CS5562: Trustworthy Machine Learning

Part III Lecture 2: Fairness → Satisfying Fairness Criteria

Reza Shokri[a]

Aug 2023

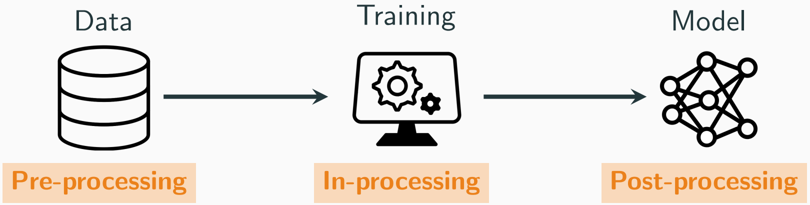
[a]Acknowledgment. The wonderful teaching assistants: Hongyan Chang, Martin Strobel, Jiashu Tao, Yao Tong, Jiayuan Ye

Contents

How to achieve group fairness

Unfairness mitigation

Addressing bias can be categorized

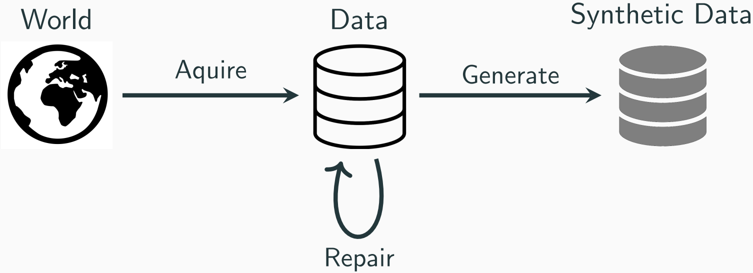


Source: [Bellamy et al., 2019]

Pre-processing: Preparing Unbiased data

Main idea: Fix the problem before training a model

Pro Potentially remove the root source of bias



Source: For overview and additional references see Ding [2021]

Generating fair representations

Can we release representations of the original data that follow certain fairness criteria?

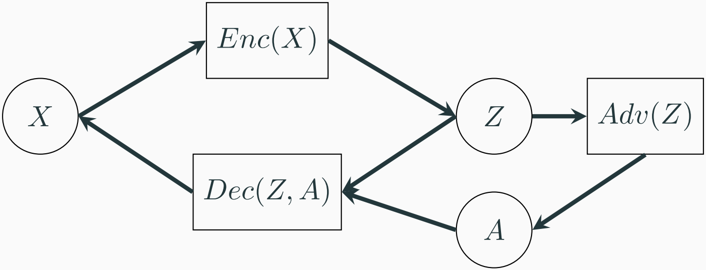
Goal Generate **useful** representations for arbitrary downstream tasks that also satisfy fairness constraints.

Approach Adversarial training with adversary that tries to guess sensitive attribute, penalizes unfair representations.

Guarantees Find best representations of the data such that the fairness constraints will be satisfied by an arbitrary downstream classifier trained on these representations.

Source: [Madras et al., 2018]

Adversarial Framing



Adversary $Adv(Z)$ tries to maximize unfairness & reconstruct A from Z

Encoder $Enc(X)$ is forced to maximize the utility of Z to $Dec(Z, A)$ while minimizing adversary's ability to reconstruct A

Source: [Madras et al., 2018]

Formalization

Objective:

$$\min_{Enc,Dec} \max_{Adv} \mathbb{E}_{X,Y,A}[L(Enc, Dec, Adv)]$$

$$L(Enc, Dec, Adv) = \alpha L_{Dec}(Enc(X), A, Y) + \beta L_{Adv}(Adv(Enc(X)), A)$$

Translating Fairness to Adversarial Objective

We can use any of our fairness notions for the adversarial loss. Here we consider demographic parity ($Z \perp A$):

$$L_{Adv}^{DP}(h) = \sum_{i \in \{0,1\}} \frac{1}{|S_i|} \sum_{(x,y,a) \in S_i} |Adv(Enc(x)) - a|$$

$$(S_i = \{(x, y, i) \in S\})$$

Source: [Madras et al., 2018]

