Disparate Conditional Prediction in Multiclass Classifiers

ICML 2025

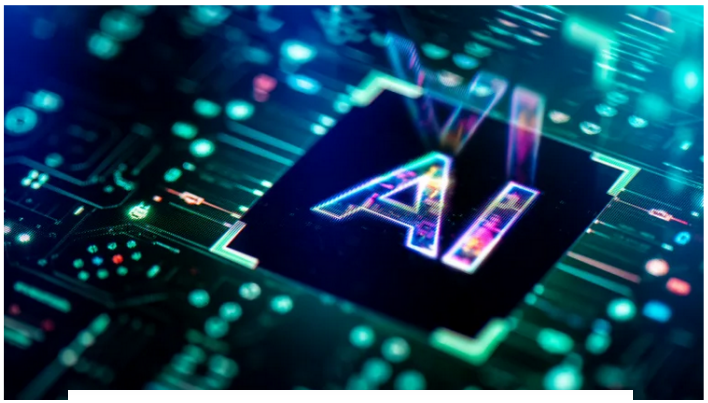
Sivan Sabato, Eran Treister and Elad Yom-Tov

Discrimination by AI is widespread

EEOC Settles First-Ever AI Discrimination Lawsuit

By Raeann Burgo and Wendy Hughes © Fisher Phillips
August 17, 2023

LIKE SAVE PRINT EMAIL



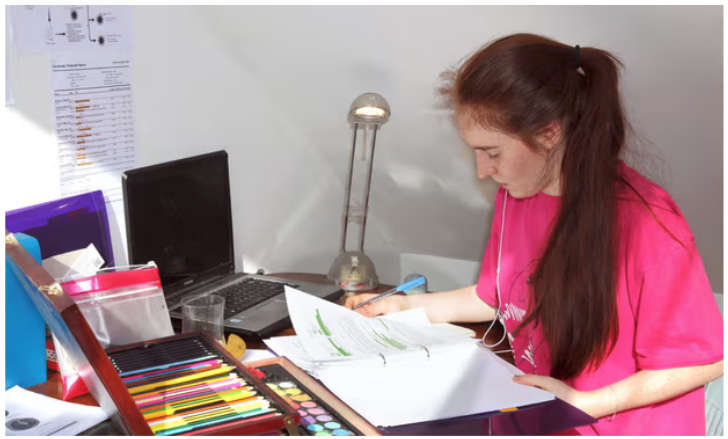
FIRST OPINION

How artificial intelligence could make pregnancy discrimination in employment more common

By Anya E.R. Prince Nov. 30, 2023 Reprints

Programs to detect AI discriminate against non-native English speakers, shows study

Over half of essays written by people were wrongly flagged as AI-made, with implications for students and job applicants



BUSINESS ›

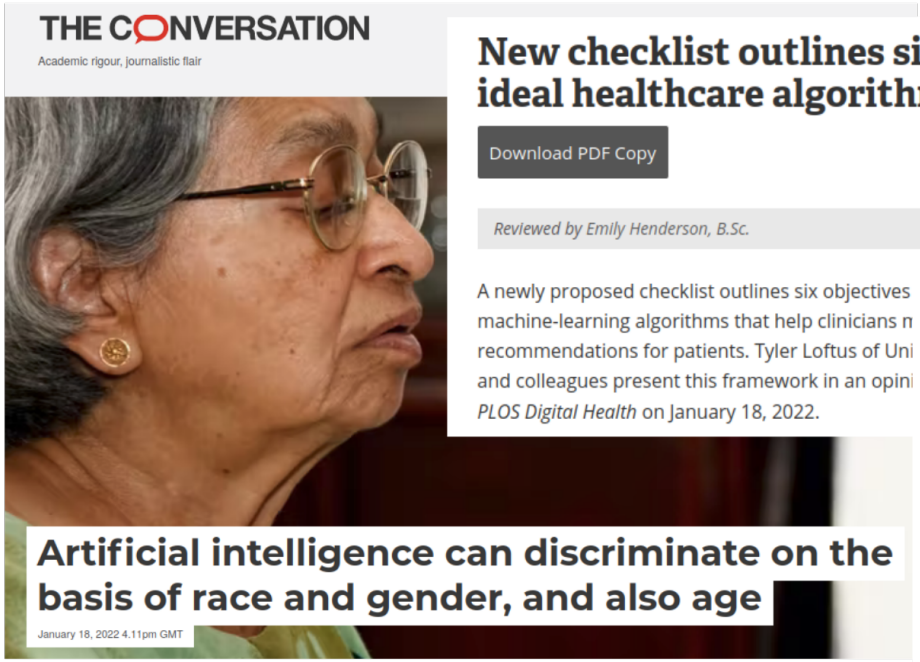
Stanford study indicates AI chatbots used by health providers are perpetuating racism

