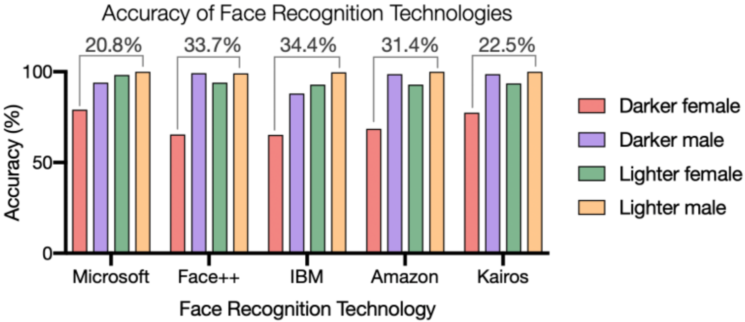
When Statistics Eclipse Fairness

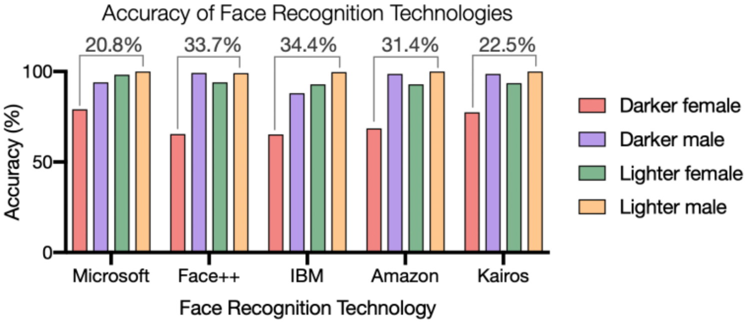
Cyrus Cousins

University of Massachusetts Amherst

The Future of Automated Discrimination is Here!



The Future of Automated Discrimination is Here!



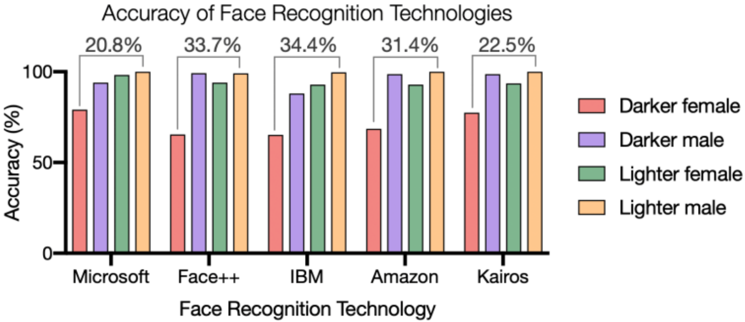
RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

The Future of Automated Discrimination is Here!



RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 5 YEARS AGO

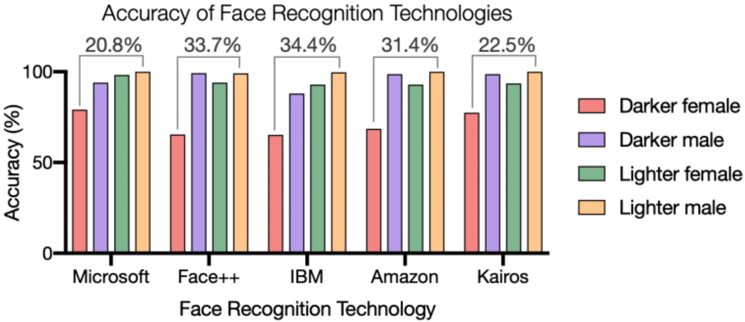
Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

TECH / GOOGLE / ARTIFICIAL INTELLIGENCE

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech / Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

The Future of Automated Discrimination is Here!



RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

TECH / GOOGLE / ARTIFICIAL INTELLIGENCE

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech / Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

How Wrongful Arrests Based on AI Derailed 3 Men's Lives

Robert Williams, Michael Oliver, and Nijeer Parks were misidentified by facial recognition software. The impact cast a long shadow.

Fairness, Discrimination, and Machine Learning

Bias can arise from any step in the machine learning pipeline

- Replicate discrimination in training data
- Data quality, data quantity issues
- Concerns of modeler leak into objective
- Model selection and deployment favor *profit* over social *equity*

Setting the Boundaries

In This Talk

- Assume human impact of model is understood (through the loss function)
  - Work with economists, sociologists
  - Listen to marginalized communities

Setting the Boundaries

In This Talk

- Assume human impact of model is understood (through the loss function)
  - Work with economists, sociologists
  - Listen to marginalized communities
- How do we codify fairness?
  - Compromise between *protected groups*
  - Race, gender, etc.

