Learning to Walk Impartially on the Pareto Frontier of Fairness, Privacy, and Utility

Mohammad Yaghini, Patty Liu, Franziska Boenisch, Nicolas Papernot

University of Toronto & Vector Institute

Motivation: Specification Problem

- We want to deploy a model for chest X-rays to all regional hospitals
- Building the model is outsourced to a private company
- Regulators have privacy and fairness concerns
- They want to specify acceptable levels of trustworthiness guarantees for the model

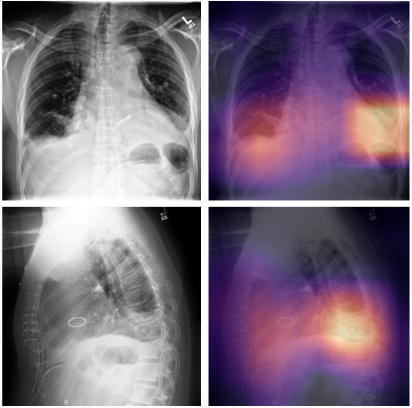


Figure 1: CheXpert

Motivation: Specification Problem

- We want to deploy a model for chest X-rays to all regional hospitals
- Building the model is outsourced to a private company
- Regulators have privacy and fairness concerns
- They want to specify acceptable levels of trustworthiness guarantees for the model

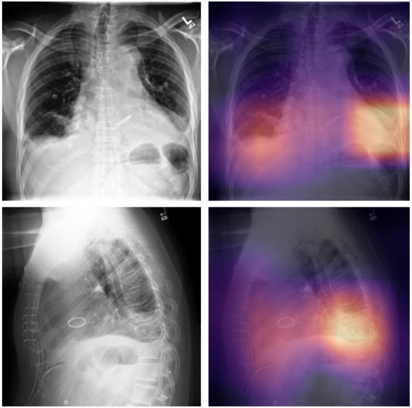


Figure 1: CheXpert

What are trustworthiness objectives for ML?

Values	Objective Examples	Mechanisms
Utility	Accuracy	Architecture search, optimizer search, etc.
Privacy	Differential Privacy (DP Loss) Unlearning	DP mechanisms: Noising, Randomized Response, etc.
Fairness	Demographic Parity (DemParity) Equality of Odds Disparate Impact	DemParity processors and regularizers

and many more (interpretability, robustness to distribution shifts/adversarial examples, etc.)

What are trustworthiness objectives for ML?

Values	Objective Examples	Mechanisms
Utility	Accuracy	Architecture search, optimizer search, etc.
Privacy	Differential Privacy (DP Loss) Unlearning	DP mechanisms: Noising, Randomized Response, etc.
Fairness	Demographic Parity (DemParity) Equality of Odds Disparate Impact	DemParity processors and regularizers

and many more (interpretability, robustness to distribution shifts/adversarial examples, etc.)

Privacy Objective: Differential Privacy

Definition ((ε, δ)-Differential Privacy)

Let $\mathcal{M} \colon \mathcal{D}^* \to \mathcal{R}$ be a randomized algorithm that satisfies (ε, δ)-DP with $\varepsilon \in \mathbb{R}_+$ and $\delta \in [0, 1]$ if for all neighboring datasets $D \sim D'$, and for all possible subsets $R \subseteq \mathcal{R}$ of the result space $\mathcal{M}$ satsifies

$$\mathbb{P}\left[\mathcal{M}(D) \in R\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(D') \in R\right] + \delta$$

Fairness Objective: Demographic Parity

Definition (Demographic Disparity)

$$\Gamma_{\text{DemParity}}(k, z) = \mathbb{P}[\hat{Y} = k' \mid Z = z] - \mathbb{P}[\hat{Y} = k' \mid Z \neq z]$$

where $\hat{Y} = \omega(\mathbf{x}, z)$ are model $\omega : \mathcal{X} \times \mathcal{Z} \mapsto \mathcal{K}$ predictions for samples with sensitive attribute z.

