **policy** is **γ-superhuman** if it has smaller **metrics** $f_1, f_2, \ldots$ than γ% of **human demonstrations**

Subdominance measures how far a policy is from superhuman by some **margins**

Minimally subdominant policy tends to reside close to the **Pareto frontier**

Subdominance bounds the **superhuman percentile**

Subdominance

$\hat{\mathbf{y}} = \{\hat{y}_j\}_{j=1}^{\mathrm{M}}$

Model Predictions

$\tilde{\mathbf{y}} = \{\tilde{y}_j\}_{j=1}^{\mathrm{M}}$

demonstrations

The minimally subdominant policy:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \min_{\boldsymbol{\alpha} \succeq 0} \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{x} \sim P_{\boldsymbol{\theta}}} \left[\operatorname{subdom}_{\boldsymbol{\alpha}} \left(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a} \right) \right] + \lambda \|\boldsymbol{\alpha}\|_1$$

$\boldsymbol{\theta}$: **Model parameter**

α: **Sensitivity to underperform demonstrations**

If we have metrics inacc, dp, eqodds:

$\text{subdom}_{\alpha} = \alpha_{\text{inacc}}\ \text{subdom}_{\text{inacc}} + \alpha_{\text{dp}}\ \text{subdom}_{\text{dp}} + \alpha_{\text{eqodds}}\ \text{subdom}_{\text{eqodds}}$

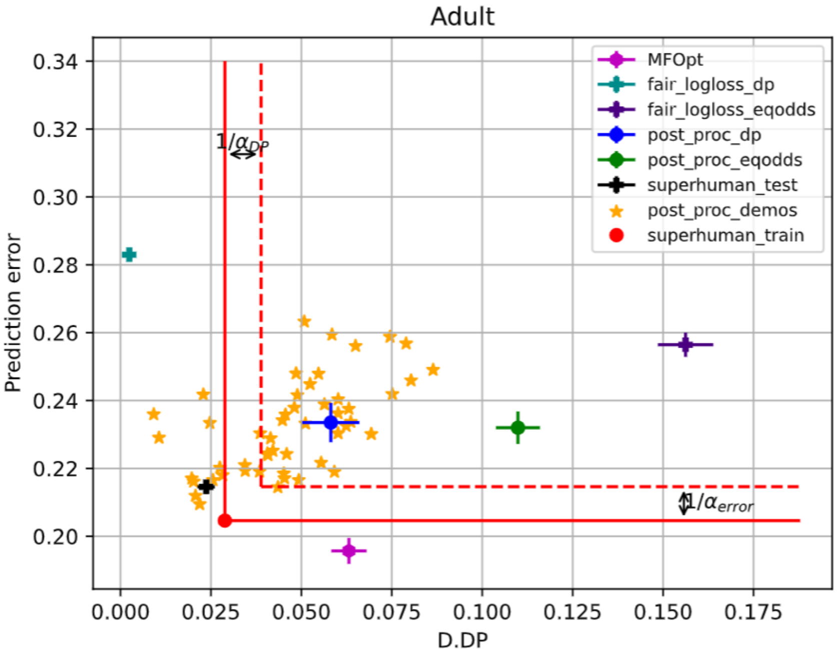
Experiments

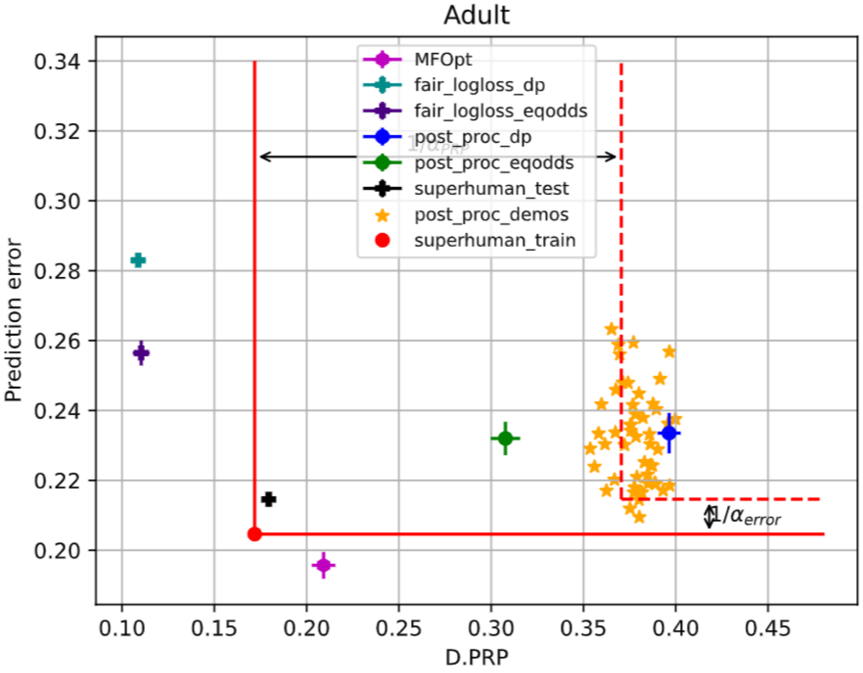
Metrics:

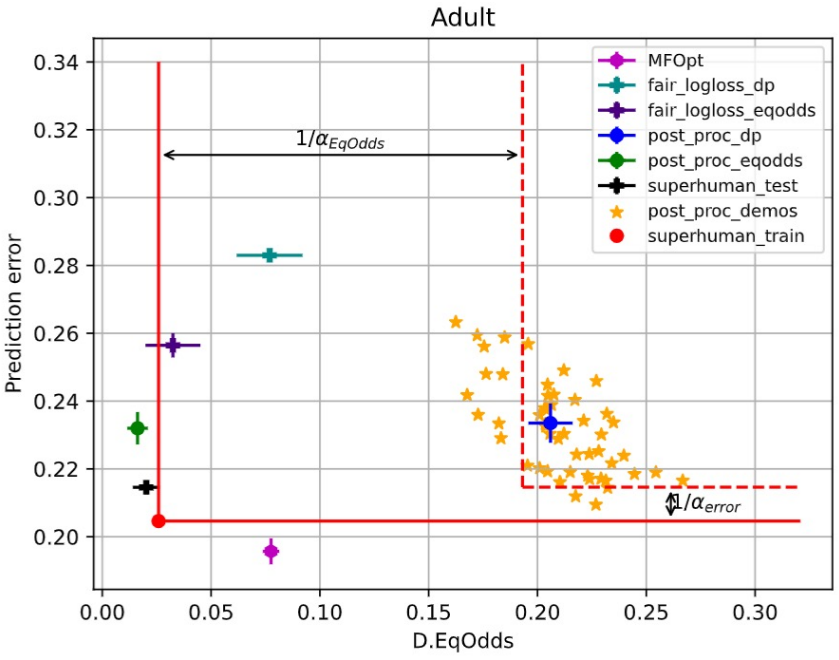
(In)Accuracy (Prediction error)

vs

[DP, EqOdds, PRP]







Thank you!