On the Interplay between Fairness and Explainability

Stephanie Brandl   Emanuele Bugliarello   Ilias Chalkidis

Department of Computer Science, University of Copenhagen

1. Motivation

💡 In order to build reliable and trustworthy NLP applications, models need to be **both fair across different demographics and explainable**.

🤔 Usually these two, **fairness** and **explainability**, are optimized and/ or examined independently of each other. Instead, we argue that forthcoming, trustworthy NLP systems should **consider both**.

Contributions:

I. We examine the **interplay** between **two crucial dimensions of trustworthiness**: fairness and explainability, by comparing models that were fine-tuned using **fairness-promoting techniques** or **rationale extraction frameworks**.

II. Our experiments on multi-class classification datasets (BIOS, ECtHR):

A. confirm recent findings on the **independence of bias mitigation and empirical fairness** (Cabello et al., 2023), and

B. show that also **empirically fairness and explainability are independent**.

2. Datasets

We experiment with two multi-class classification datasets:

(a) **MED-BIOS** (Eberle et al., 2023)

Medical Occupation Classification 👨‍⚕️

+ Gender: 👨 / 👩

(b) **ECtHR** (Chalkidis et al., 2021)

ECHR Judgment Forecasting 👨‍⚖️

+ Nationality: 🇪🇺 / {🇷🇺🇺🇦🇹🇷}

	BIOS	
Occupation	Male	Female
Psychologist	822 (37%)	1378 (63%)
Surgeon	1090 (85%)	190 (15%)
Nurse	152 (09%)	1486 (91%)
Dentist	996 (65%)	537 (35%)
Physician	650 (48%)	699 (52%)
Total	3710 (46%)	4290 (54%)
	ECtHR	
ECHR Article	E. European	Rest
3 – Proh. Torture	303 (88%)	42 (12%)
5 – Liberty	382 (88%)	51 (12%)
6 – Fair Trial	1776 (80%)	454 (20%)
8 – Private Life	129 (55%)	104 (45%)
P1.1 – Property	228 (88%)	31 (12%)
Total	2818 (80%)	682 (20%)

Table 1: Label and demographic attribute distribution across the training sets of the BIOS and ECtHR datasets.

3. Methods

We work with two groups of methods:

(a) Optimizing for **fairness**

1. Group Parity (Sun et al., 2009)
2. Group Neutralization (Brandl et al., 2022)
3. Group DRO (Sagawa et al., 2020)
4. Spectral Decoupling (Pezeshki et al., 2021)
5. Debiased Focal Loss (Orgav & Belinkov, 2022)

(b) Optimizing for **explainability**

1. Baseline REF (Lei et al., 2016)
2. 3-Player Game REF (Yu et al., 2019)
3. 3-Player+ Game REF (Chalkidis et al., 2021)

Optimize for Fairness

Representational Bias

Group Parity (FAIR-GP): 50% 👨‍⚕️ – 50% 👩‍⚕️
Group Neutralization (FAIR-GN): 👨‍⚕️ / 👩‍⚕️ → 🧑‍⚕️
Group DRO (FAIR-DRO): 50% 👨‍⚕️ – 50% 👩‍⚕️ + Adaptive losses

Penalize Over-confidency

Spectral Decoupling (FAIR-SD): CLS Pred. 99% ✅ → |L2|☝️ → Loss☝️
Debiased Focal Loss (FAIR-DFL): Detect Pred. 99% ✅ → Loss☝️

Optimize for Explainability

Rationale Extraction Frameworks

X → Rationale Extractor → 😇 **R** → 😇 Predictor → **Y**
😈 **R** → 😈 Predictor → **Y'**

Baseline: 😇 (Concise + Informative Rationales **R**)
3-Player Game: 😇 **R** vs. Complement-based 😈 **R**
3-Player+ Game: 😇 **R** vs. Random-choice 😈 **R**
Rationales 2 Attentions: Binary 😇 = Continuous 😇