Setting the Boundaries

In This Talk

- Assume human impact of model is understood (through the loss function)
  - Work with economists, sociologists
  - Listen to marginalized communities
- How do we codify fairness?
  - Compromise between *protected groups*
  - Race, gender, etc.
- Statistical learning theory for fairness
  - Overfitting to fairness

Setting the Boundaries

In This Talk

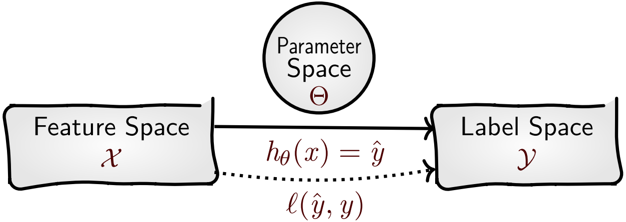
- Assume human impact of model is understood (through the loss function)
  - Work with economists, sociologists
  - Listen to marginalized communities
- How do we codify fairness?
  - Compromise between *protected groups*
  - Race, gender, etc.
- Statistical learning theory for fairness
  - Overfitting to fairness

Professional Theorist

Amateur Humanist

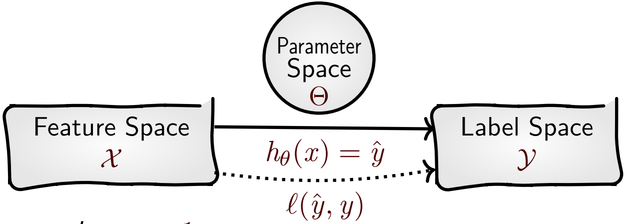
Loss, Data, Machine Learning, and Humanity

- Loss functions are *proxies* for the *impact* a model has on real humans
  - Domain $\mathcal{X}$
  - Codomain $\mathcal{Y}$
  - Parameter space Θ
  - Loss function $\ell(\hat{y}, y)$

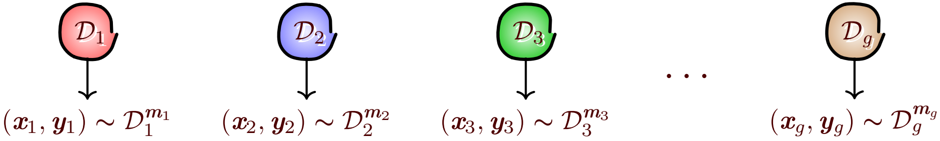


Loss, Data, Machine Learning, and Humanity

- Loss functions are *proxies* for the *impact* a model has on real humans
  - Domain $\mathcal{X}$
  - Codomain $\mathcal{Y}$
  - Parameter space Θ
  - Loss function $\ell(\hat{y}, y)$

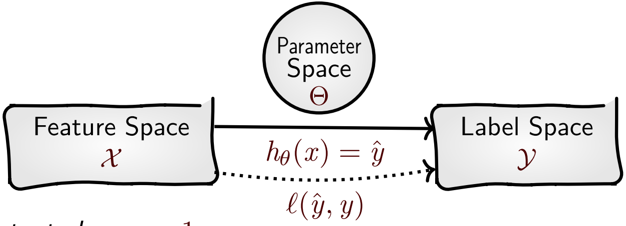


- **Group fairness:** Assume *protected groups* $1, \ldots, g$
  - Distribution $\mathcal{D}_i$ over $\mathcal{X} \times \mathcal{Y}$ captures experiences of each group i
  - Per-group training samples $(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim \mathcal{D}_i^{m_i}$

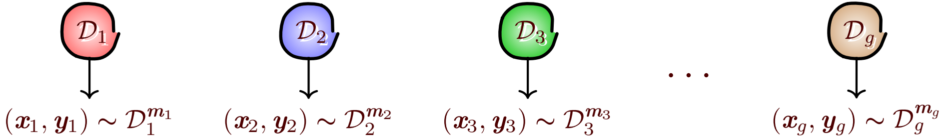


Loss, Data, Machine Learning, and Humanity

- Loss functions are *proxies* for the *impact* a model has on real humans
  - Domain $\mathcal{X}$
  - Codomain $\mathcal{Y}$
  - Parameter space Θ
  - Loss function $\ell(\hat{y}, y)$



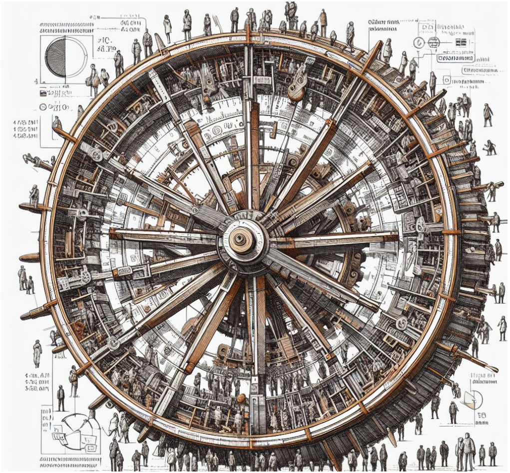
- **Group fairness:** Assume *protected groups* $1, \dots, g$
  - Distribution $\mathcal{D}_i$ over $\mathcal{X} \times \mathcal{Y}$ captures experiences of each group i
  - Per-group training samples $(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim \mathcal{D}_i^{\boldsymbol{m}_i}$



- Summarize model h_θ impact for group i as the *expected risk* or *empirical risk*