CBS NEWS BAY AREA

UPDATED ON: OCTOBER 20, 2023 / 6:32 PM PDT / AP

Image source: CBS news, STAT news, the Guardian, SHRM

Checking for fairness



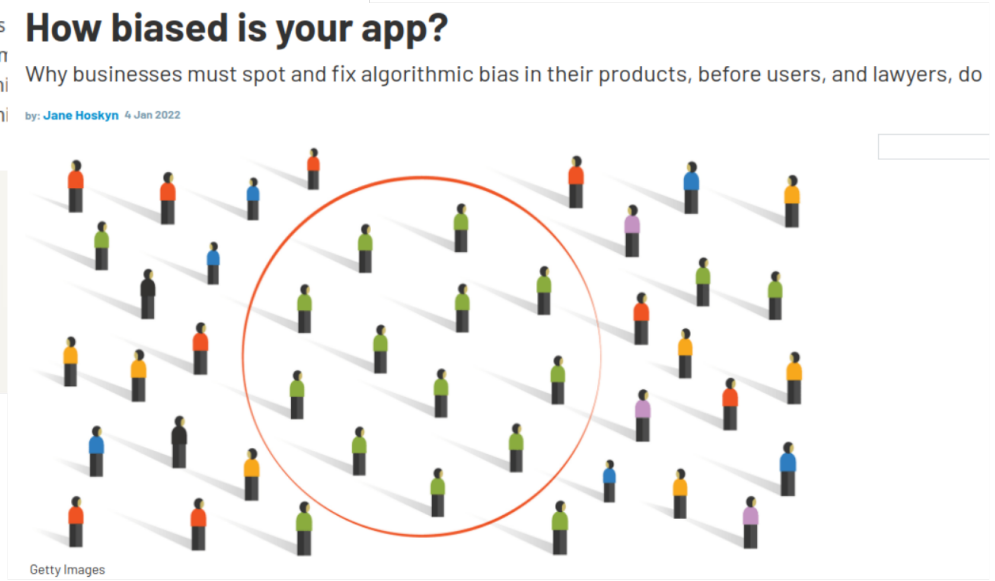


Image source: theconversation.com,www.itpro.co.uk,www.news-medical.net,wikiquote.org,wikipedia.org

Is the classifier fair?

- The protected attribute A has several possible values.
- A confusion matrix for each value a of A:

$$\mathcal{C}_a = \begin{pmatrix} \mathbb{P}[\hat{Y} = 1 \mid Y = 1], & \dots & \mathbb{P}[\hat{Y} = k \mid Y = 1] \\ & \dots & \\ \mathbb{P}[\hat{Y} = k \mid Y = 1], & \dots & \mathbb{P}[\hat{Y} = k \mid Y = k] \end{pmatrix}$$

- Fairness under **multiclass equalized odds**: all matrices are the same.

Image source: wikipedia.org

Beyond exact fairness

- In practice, exact fairness may be impractical.
- **How** fair/unfair is a given classifier?
- Which classifier is more fair?
- **How to quantify unfairness in an interpretable way?**

Unfairness measures

- Previous work
  - Most previous unfairness measures are ad-hoc
    - ⋆ Difference-based (e.g., Donini et al. 2018, Want et al. 2024)
    - ⋆ Ratio-based (e.g., Calmon et al. 2017, Alghamdi et al. 2022)
  - The value they provide is not directly interpretable.

Unfairness measures

- Previous work
  - Most previous unfairness measures are ad-hoc
    - ★ Difference-based (e.g., Donini et al. 2018, Want et al. 2024)
    - ★ Ratio-based (e.g., Calmon et al. 2017, Alghamdi et al. 2022)
  - The value they provide is not directly interpretable.
  - Sabato et al. 2020 proposed an interpretable measure.
    - ★ **DCP**: Disparate Conditional Prediction.
    - ★ Provides a directly interpretable unfairness measure.
    - ★ However, that work only considered binary classifiers.

Unfairness measures

- Previous work
  - Most previous unfairness measures are ad-hoc
    - ★ Difference-based (e.g., Donini et al. 2018, Want et al. 2024)
    - ★ Ratio-based (e.g., Calmon et al. 2017, Alghamdi et al. 2022)
  - The value they provide is not directly interpretable.
  - Sabato et al. 2020 proposed an interpretable measure.
    - ★ **DCP**: Disparate Conditional Prediction.
    - ★ Provides a directly interpretable unfairness measure.
    - ★ However, that work only considered binary classifiers.
- Our contributions
  - We generalize DCP to multiclass classifiers
  - We develop a computational approach for calculating the DCP in the multiclass case.
  - We show how to find the best-case DCP without access to the confusion matrices.

Disparate Conditional Prediction [Sabato et al., 2020]

Definition: Unfairness

The *unfairness of a classifier* is the **fraction of the population** that this classifier treats differently from the baseline.

Disparate Conditional Prediction [Sabato et al., 2020]

Definition: Unfairness

The *unfairness of a classifier* is the **fraction of the population** that this classifier treats differently from the baseline.

- $\mathcal{B}_{\hat{Y}|Y}$: The **baseline** conditional distribution of the classifier's prediction given the true label.
- $\mathcal{N}^a_{\hat{Y}|Y}$: A **nuisance** conditional distribution
  - Can be different for each value of the protected attribute.