Setting

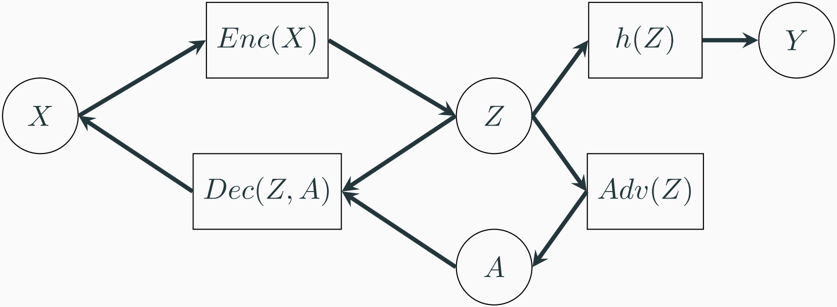
How could our representations be used? Consider two scenarios:

1. An **indifferent user** who doesn't care about A and simply wants the best utility regardless of fairness.
2. A **malicious user** who wants to discriminate against a certain group regardless of utility.

So far
We prevent malicious users from reconstructing A. But can we also optimize utility for the indifferent user?

Source: [Madras et al., 2018]

Preserving Utility



Add a classifier $h(Z)$ and an additional loss term to our pre-processing:

$$
\begin{aligned}
L(Enc, Dec, Adv, h) = &\alpha L_{Dec}(Enc(X), A, Y) + \beta L_{Adv}(Adv(Enc(X)), A) \\
&+ \gamma L_h(h(Enc(X)), Y)
\end{aligned}
$$

Guarantees

Definition
For $Z_0 = \{Enc(x) | (x, y, 0) \in S\}, Z_1 = \{Enc(x) | (x, y, 1) \in S\}$ we define violation of demographic parity by h:

$$\Delta_{DP}(h) \triangleq |\mathbb{E}_{Z_0}[h] - \mathbb{E}_{Z_1}[h]|$$

- If $h(Z)$ has large Δ_{DP}, its output will be correlated with A and an adversary Adv using only $h(Z)$ can partially reconstruct A
- Formally, we can show that $\Delta_{DP} = 1 - L_{Adv}$
- For an optimal adversary Adv^*: $L_{Adv^*} < L_{Adv} \Rightarrow \Delta_{DP} \leq 1 - L_{Adv^*}$
- Therefore by bounding the performance of an optimal adversary, we can minimize Δ_{DP} for any classifier trained on Z

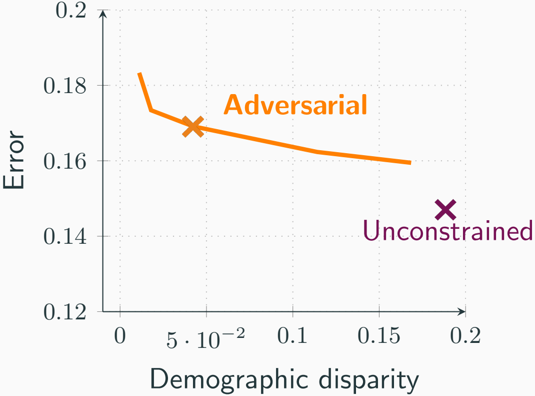
Source: Thm. 1, [Madras et al., 2018]

Experimental results

Training a feed forward MLP on the Adult dataset with SGD

Setup:

1. Learn encoder f
2. Train a classifier g (without fairness constraints) on encoded new data
3. Evaluate fairness of g on holdout test set



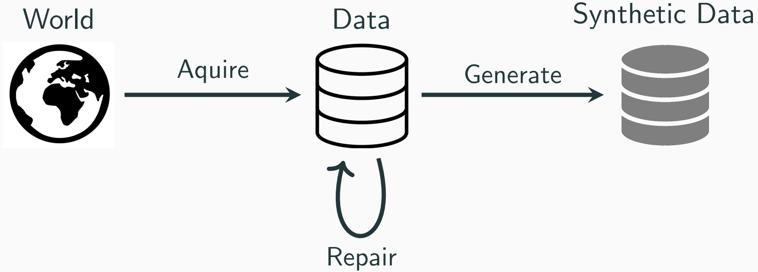
Changing the importance of individual loss functions (via α, β, γ) allows **different error-fairness tradeoffs**.