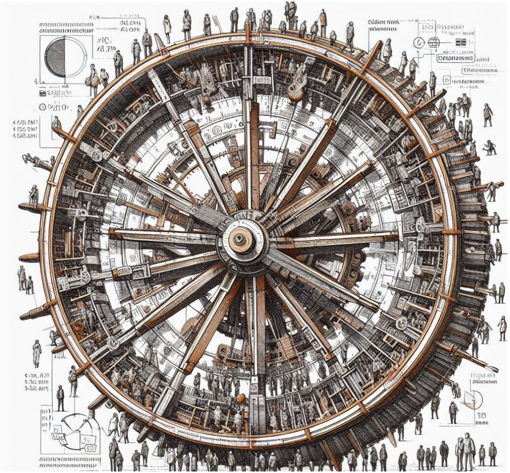
$$\mathrm{R}_i(\theta) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\Big[\ell(h_\theta(x), y)\Big] \quad , \quad \hat{\mathrm{R}}_i(\theta) = \mathop{\widehat{\mathbb{E}}}_{(x,y)\in(\boldsymbol{x}_i,\boldsymbol{y}_i)}\left[\frac{1}{\boldsymbol{m}_i}\sum_{i=1}^{\boldsymbol{m}_i}\ell(h_\theta(x), y)\right]$$

Human-Centric Machine Learning

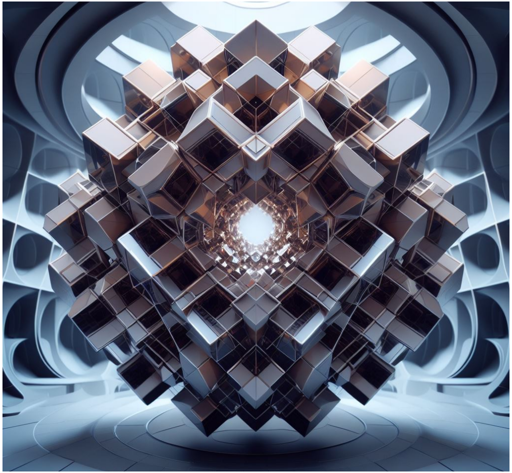


- Center those impacted, not the modeler!
  - Risk $\mathrm{R}_i(\theta)$ is *harm to* group i by model θ
  - Data derived from *impacted humans*, not *decisions about them*

Human-Centric Machine Learning



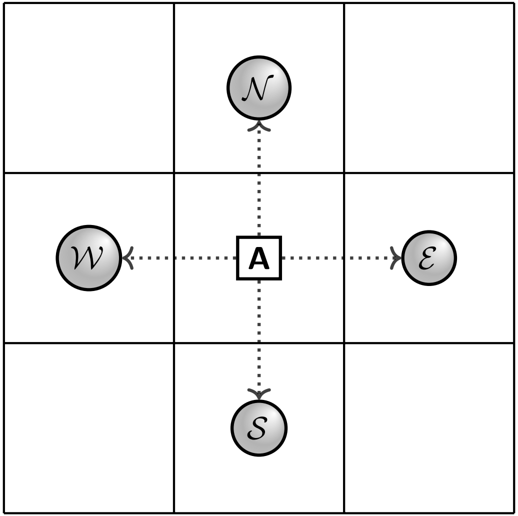
- Center those impacted, not the modeler!
  - Risk $\mathrm{R}_i(\theta)$ is *harm to* group i by model θ
  - Data derived from *impacted humans*, not *decisions about them*



- Contrast with *constraint-based fairness*
  - Primary objective: Given by modeler
  - Secondary concern: Human-centric fairness constraints

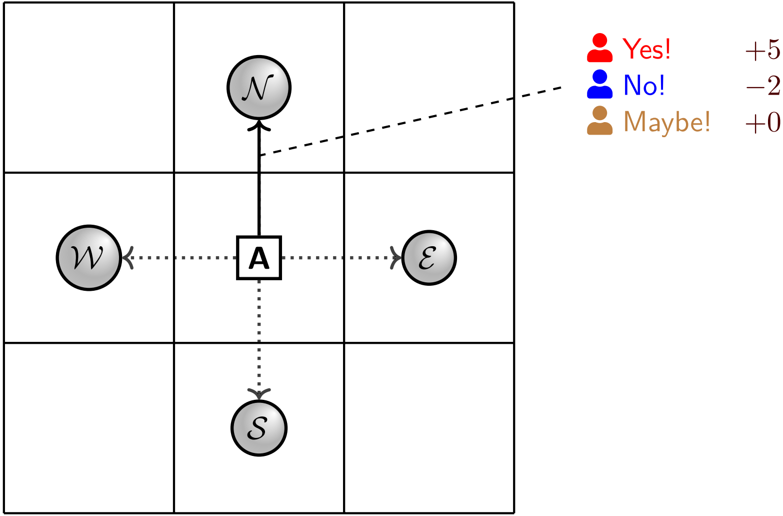
Vignette: Group-Fair Reinforcement Learning

- Agent **A** receives *vector-valued* reward $\boldsymbol{r}(s, a) \in \mathbb{R}^g$ representing all groups



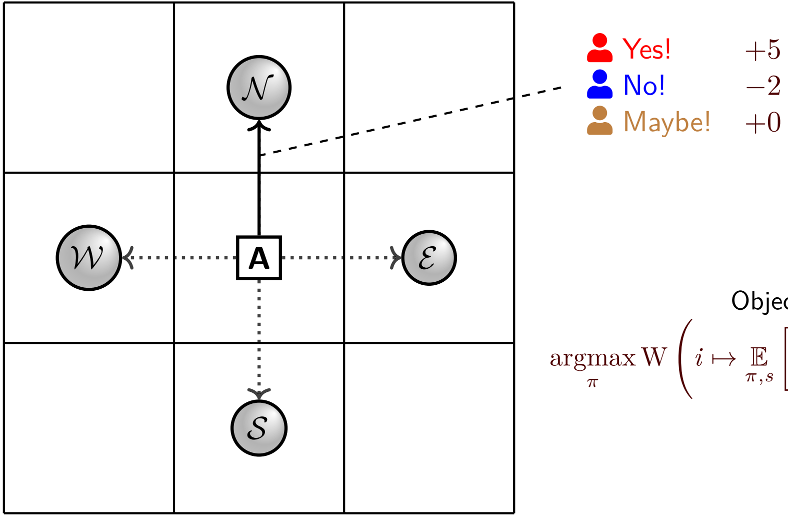
Vignette: Group-Fair Reinforcement Learning