4. Experiments & Results

(a) Synthetic Data

Method	Empirical Fairness (mF1) M ↑ / F ↑ / Diff. ↓	Nurse (M) ↑	Surgeon (F) ↑
	BIOS$_{\text{biased}}$ *(Artificially Unbalanced)*		
BASELINE	45.9 / 34.6 / 11.3	0.0	14.8
FAIR-GN	81.7 / 82.1 / 0.4	61.5	69.1
FAIR-DRO	53.5 / 60.6 / 7.1	0.0	48.5
FAIR-SD	48.7 / 50.5 / 1.8	0.0	38.7
FAIR-DFL	45.7 / 47.5 / 1.8	0.0	14.8
	BIOS$_{\text{balanced}}$ *(Artificially Balanced)*		
BASELINE	83.6 / 84.4 / 0.8	76.9	73.9
FAIR-GN	84.8 / 84.2 / 0.6	74.1	73.5
FAIR-DRO	84.8 / 85.0 / 0.2	74.1	79.2
FAIR-SD	83.5 / 86.2 / 2.6	71.4	80.0
FAIR-DFL	82.6 / 85.8 / 3.2	74.1	76.6

Table 2: Fairness-related metrics: macro-F1 (mF1) per group (male/female) and their absolute difference (Diff.), and worst-performing class (profession) per group, of fairness-promoting methods on the *ultra-biased* or *debiased* version of BIOS.

(b) Real Data

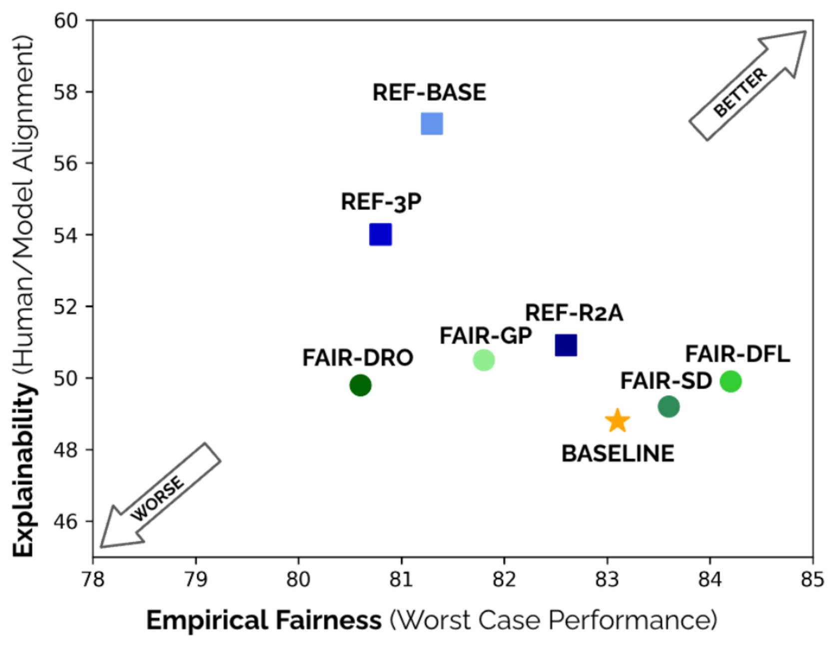

Figure 1: Interplay between *empirical fairness*, measured via worst-case performance, and *explainability* measured via human/model alignment, of different methods (Section 4) optimizing for fairness (FAIR), explainability (REF), or none (BASELINE) on the ECtHR

Method	BIOS mF1	BIOS Empirical Fairness mF1 (M / F / Diff.)	BIOS Explainability AOPC	BIOS Explainability R@k	ECtHR mF1	ECtHR Empirical Fairness mF1 (EE / R / Diff.)	ECtHR Explainability AOPC	ECtHR Explainability R@k
BASELINE	**88.1**	85.5 / **87.5** / 2.0	88.5	**52.0**	83.5	83.1 / 83.3 / **0.2**	77.4	48.8
Optimizing for Fairness								
FAIR-GP	87.8	83.8 / **87.5** / 3.7	88.0	47.8	83.9	83.5 / 81.8 / 2.5	77.0	50.5
FAIR-GN	87.8	82.5 / 86.8 / 4.2	88.0	48.7	Not Applicable (N/A)[4]			
FAIR-DRO	87.6	84.2 / 86.4 / 2.2	88.4	48.8	83.9	83.6 / 80.6 / 3.0	77.9	49.8
FAIR-SD	87.9	85.6 / 86.6 / **1.0**	88.5	49.4	**84.9**	**84.2** / **87.1** / 2.9	**78.8**	49.9
FAIR-DFL	87.6	84.5 / 86.4 / 1.9	87.3	45.5	84.3	84.1 / 83.6 / 0.5	78.2	49.2
Optimizing for Explainability								
REF-BASE	85.3	82.2 / 83.9 / 1.7	78.1	45.7	81.8	81.9 / 81.3 / 0.6	73.2	**57.1**
REF-3P	86.4	81.8 / 85.0 / 3.1	79.6	44.3	83.1	83.3 / 80.8 / 2.5	73.3	54.0
REF-R2A	86.1	82.4 / 85.4 / 3.0	82.9	50.7	82.8	82.6 / 83.4 / 0.8	74.5	50.9