Pre-processing: Preparing Unbiased data

Main idea: Fix the problem before training a model

Pro Potentially remove the root source of bias

Con No control over dataset after release

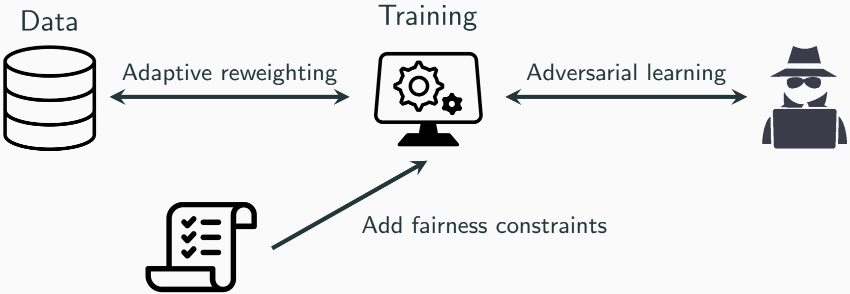


In-processing: Training an unbiased model with biased data

Main idea: Ensure fairness during training

Pro Potential to optimize for performance as well as fairness

Con Requires access and potentially large changes to the training process

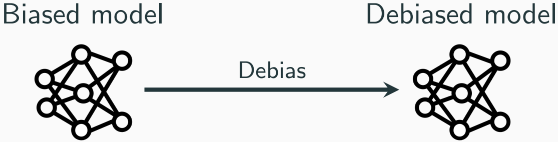


Post-processing: Fixing a biased model

Main idea: Adapt model output to ensure fairness

Pro No access to training data or training process required

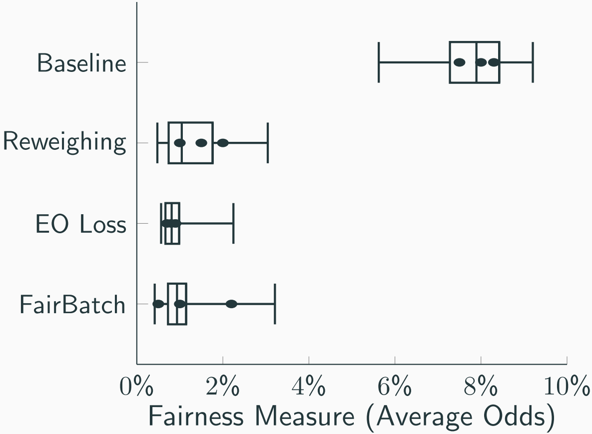
(Appears in Assignment 5)



Limits of group fairness

Fairness Measures Aren't Stable!

Model fairness can vary significantly due to the training randomness (i.e., in every run the fairness gap can change).



Source: [Amir et al., 2021, Sellam et al., 2021, Baldini et al., 2021, Ganesh et al., 2023]

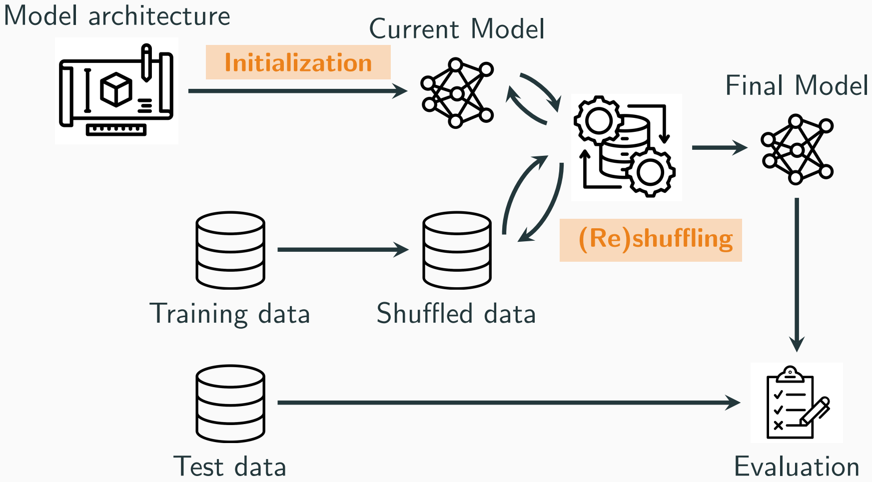
Standard Solutions

Executing multiple training runs with changing random seeds to capture overall fairness variance.