- Agent **A** receives *vector-valued* reward $\boldsymbol{r}(s, a) \in \mathbb{R}^g$ representing all groups



Vignette: Group-Fair Reinforcement Learning

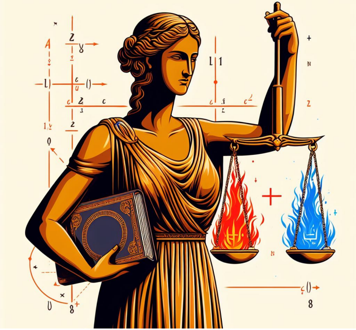
- Agent **A** receives *vector-valued* reward $\boldsymbol{r}(s, a) \in \mathbb{R}^g$ representing all groups
- Optimize not the value of *what I want*, but the *welfare* of value functions



Objective:

$$\operatorname*{argmax}_{\pi} \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\pi, s}\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_i(s_t, \pi(s_t)) \right] \right)$$

Welfare-Centric Fair Machine Learning



- How should we consolidate per-group risk or utility?
  - Studied by moral philosophers and economists
  - Cardinal welfare theory generally treats equitable distribution of *utility* to *individuals*

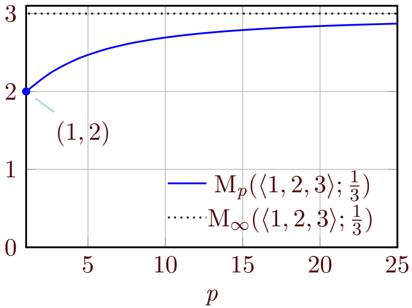
Welfare-Centric Fair Machine Learning

- How should we consolidate per-group risk or utility?
  - Studied by moral philosophers and economists
  - Cardinal welfare theory generally treats equitable distribution of *utility* to *individuals*
- I axiomatically characterize equitable distribution of *disutility* to *weighted groups*
  - Power-mean malfare: Disutility vector $\boldsymbol{\ell} \in \mathbb{R}^g_{0+}$, weights probability vector $\boldsymbol{w} \in \triangle_g$

$$\mathrm{M}_p(\boldsymbol{\ell}; \boldsymbol{w}) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{\ell}_i^p}$$

$$\lim_{p \to \infty} \mathrm{M}_p(\boldsymbol{\ell}; \boldsymbol{w}) = \max_{i \in 1, \ldots, g} \boldsymbol{\ell}_i$$

  - $p \geq 1$ is convex, incentivizes equitable redistribution



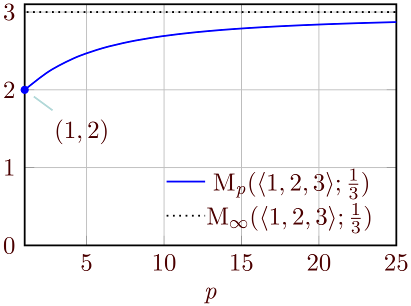
Welfare-Centric Fair Machine Learning

- How should we consolidate per-group risk or utility?
  - Studied by moral philosophers and economists
  - Cardinal welfare theory generally treats equitable distribution of *utility* to *individuals*
- I axiomatically characterize equitable distribution of *disutility* to *weighted groups*
  - Power-mean malfare: Disutility vector $\boldsymbol{\ell} \in \mathbb{R}^g_{0+}$, weights probability vector $\boldsymbol{w} \in \triangle_g$

$$\mathrm{M}_p(\boldsymbol{\ell}; \boldsymbol{w}) = \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \boldsymbol{\ell}_i^p}$$

$$\lim_{p \to \infty} \mathrm{M}_p(\boldsymbol{\ell}; \boldsymbol{w}) = \max_{i \in 1, \ldots, g} \boldsymbol{\ell}_i$$

  - $p \geq 1$ is convex, incentivizes equitable redistribution
- Welfare and malfare encode *social values*
  - Optimizing is *intersubjectively fair* for *shared values*

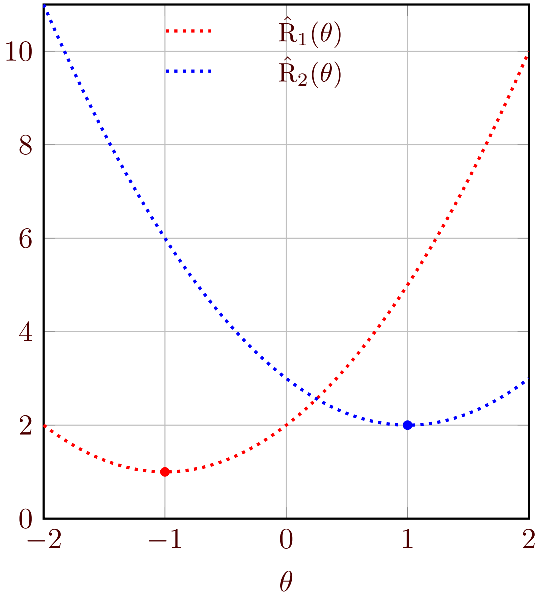


A Generic Fair Machine Learning Algorithm

Empirical Risk Minimization

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \hat{\mathrm{R}}_i(\theta)$$