Definition (γ-disparity)

$\forall z \in \mathcal{Z}, \forall k \in \mathcal{K}$,

$$\Gamma_{\text{DemParity}}(k, z) \leq \gamma$$

ML trustworthiness as multi-objective optimization

$$\begin{aligned} \min_{\omega} \quad & \ell_{\mathsf{acc}}(\omega) \\ \text{subject to} \quad & \ell_{\mathsf{priv}}(\omega) \leq \varepsilon \\ & \ell_{\mathsf{fair}}(\omega) \leq \gamma \end{aligned} \tag{1}$$

where $(\ell_{\mathsf{acc}}, \ell_{\mathsf{priv}}, \ell_{\mathsf{fair}}) \in \mathbb{R}^3_{\geq 0}$ are the loss functions for each of the utility, privacy and fairness criteria, respectively.

ML trustworthiness as multi-objective optimization

$$\begin{array}{cc} \min_{\omega} & \ell_{\mathsf{acc}}(\omega) \\ \text{subject to} & \ell_{\mathsf{priv}}(\omega) \leq \varepsilon \\ & \ell_{\mathsf{fair}}(\omega) \leq \gamma \end{array} \tag{1}$$

where $(\ell_{\mathsf{acc}}, \ell_{\mathsf{priv}}, \ell_{\mathsf{fair}}) \in \mathbb{R}^3_{\geq 0}$ are the loss functions for each of the utility, privacy and fairness criteria, respectively.

- Problem 1: Privacy is ensured at the level of mechanism (here, the ML pipeline)

ML trustworthiness as multi-objective optimization

$$
\begin{aligned}
\min_{\omega} \quad & \ell_{\text{acc}}(\omega) \\
\text{subject to} \quad & \cancel{\ell_{\text{priv}}(\omega) \le \varepsilon} \\
& \ell_{\text{fair}}(\omega) \le \gamma
\end{aligned}
\tag{1}
$$

where $(\ell_{\text{acc}}, \ell_{\text{priv}}, \ell_{\text{fair}}) \in \mathbb{R}^3_{\geq 0}$ are the loss functions for each of the utility, privacy and fairness criteria, respectively.

- Problem 1: Privacy is ensured at the level of mechanism (here, the ML pipeline) $\Rightarrow$ We do not have a sample-based privacy loss

ML trustworthiness as multi-objective optimization

Treated as hyper-parameters

$$\begin{aligned} \min_{\omega} \quad & \ell_{\text{acc}}(\omega) \\ \text{subject to} \quad & \cancel{\ell_{\text{priv}}(\omega) \leq \varepsilon} \\ & \ell_{\text{fair}}(\omega) \leq \gamma \end{aligned} \tag{1}$$

where $(\ell_{\text{acc}}, \ell_{\text{priv}}, \ell_{\text{fair}}) \in \mathbb{R}^3_{\geq 0}$ are the loss functions for each of the utility, privacy and fairness criteria, respectively.

- Problem 1: Privacy is ensured at the level of mechanism (here, the ML pipeline) $\Rightarrow$ We do not have a sample-based privacy loss
- Problem 2: Trustworthy parameters are treated as hyper-parameters, not first-class objectives

ML trustworthiness as multi-objective optimization

Treated as hyper-parameters

$$\begin{aligned} \min_\omega \quad & \ell_{\text{acc}}(\omega) \\ \text{subject to} \quad & \cancel{\ell_{\text{priv}}(\omega) \leq \varepsilon} \\ & \ell_{\text{fair}}(\omega) \leq \gamma \end{aligned} \tag{1}$$

where $(\ell_{\text{acc}}, \ell_{\text{priv}}, \ell_{\text{fair}}) \in \mathbb{R}^3_{\geq 0}$ are the loss functions for each of the utility, privacy and fairness criteria, respectively.