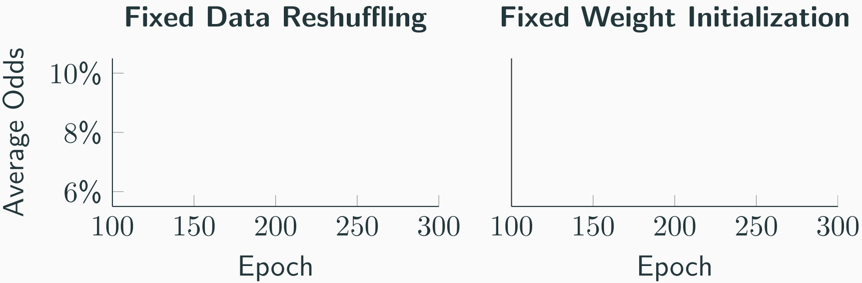
Blindly executing training runs

- is expensive,
- raises the bar to do fair ML research,
- **lacks the understanding of the underlying cause for high fairness variance.**

Weight Initialization and Data Reshuffling

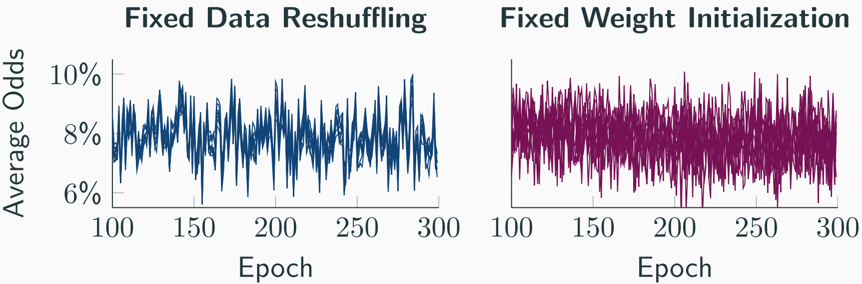


Variance Across Epochs

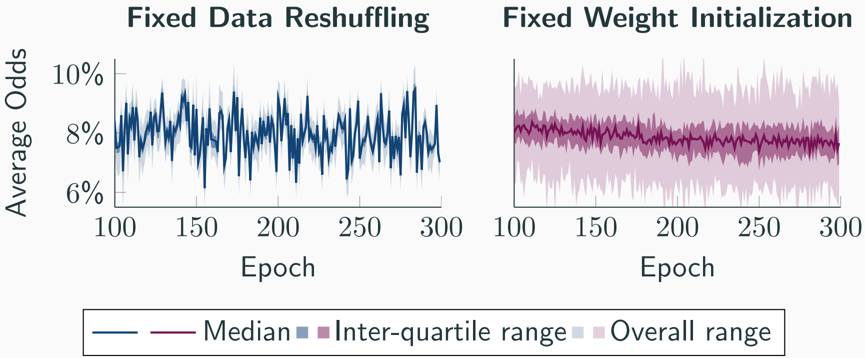


Variance Across Epochs

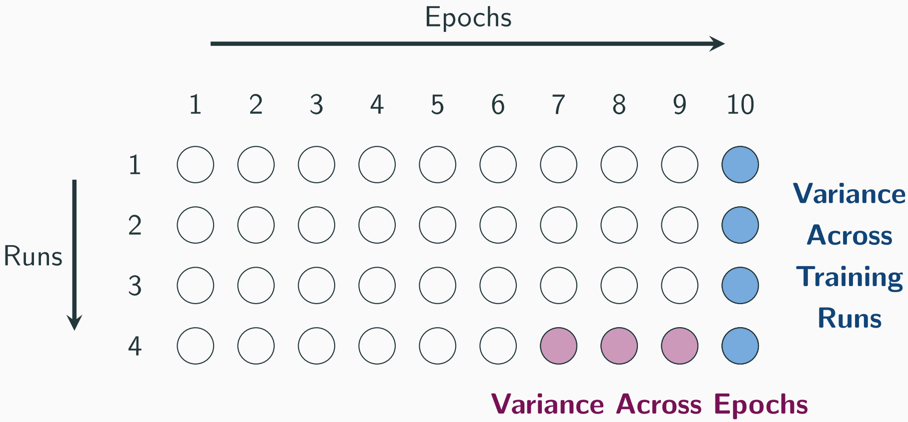
1 Run 2 Runs 3 Runs 4 Runs 5 Runs 6 Runs 7 Runs 8 Runs 9 Runs 10 Runs