Univariate Linear Regression



A Generic Fair Machine Learning Algorithm

Empirical Risk Minimization

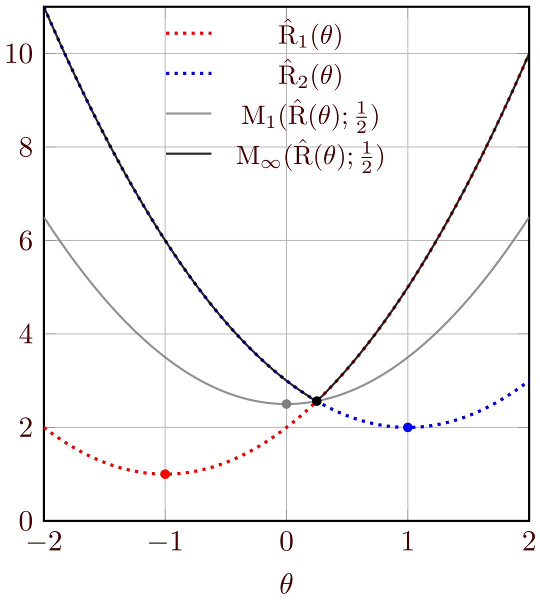
$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \hat{\mathrm{R}}_i(\theta)$$

Empirical Malfare Minimization

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{M}\left(i \mapsto \hat{\mathrm{R}}_i(\theta)\right)$$

- EMM generalizes ERM
  - Convex optimization
  - Intuitive hyperparameter $\mathcal{M}(\cdot)$

Univariate Linear Regression



A Generic Fair Machine Learning Algorithm

Empirical Risk Minimization

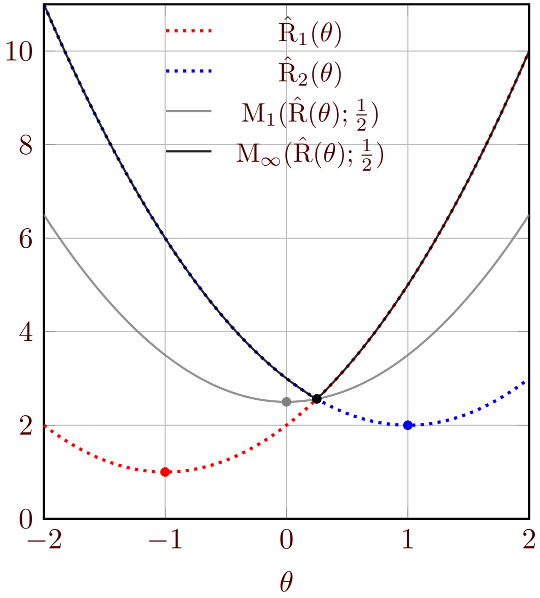
$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \hat{\mathrm{R}}_i(\theta)$$

Empirical Malfare Minimization

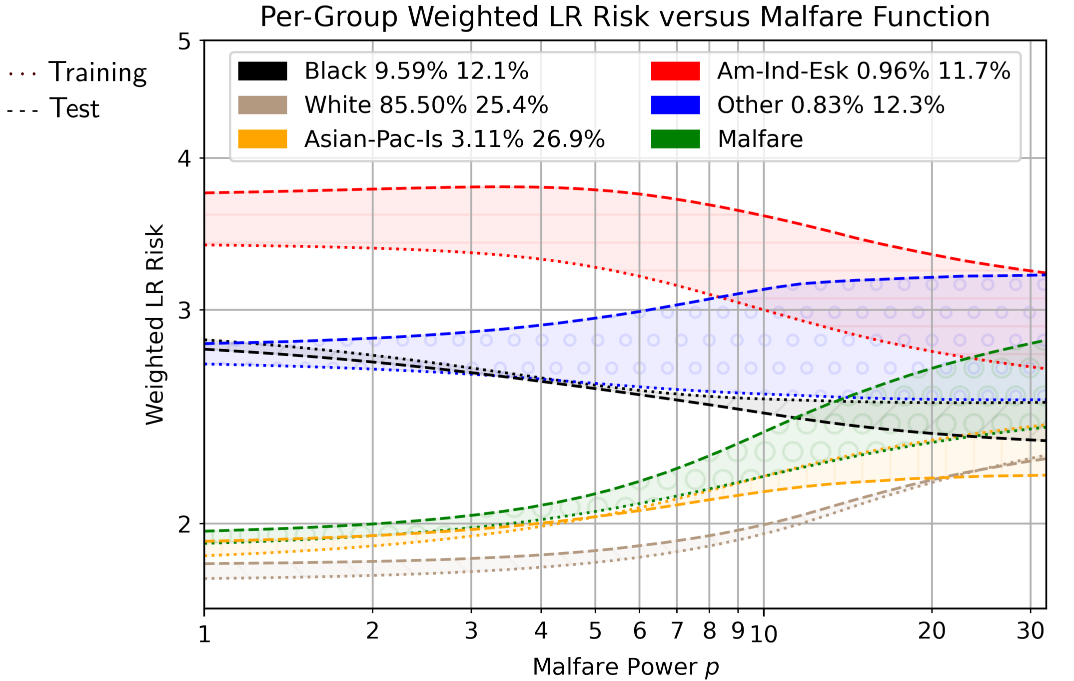
$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \mathrm{M}\left(i \mapsto \hat{\mathrm{R}}_i(\theta)\right)$$