- Problem 1: Privacy is ensured at the level of mechanism (here, the ML pipeline) $\Rightarrow$ We do not have a sample-based privacy loss
- Problem 2: Trustworthy parameters are treated as hyper-parameters, not first-class objectives $\Rightarrow$ Pre-Selection Bias

Pre-selection Bias

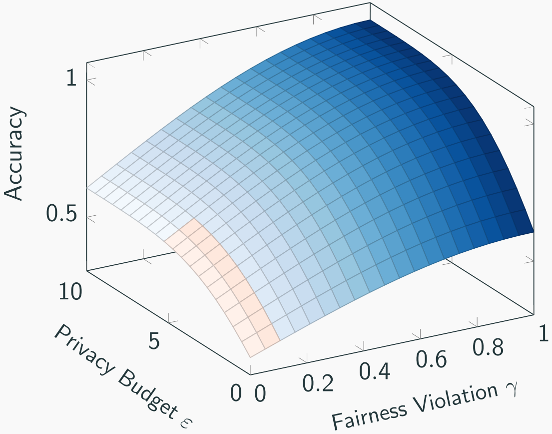
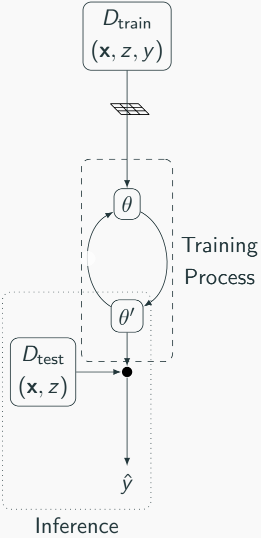
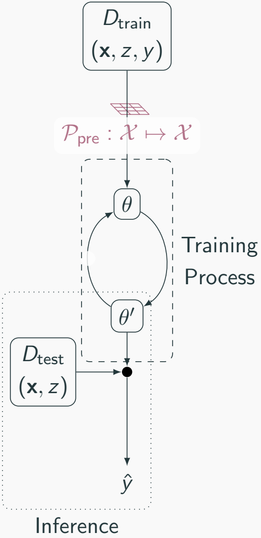


Figure 2: Pre-selection of trustworthiness parameters only recovers a portion of the Pareto frontier. The remaining parts of frontier (shaded blue) are never explored.

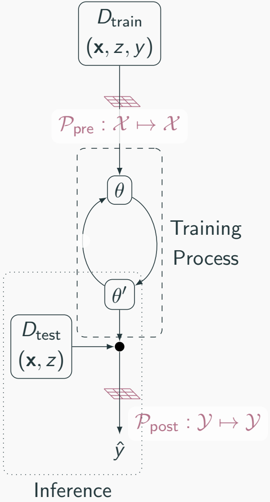
DP Learning to Fair DP Learning: Fairness Intervention



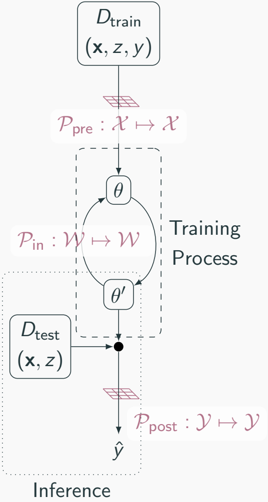
DP Learning to Fair DP Learning: Fairness Intervention