50 Runs

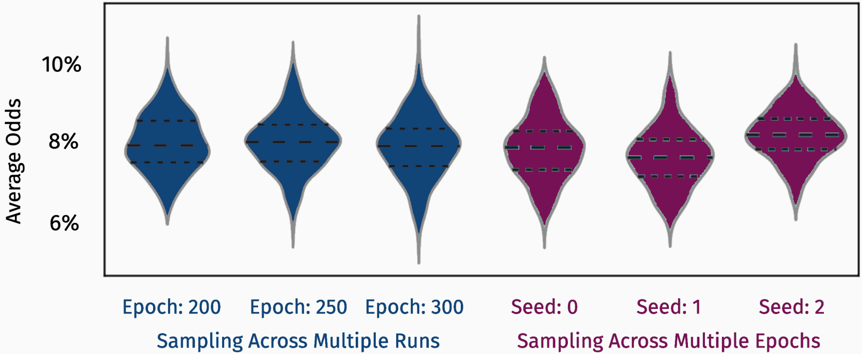

Fixed Data Reshuffling
Fixed Weight Initialization
Average Odds
10%
8%
6%
100
150
200
250
300
Epoch
Median
Inter-quartile range
Overall range

Variance Across Epochs vs Training Runs



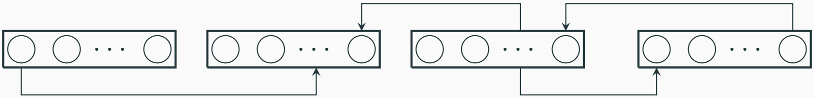
Variance Across Epochs vs Training Runs

The distribution of fairness scores **across multiple runs** is similar to the distribution of fairness scores **across epochs in any single run.**



Manipulating Fairness with Data Order

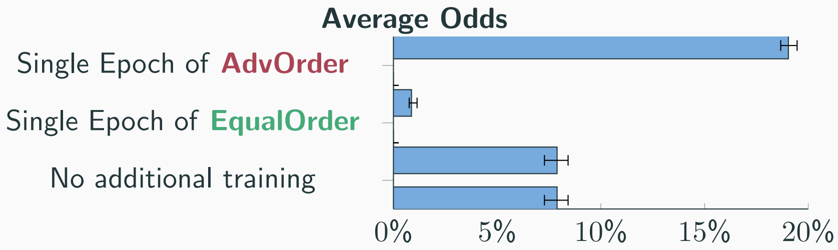
Guiding Principle: The most recent gradient updates seen by the model have a significant influence on its fairness scores!

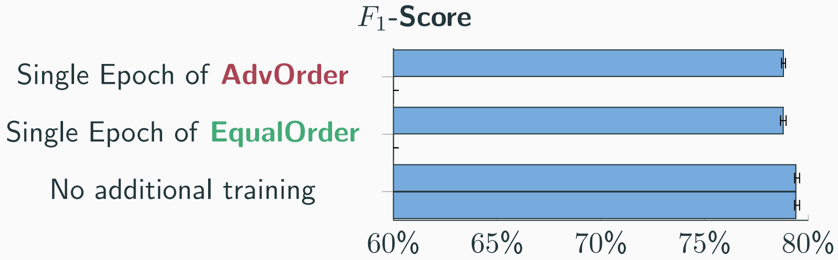


EqualOrder
To improve fairness scores

AdvOrder
To adversarially introduce bias

Bias Mitigation with Data Order





Takeaways

1. Which bias mitigation method we choose depends on the access we have to a model's training
2. We cannot achieve both ideal fairness and accuracy
3. These technical tradeoffs encourage non-technical solutions (e.g., collecting higher quality data)