Disparate Conditional Prediction [Sabato et al., 2020]

Definition: Unfairness

The *unfairness of a classifier* is the **fraction of the population** that this classifier treats differently from the baseline.

- $\mathcal{B}_{\hat{Y}|Y}$: The **baseline** conditional distribution of the classifier's prediction given the true label.
- $\mathcal{N}^a_{\hat{Y}|Y}$: A **nuisance** conditional distribution
  - Can be different for each value of the protected attribute.
- The conditional label distribution of the classifier:

$$\mathbb{P}[\hat{Y} \mid Y, A] = \eta_{A,Y} \cdot \mathcal{N}^A_{\hat{Y}|Y} + (1 - \eta_{A,Y}) \cdot \mathcal{B}_{\hat{Y}|Y}.$$

Disparate Conditional Prediction [Sabato et al., 2020]

Definition: Unfairness

The *unfairness of a classifier* is the **fraction of the population** that this classifier treats differently from the baseline.

- $\mathcal{B}_{\hat{Y}|Y}$: The **baseline** conditional distribution of the classifier's prediction given the true label.
- $\mathcal{N}^a_{\hat{Y}|Y}$: A **nuisance** conditional distribution
  - Can be different for each value of the protected attribute.
- The conditional label distribution of the classifier:

$$\mathbb{P}[\hat{Y} \mid Y, A] = \eta_{A,Y} \cdot \mathcal{N}^A_{\hat{Y}|Y} + (1 - \eta_{A,Y}) \cdot \mathcal{B}_{\hat{Y}|Y}.$$

- **DCP** $:= \min \sum_{a,y} \mathbb{P}[A = a, Y = y] \cdot \eta_{a,y}$, where the minimum is taken over **all possible decompositions** of the classifier's true conditional label distribution into $\mathcal{B}_{\hat{Y}|Y}$, $\mathcal{N}^A_{\hat{Y}|Y}$.

Multiclass DCP

Theorem

For a multiclass classifier $\mathcal{C}$,

$$\mathrm{DCP}(\mathcal{C}) = \sum_{y \in \mathcal{Y}} \min_{\mathcal{C}_b[y] \in \Delta_k} \sum_{a \in \mathcal{A}} w_a \pi_a^y \max_{\hat{y} \in \mathcal{Y}} \eta(\mathcal{C}^{y\hat{y}}_{\text{baseline}}, \mathcal{C}^{y\hat{y}}_a),$$

where

$$\eta(a, b) = \begin{cases} 1 - b/a & b < a, \\ 1 - (1-b)/(1-a) & b > a, \\ 0 & b = a. \end{cases}$$

Bounding the DCP

- Unlike the binary case, the minimization of multiclass DCP is not known to be computationally tractable.
- We provide an analytical **lower bound**.
- We provide a local minimization procedure, which generates an **upper bound**.
  - The objective is non-smooth and non-convex
  - It also has regions with large gradients.
  - The local minimization procedure is based on sequential solutions of linear programming approximations to the objective.

Bounding the DCP without confusion matrices

- Sometimes, confusion matrices cannot be estimated
  - Lack of access to the classifier
  - Lack of quality validation data
- We can still bound the best-case DCP, using only high level statistics.
- This can be used to audit proprietary non-public classifiers for possible fairness issues.

Experiments

- We compare several approaches for generating the upper bound.
- The results show the advantage of our local minimization along with a greedy initialization.
- In most cases, the approximation factor of the bounds is close to 1.
- See paper for more experiments!

# Labels	Error	Lower Bound	Upper Bounds				Best Ratio
			Average	Greedy	Average+LM	Greedy+LM	
3	11.74%	5.39%	27.19%	9.65%	14.07%	**7.65%**	1.42
3	5.71%	4.35%	42.17%	5.92%	32.39%	**5.28%**	1.21
3	3.96%	3.24%	43.63%	5.07%	16.95%	**3.25%**	1.00
3	5.15%	4.24%	39.05%	5.40%	14.32%	**4.25%**	1.00
3	3.36%	2.65%	48.04%	5.20%	5.49%	**3.81%**	1.44
3	1.85%	1.85%	59.64%	8.22%	17.85%	**1.85%**	1.00
3	1.96%	1.96%	51.86%	7.88%	13.31%	**1.96%**	1.00
3	2.32%	2.28%	48.50%	6.57%	10.56%	**2.28%**	1.00
3	14.00%	4.57%	28.85%	6.47%	13.13%	**6.10%**	1.34
4	3.91%	1.48%	27.55%	8.12%	1.86%	**1.49%**	1.00
5	5.61%	2.06%	7.14%	43.01%	6.61%	**3.91%**	1.90
5	11.83%	4.57%	25.28%	34.40%	22.18%	**8.28%**	1.81
6	0.87%	0.86%	21.96%	24.65%	9.55%	**0.86%**	1.00
8	22.61%	8.29%	84.54%	32.03%	38.20%	**23.61%**	2.85
9	21.54%	5.03%	86.17%	12.50%	6.84%	**6.17%**	1.23

Poster and paper

https://icml.cc/virtual/2025/poster/45683