DP Learning to Fair DP Learning: Fairness Intervention



DP Learning to Fair DP Learning: Fairness Intervention

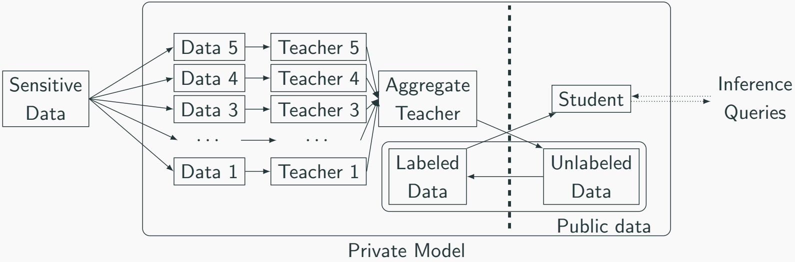


Fairness pre-processing increases the cost of private training

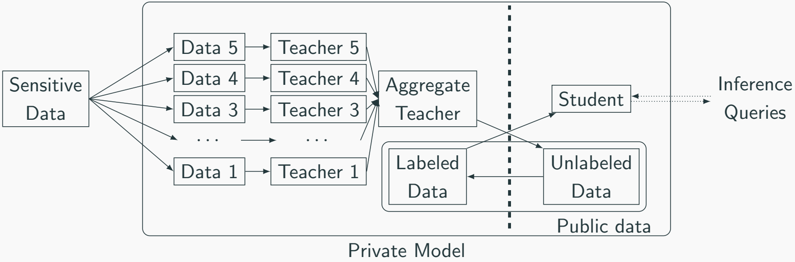
Theorem

Assume the training dataset $D = \{(\mathbf{x}, z, y) \mid \mathbf{x} \in \mathcal{X}, z \in \mathcal{Z}, y \in \mathcal{Y}\}$ is fed through the demographic parity pre-processor $\mathcal{P}_{pre}$ following an ordering defined over the input space $\mathcal{X}$. Let $\mathcal{P}_{pre}$ enforce a maximum violation γ, and $|Z| = 2$. Suppose now $\mathcal{M}$ is an (ε, δ) training mechanism, then $\mathcal{M} \circ \mathcal{P}_{pre}$ is $(K_\gamma \varepsilon, K_\gamma e^{K_\gamma \varepsilon} \delta)$-DP where $K_\gamma = 2 + \left\lceil \frac{2\gamma}{1-\gamma} \right\rceil$.