Fairness and Trustworthy ML


Minority

Robustness vs. Fairness

Recall: Equalizing error rates
For any two groups a, b, we require

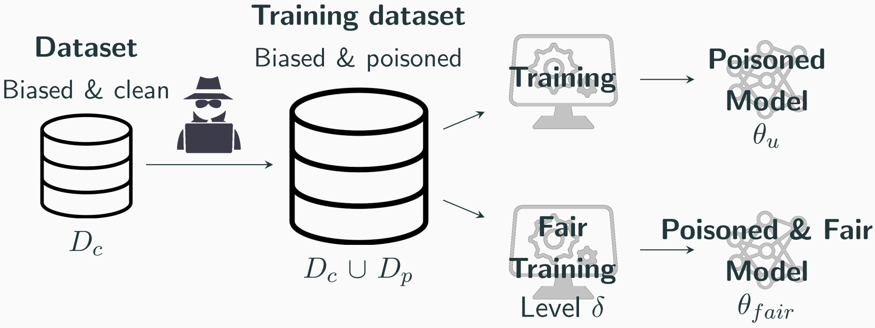
$$\mathbb{P}[D = 1|Y = 0, A = a] = \mathbb{P}[D = 1|Y = 0, A = b] \qquad \text{(equal FPR)}$$
$$\mathbb{P}[D = 0|Y = 1, A = a] = \mathbb{P}[D = 0|Y = 1, A = b] \qquad \text{(equal FNR)}$$

Quantifying Fairness
The level of unfairness can be measured as a difference.

$$\delta = \max(|\mathbb{P}[D = 1|Y = 0, A = a] - \mathbb{P}[D = 1|Y = 0, A = b]\}| \,, \\ |\mathbb{P}[D = 0|Y = 1, A = a] - \mathbb{P}[D = 0|Y = 1, A = b]|)$$

Robustness of fair models

What's the robustness cost of ensuring fairness?

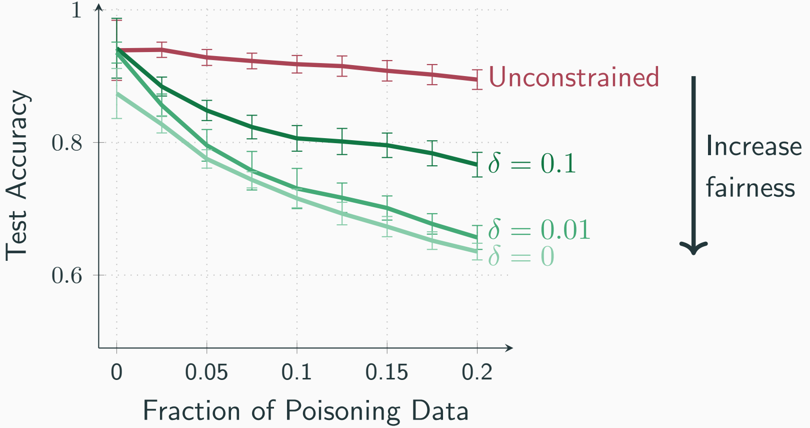


Attack objective

$$\max_{D_p} L(\theta_u; D_{test}) \text{ where } \theta_u = \arg\min_{\theta} L(\theta; D_c \cup D_p)$$

Fairness hurts robustness

Note: δ measures the fairness gap

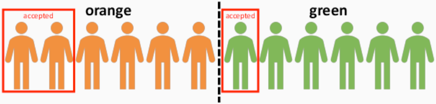


Source: [Chang et al., 2020]

Privacy and Fairness

Can fair models leak more information about their training data?

(Case study of group fairness)



Same prediction error across groups

Source: [Chang and Shokri, 2021]

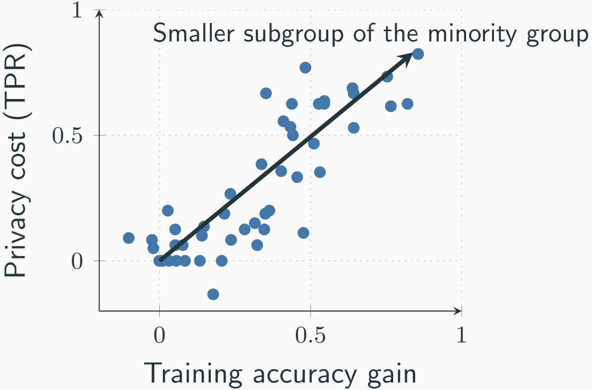
Privacy Across Different Subgroups

What are the most vulnerable points on the fair model?