- EMM generalizes ERM
  - Convex optimization
  - Intuitive hyperparameter $\mathrm{M}(\cdot)$
- Interesting special cases
  - $p = \infty$ is *minimax fair learning*
  - $p = 1$ is $\boldsymbol{w}$-weighted risk minimization

Univariate Linear Regression



Fair Logistic Regression on the Adult Dataset



Overfitting to Fairness

- *Rademacher averages* bound *risk* generalization gap
  - Suppose range r loss function, parameter space Θ
  - **Supremum Deviation Bound**: With probability at least $1 - \delta$:

$$\text{For all } \theta \in \Theta\text{:} \quad \left| \mathrm{R}_i(\theta) - \hat{\mathrm{R}}_i(\theta) \right| \leq \boldsymbol{\varepsilon}_i = 2\mathfrak{R}_{\boldsymbol{m}_i}(\ell \circ \Theta, \mathcal{D}_i) + r\sqrt{\frac{\ln \frac{1}{\delta}}{2\boldsymbol{m}_i}}$$

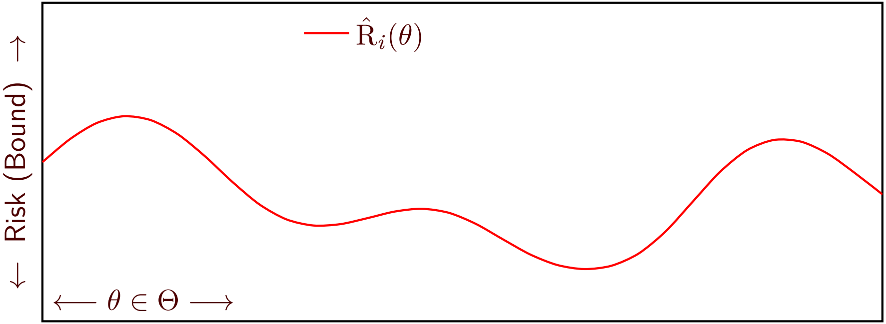
  - *Can overfit more to smaller groups*

Overfitting to Fairness

- *Rademacher averages* bound *risk* generalization gap
  - Suppose range r loss function, parameter space Θ
  - **Supremum Deviation Bound**: With probability at least $1-\delta$:

$$\text{For all } \theta \in \Theta\text{:} \quad \left|\mathrm{R}_i(\theta) - \hat{\mathrm{R}}_i(\theta)\right| \leq \boldsymbol{\varepsilon}_i = 2\mathfrak{R}_{\boldsymbol{m}_i}(\ell \circ \Theta, \mathcal{D}_i) + r\sqrt{\frac{\ln \frac{1}{\delta}}{2\boldsymbol{m}_i}}$$

  - *Can overfit more to smaller groups*

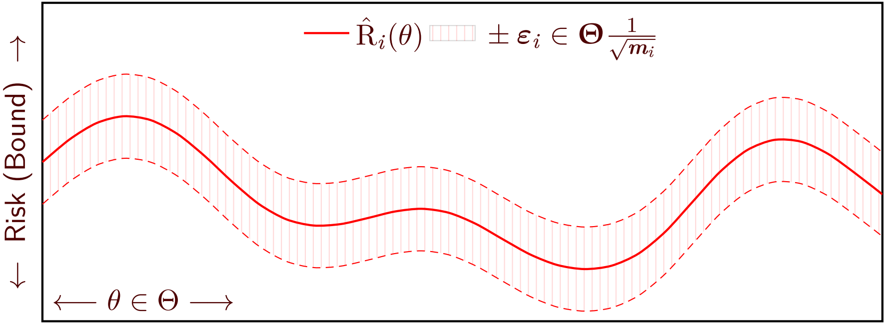


Overfitting to Fairness

- *Rademacher averages* bound *risk* generalization gap
  - Suppose range r loss function, parameter space Θ
  - **Supremum Deviation Bound**: With probability at least $1-\delta$:

$$\text{For all } \theta \in \Theta: \quad \left|\mathrm{R}_i(\theta) - \hat{\mathrm{R}}_i(\theta)\right| \leq \varepsilon_i = 2\mathfrak{R}_{m_i}(\ell \circ \Theta, \mathcal{D}_i) + r\sqrt{\frac{\ln \frac{1}{\delta}}{2m_i}}$$