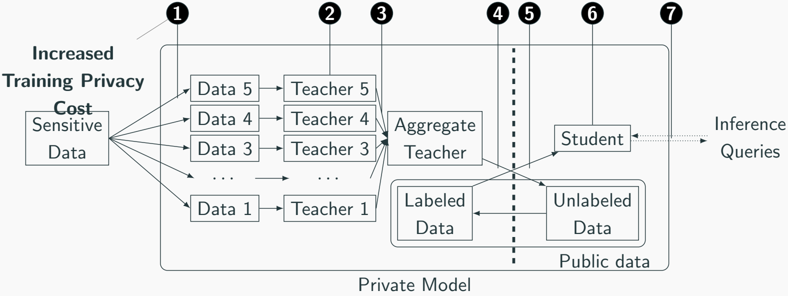
Example: PATE



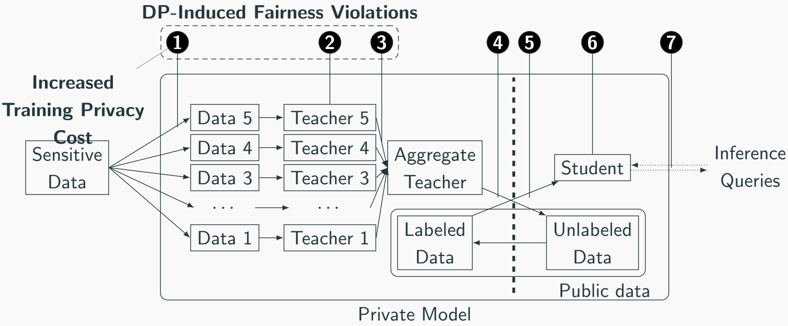
Example: PATE



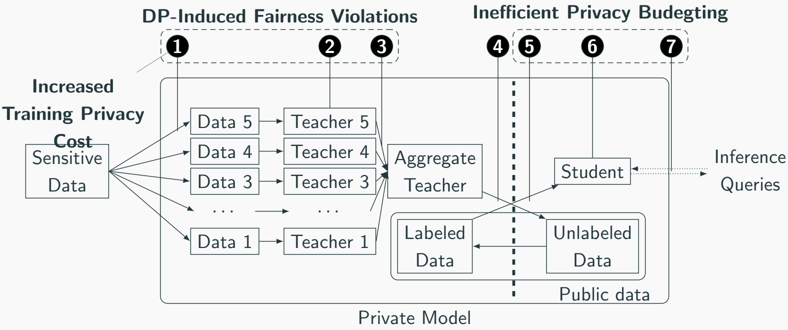
Example: PATE



Example: PATE



Example: PATE



Algorithm 1 Confident-GNMax Aggregator

Input: query data point x, sensitive attribute z, predicted class label k, subpopulation subclass counts $m : \mathcal{Z} \times \mathcal{K} \mapsto \mathbb{Z}_{\geq 0}$

Require: minimum count M, threshold T, noise parameters σ_1, σ_2, fairness violation margin γ

```
if max_j{n_j(x)} + N(0, σ_1^2) ≥ T then
    k ← arg max_j {n_j(x) + N(0, σ_2^2)}
    return k
else
    return ⊥
```

Algorithm 2 Confident&Fair-GNMax Aggregator

Input: query data point x, sensitive attribute z, predicted class label k, subpopulation subclass counts $m : \mathcal{Z} \times \mathcal{K} \mapsto \mathbb{Z}_{\geq 0}$

Require: minimum count M, threshold T, noise parameters σ_1, σ_2, fairness violation margin γ

if $\max_j\{n_j(x)\} + \mathcal{N}(0, \sigma_1^2) \geq T$ **then**

  $k \leftarrow \arg\max_j \left\{n_j(x) + \mathcal{N}(0, \sigma_2^2)\right\}$

  if $\sum_{\tilde{k}} m(z, \tilde{k}) < M$ **then**

    $m(z, k) \leftarrow m(z, k) + 1$

    return k

else

  return $\perp$

Algorithm 2 Confident&Fair-GNMax Aggregator

Input: query data point x, sensitive attribute z, predicted class label k, subpopulation subclass counts $m : \mathcal{Z} \times \mathcal{K} \mapsto \mathbb{Z}_{\geq 0}$

Require: minimum count M, threshold T, noise parameters σ_1, σ_2, fairness violation margin γ

if $\max_j\{n_j(x)\} + \mathcal{N}(0, \sigma_1^2) \geq T$ **then**

  $k \leftarrow \arg\max_j \left\{n_j(x) + \mathcal{N}(0, \sigma_2^2)\right\}$

  if $\sum_{\tilde{k}} m(z, \tilde{k}) < M$ **then**

    $m(z, k) \leftarrow m(z, k) + 1$

    return k

  else

    if $\left(\frac{m(z,k)+1}{\left(\sum_{\tilde{k}} m(z,\tilde{k})\right)+1} - \frac{\sum_{\tilde{z}\neq z} m(\tilde{z},k)}{\sum_{\tilde{z}\neq z,\tilde{k}} m(\tilde{z},\tilde{k})}\right) < \gamma$ **then**

      $m(z, k) \leftarrow m(z, k) + 1$

      return k

    else

      return $\perp$

else

  return $\perp$

Closing the fairness gap with Reject-Option for Fairness

- Optimizing for fairness during the training process does not guarantee that fairness is obtained at inference time
- What if there were a **hard constraint** on fairness violations at inference time?
- A **reject-option** allows to refuse to answer a query at inference time for fairness purposes.
- Introduces a new utility dimension:
  $\textbf{Coverage} := \frac{\textbf{\# Queries Answered}}{\textbf{\# Queries}}$

Algorithm 5 Inference-time Demographic Parity Post-Processor (IDP^3)