Gain and Risks of Group Fairness

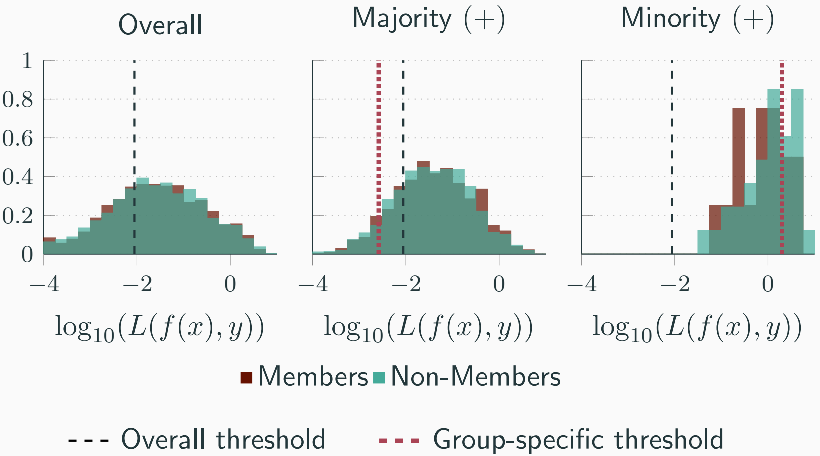
Members of the minority group potentially gain higher **influence** on the fair model

Insight
This can significantly increase the privacy risk on minorities



Our attack strategy

Our proposal: Find an attack model for each subgroup (defined by the label and sensitive attribute)



Attack accuracy

Synthetic data the model satisfies equalized odds (Fairness gap ≤ 0.001)

Attack	Target model	Min (+)	Maj (+)	Min (-)	Maj (-)
Single model	Standard	0.529	0.512	0.518	0.512
	Fair	0.608	0.528	0.524	0.522
Subgroup-based	Standard	0.618	0.528	0.524	0.522
	Fair	0.692	0.534	0.525	0.515

How else could we solve these problems?

Example Imagine a scenario where we are asked to predict likelihood that people will show up for their court dates. Those who are predicted to not appear will be jailed.

Problem People with young children are much more likely to be predicted high risk. They would receive much more disruptive outcomes.

Observation Classifying risk doesn't solve the actual problem here (we want people to show up for their court dates).

Solution Rather than using this prediction system, offer childcare, transportation vouchers to enable more people to make it to their appointments.

Source: Despart [2019]

Overall Takeaways

What is fairness?

Amir, S., van de Meent, J.-W., and Wallace, B. C. (2021).
On the impact of random seeds on the fairness of clinical classifiers.
arXiv preprint arXiv:2104.06338.

Baldini, I., Wei, D., Ramamurthy, K. N., Yurochkin, M., and Singh, M. (2021).
Your fairness may vary: Pretrained language model fairness in toxic text classification.
arXiv preprint arXiv:2108.01250.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. (2019).
Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias.
IBM Journal of Research and Development, 63(4/5):4–1.

Chang, H., Nguyen, T. D., Murakonda, S. K., Kazemi, E., and Shokri, R. (2020).
On adversarial bias and the robustness of fair machine learning.
arXiv preprint arXiv:2006.08669.

Chang, H. and Shokri, R. (2021).
On the privacy risks of algorithmic fairness.
In 2021 IEEE European Symposium on Security and Privacy (EuroS&P), pages 292–303. IEEE.

Ganesh, P., Chang, H., Strobel, M., and Shokri, R. (2023).
On the impact of machine learning randomness on group fairness.
In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pages 1789–1800.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018).
Learning adversarially fair and transferable representations.
In International Conference on Machine Learning, pages 3384–3393. PMLR.

Sellam, T., Yadlowsky, S., Wei, J., Saphra, N., D'Amour, A., Linzen, T., Bastings, J., Turc, I., Eisenstein, J., Das, D., et al. (2021).
The multiberts: Bert reproductions for robustness analysis.
arXiv preprint arXiv:2106.16163.

Icon credits

- Data by shashank singh from NounProject.com
- Neural Network by Ian Rahmadi Kurniawan from NounProject.com
- Software Engineering by Eli Magaziner from NounProject.com
- Data by IconPai from NounProject.com
- Rules by Adrien Coquet from NounProject.com
- Lightning by RULI from NounProject.com
- Lock by Abdan Bagus Panuntun from NounProject.com