Table 3: Test Results for all examined methods. We report the overall macro-F1 (mF1), alongside fairness-related metrics: macro-F1 (mF1) per group and their absolute difference (Diff.), also referred to as group disparity; and explainability-related scores: AOPC for faithfulness and token R@k for human-model rationales alignment. The best scores across all models in the same group (FAIR-, REF-) are underlined, and the best scores overall are in **bold**.

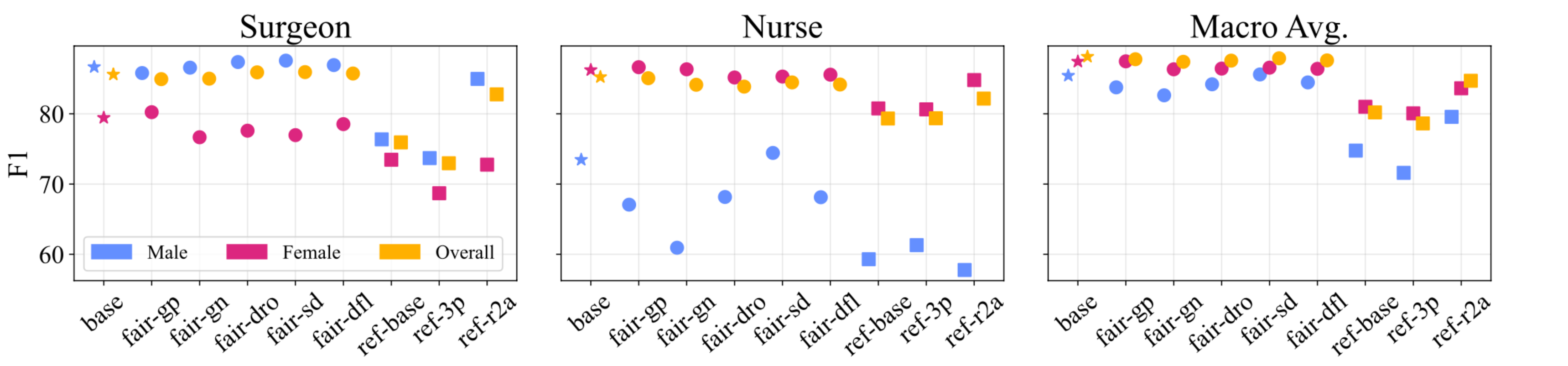

Figure 3: F1 and macro-F1 scores for the classes *surgeon* and *nurse* from the BIOS dataset for all methods per gender. Baseline is marked as ⋆, fairness-promoting methods as ∘, and REFs as □. We see a severe drop in performance for the underrepresented class (female surgeons and male nurses).

(c) Bias Mitigation

Method	Fairness (mF1) WC ↑	Diff. ↓	Bias Proxies $\|L2\|$ ↓	Group Acc. ↓
	BIOS – Occupation Classification			
BASELINE	85.5	2.0	12.6	93.2
FAIR-GP	83.8	3.7	18.6	96.6
FAIR-GN	82.5	4.2	11.6	65.4
FAIR-DRO	84.2	2.2	21.2	98.2
FAIR-SD	85.6	1.0	00.7	96.0
FAIR-DFL	84.5	1.9	06.5	96.2
	ECtHR – ECHR Violation Prediction			
BASELINE	83.1	0.2	10.7	75.0
FAIR-GP	81.8	2.7	11.3	69.6
FAIR-DRO	80.6	3.0	16.7	76.2
FAIR-SD	84.2	2.9	00.4	72.4
FAIR-DFL	83.6	0.5	04.5	63.0

Table 4: Fairness- and bias-related metrics. We show again downstream task performance for *Worst-Case* (WC) and the group-wise difference as indicators for empirical fairness. We further add $L2$ norm of the classification logits as an indicator for (over-)confidency and accuracy for group classification both as bias proxies.

5. Takeaways

A. Improving either empirical fairness or explainability does not improve the other.

B. Many fairness-promoting methods do not mitigate bias, nor promote fairness as intended (Figure 1).

C. Gender information is encoded to a high amount in the occupation classification task, and the only successful strategy to prevent this seems to be the normalization across genders during training.