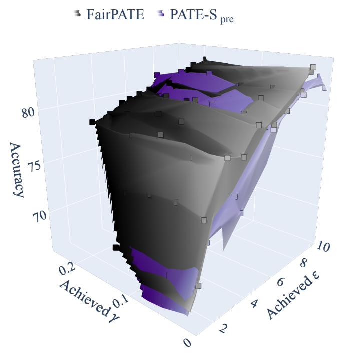
Input: data point x, sensitive attribute z, predicted label $\hat{y}$, subpopulation-class counts $m : \mathcal{Z} \times \mathcal{Y} \mapsto \mathbb{Z}_{\geq 0}$

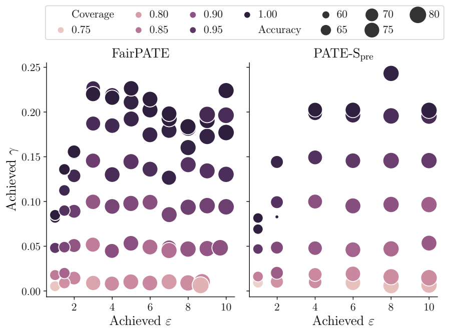
Require: minimum count M, fairness violation margin γ

if $\sum_{\tilde{y}} m(z, \tilde{y}) < M$ **then**
    $m(z, y) \leftarrow m(z, \hat{y}) + 1$
    return $\hat{y}$
else
    if $\left(\frac{m(z,\hat{y})+1}{\left(\sum_{\tilde{y}} m(z,\tilde{y})\right)+1} - \frac{\sum_{\tilde{z} \neq z} m(\tilde{z},\hat{y})}{\sum_{\tilde{z} \neq z, \tilde{y}} m(\tilde{z},\tilde{y})} \right) < \gamma$ **then**
        $m(z, y) \leftarrow m(z, \hat{y}) + 1$
        return $\hat{y}$
    else
        return $\perp$

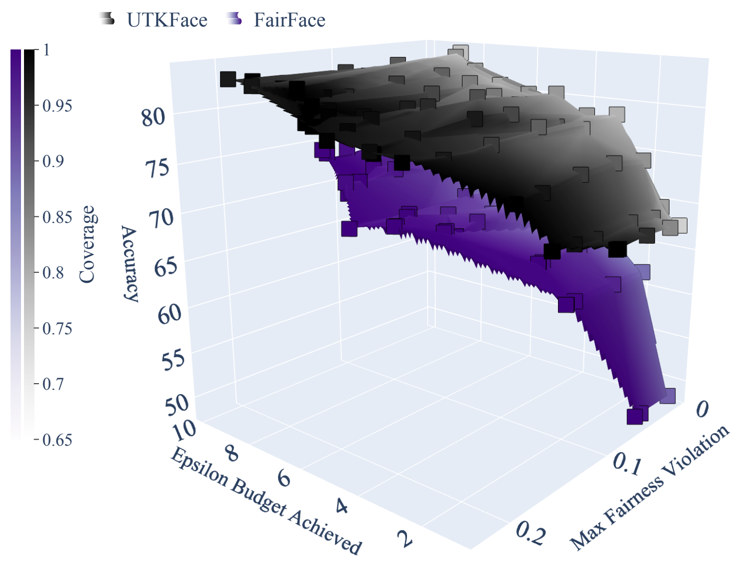
Results

FairPATE Pareto-dominates similar designs in most contexts





Specification without direct data access is possible



Conclusion: Specification requires objective-impartiality!

Pre-Selection Bias → Pareto Inefficiencies → Non-Impartial Designs → Inaccurate Specification

Conclusion: Specification requires objective-impartiality!

Conclusion: Specification requires objective-impartiality!

Conclusion: Specification requires objective-impartiality!

Conclusion: Specification requires objective-impartiality!

Thank you!