  - *Can overfit more to smaller groups*

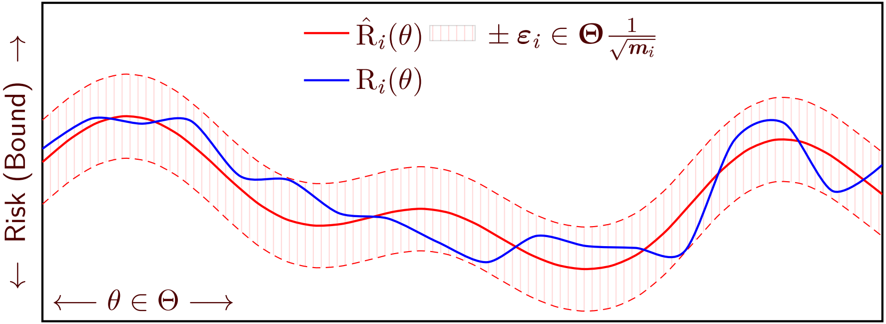


Overfitting to Fairness

- *Rademacher averages* bound *risk* generalization gap
  - Suppose range r loss function, parameter space Θ
  - **Supremum Deviation Bound**: With probability at least $1 - \delta$:

$$\text{For all } \theta \in \Theta\text{:} \quad \left|\mathrm{R}_i(\theta) - \hat{\mathrm{R}}_i(\theta)\right| \leq \boldsymbol{\varepsilon}_i = 2\mathfrak{R}_{\boldsymbol{m}_i}(\ell \circ \Theta, \mathcal{D}_i) + r\sqrt{\frac{\ln \frac{1}{\delta}}{2\boldsymbol{m}_i}}$$

  - *Can overfit more to smaller groups*

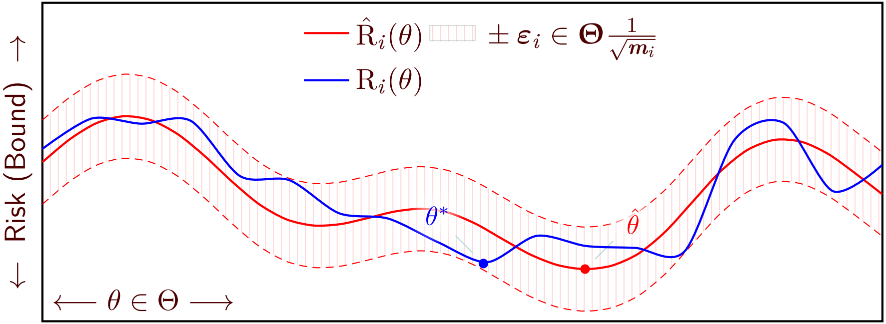


Overfitting to Fairness

- *Rademacher averages* bound *risk* generalization gap
  - Suppose range r loss function, parameter space Θ
  - **Supremum Deviation Bound**: With probability at least $1-\delta$:

$$\text{For all } \theta \in \Theta: \quad \left|\mathrm{R}_i(\theta) - \hat{\mathrm{R}}_i(\theta)\right| \leq \varepsilon_i = 2\mathfrak{R}_{m_i}(\ell \circ \Theta, \mathcal{D}_i) + r\sqrt{\frac{\ln \frac{1}{\delta}}{2m_i}}$$

  - *Can overfit more to smaller groups*

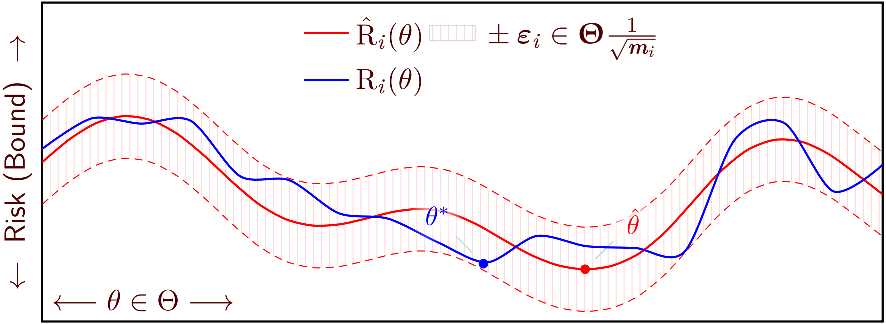


Overfitting to Fairness

- *Rademacher averages* bound *risk* generalization gap
  - Suppose range r loss function, parameter space Θ
  - **Supremum Deviation Bound**: With probability at least $1-\delta$:

$$\text{For all } \theta \in \Theta\text{:} \quad \left|\mathrm{R}_i(\theta) - \hat{\mathrm{R}}_i(\theta)\right| \leq \boldsymbol{\varepsilon}_i = 2\mathfrak{R}_{\boldsymbol{m}_i}(\ell \circ \Theta, \mathcal{D}_i) + r\sqrt{\frac{\ln\frac{1}{\delta}}{2\boldsymbol{m}_i}}$$

  - *Can overfit more to smaller groups*



- I generalize this result to power-mean malfare

$$\text{For all } \theta \in \Theta\text{:} \quad \left|\mathrm{M}_p\left(i \mapsto \mathrm{R}_i(\theta); \boldsymbol{w}\right) - \mathrm{M}_p\left(i \mapsto \hat{\mathrm{R}}_i(\theta); \boldsymbol{w}\right)\right| \leq \max_{i \in 1,\ldots,g} \boldsymbol{\varepsilon}_i$$

Fair-PAC Learning

- Can we learn a **Probably Approximately Correct** (malfare-optimal) model?
  - Sample complexity $\mathrm{m}_{\mathrm{M}}(\varepsilon, \delta)$ is the minimum sufficient sample size such that:
    - For any problem instance (distributions $\mathcal{D}_{1:g}$)
    - With **probability** at least $1 - \delta$
    - Learn model $\hat{\theta}$ that is ε-**approximately** optimal

$$\mathbb{P}\left(\mathrm{M}_p \left(i \mapsto \mathrm{R}_i(\hat{\theta}); \boldsymbol{w} \right) \leq \operatorname*{argmin}_{\theta^* \in \Theta} \mathrm{M}_p \left(i \mapsto \mathrm{R}_i(\theta^*); \boldsymbol{w} \right) + \varepsilon \right) \geq 1 - \delta$$

Fair-PAC Learning

- Can we learn a **Probably Approximately Correct** (malfare-optimal) model?
  - Sample complexity $\mathrm{m}_{\mathcal{M}}(\varepsilon, \delta)$ is the minimum sufficient sample size such that:
    - For any problem instance (distributions $\mathcal{D}_{1:g}$)
    - With **probability** at least $1 - \delta$
    - Learn model $\hat{\theta}$ that is ε-**approximately** optimal

$$\mathbb{P}\left(\mathcal{M}_p\left(i \mapsto \mathrm{R}_i(\hat{\theta}); \boldsymbol{w}\right) \leq \operatorname*{argmin}_{\theta^* \in \Theta} \mathcal{M}_p\left(i \mapsto \mathrm{R}_i(\theta^*); \boldsymbol{w}\right) + \varepsilon\right) \geq 1 - \delta$$

- Power-mean malfare is a *contraction function* (1-Lipschitz)

$$\left|\mathcal{M}_p\left(i \mapsto \hat{\mathrm{R}}_i(\theta); \boldsymbol{w}\right) - \mathcal{M}_p\left(i \mapsto \mathrm{R}_i(\theta); \boldsymbol{w}\right)\right| \leq \left\|i \mapsto \hat{\mathrm{R}}_i(\theta) - \mathrm{R}_i(\theta)\right\|_\infty$$

  - Comparable sample complexity (per-group) to PAC learning: $\mathrm{m}_{\mathcal{M}}(\varepsilon, \delta) \leq \mathrm{m}_{\mathrm{R}}\left(\varepsilon, \frac{\delta}{g}\right)$
  - ε-$\frac{\delta}{g}$ SD bound for all groups suffices for EMM to fair-PAC learn Θ

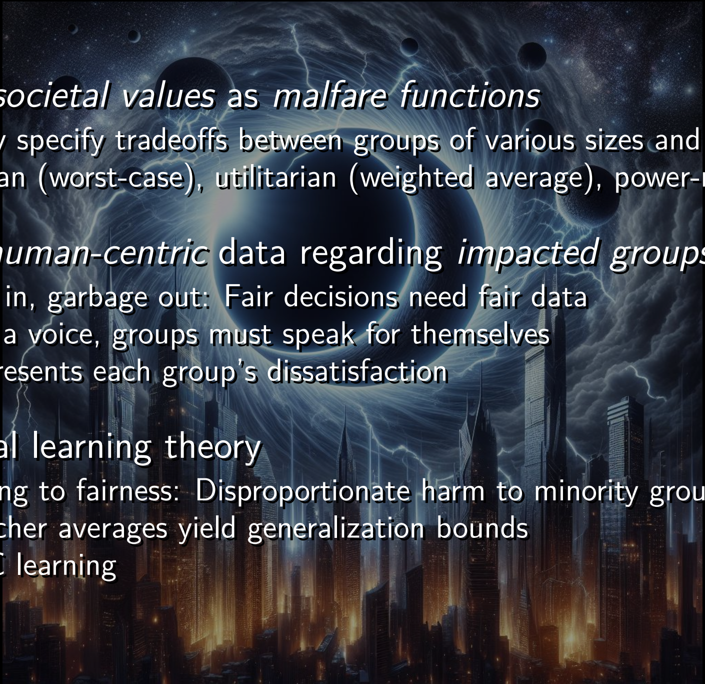
Fair Learning Overview

- Encode *societal values* as *malfare functions*
  - Implicitly specify tradeoffs between groups of various sizes and risk levels
  - Egalitarian (worst-case), utilitarian (weighted average), power-means

Fair Learning Overview

- Encode *societal values* as *malfare functions*
  - Implicitly specify tradeoffs between groups of various sizes and risk levels
  - Egalitarian (worst-case), utilitarian (weighted average), power-means
- Collect *human-centric* data regarding *impacted groups*
  - Garbage in, garbage out: Fair decisions need fair data
  - To have a voice, groups must speak for themselves
  - Risk represents each group's dissatisfaction

Fair Learning Overview



- Encode *societal values* as *malfare functions*
  - Implicitly specify tradeoffs between groups of various sizes and risk levels
  - Egalitarian (worst-case), utilitarian (weighted average), power-means
- Collect *human-centric* data regarding *impacted groups*
  - Garbage in, garbage out: Fair decisions need fair data
  - To have a voice, groups must speak for themselves
  - Risk represents each group's dissatisfaction
- Statistical learning theory
  - Overfitting to fairness: Disproportionate harm to minority groups
  - Rademacher averages yield generalization bounds
  - Fair-PAC learning