On the Interplay between Fairness and Explainability

Stephanie Brandl **Emanuele Bugliarello** **Ilias Chalkidis**
Department of Computer Science, University of Copenhagen, Denmark
{brandl, emanuele, ilias.chalkidis}@di.ku.dk

Abstract

In order to build reliable and trustworthy NLP applications, models need to be both fair across different demographics and explainable. Usually these two objectives, *fairness* and *explainability*, are optimized and/or examined independently of each other. Instead, we argue that forthcoming, trustworthy NLP systems should consider both. In this work, we perform a first study to understand how they influence each other: do *fair(er)* models rely on *more plausible* explanations? and vice versa. To this end, we conduct experiments on two English multi-class text classification datasets, BIOS and ECtHR, that provide information on gender and nationality, respectively, as well as human-annotated rationales. We fine-tune pre-trained language models with several methods for (i) bias mitigation, which aims to improve fairness; (ii) rationale extraction, which aims to produce plausible explanations. We find that bias mitigation algorithms do not always lead to fairer models. Moreover, in our analysis, we see that empirical fairness and explainability are orthogonal.

1 Introduction

Fairness and explainability are crucial factors when building trustworthy NLP applications. This is true in general, but even more so in critical and sensitive applications such as medical (Gu et al., 2020) and legal (Chalkidis et al., 2022a) domains, as well as in algorithmic hiring processes (Schumann et al., 2020). AI trustworthiness and governance are no longer wishful thinking since more and more legislatures introduce related regulations for the assessment of AI technologies, such as the EU Artificial Intelligence Act (2022), the US Algorithmic Accountability Act (2022), and the Chinese Measures on Generative AI (2023). Therefore, it is important to ask and answer challenging questions that can lead to safe and trustworthy AI systems, such as how fairness and explainability interplay when optimizing for either or both.

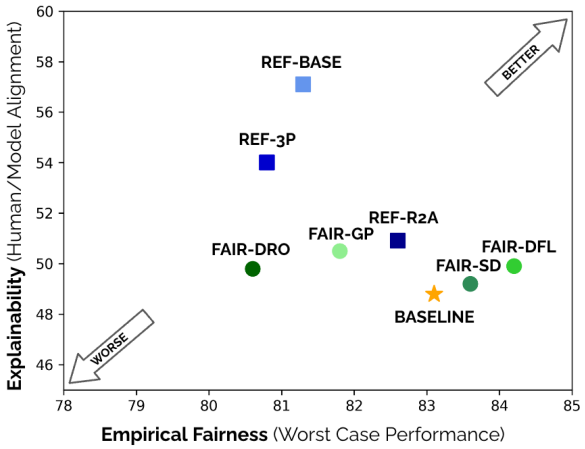

Figure 1: Interplay between *empirical fairness*, measured via worst-case performance, and *explainability* measured via human/model alignment, of different methods (Section 4) optimizing for fairness (FAIR), explainability (REF), or none (BASELINE) on the ECtHR dataset. All methods, including the baseline, are built upon fine-tuned RoBERTa models. The results here suggest that the two dimensions are independent.

So far in the NLP literature, model explanations[1] are used to detect and mitigate how fair or biased a model is (Balkir et al., 2022) or to assess a user's perception of a model's fairness (Zhou et al., 2022). Those are important use cases of explainability but we argue that we should further aim for improving one when optimizing for the other to promote trustworthiness holistically across both dimensions.

To analyze the interplay between fairness and explainability, we optimize neural classifiers for one or the other during fine-tuning, and then evaluate both afterwards (Figure 1). We do so across two English multi-class classification datasets. First, we analyze a subset of the BIOS dataset (De-Arteaga et al., 2019). This dataset contains short biographies for occupation classification. We consider a subset of 5 medical professions that also

[1]We refer to both the feature attribution scores assigned by models (binary and continuous) and the binary annotations by humans as *rationales* throughout the paper, and also use the term *(model) explanations* for the former.

includes human annotations on 100 biographies across this subset (Eberle et al., 2023). We evaluate model-based rationales extracted via (i) LRP (Ali et al., 2022) or (ii) rationale extraction frameworks (REFs; Lei et al. 2016), while also examining fairness with respect to gender. Second, we also conduct similar experiments with the ECtHR dataset (Chalkidis et al., 2021) for legal judgment forecasting on cases from the European Court of Human Rights (ECHR), both to evaluate paragraph-level rationales and to study fairness with respect to the nationality of the defendant state.

Contributions. Our main contributions in this work are the following: **(i)** We examine the *interplay* between two crucial dimensions of trustworthiness: *fairness* and *explainability*, by comparing models that were fine-tuned using five fairness-promoting techniques (Section 4.1) and three rationale extraction frameworks (Section 4.2) on two English multi-class classification dataset (BIOS and ECtHR). **(ii)** Our experiments on both datasets (a) confirm recent findings on the independence of bias mitigation and empirical fairness (Cabello et al., 2023), and (b) show that also empirical fairness and explainability are independent.

2 Related Work

Bias mitigation / fairness. The literature on inducing fairer models from biased data is rapidly growing (see Mehrabi et al. 2021; Makhlouf et al. 2021; Ding et al. 2021 for recent surveys). Fairness is often conflated with bias mitigation, although they have been shown to be orthogonal: reducing bias, such as representational bias, may not lead to a fairer model in terms of downstream task performance (Cabello et al., 2023). In this work, we follow the definition of Shen et al. (2022) and target *empirical fairness* (performance parity) that may not align with *representational fairness* (data parity). This means that we adopt a capability-centered approach to fairness and define fairness in terms of performance parity (Hashimoto et al., 2018) or equal risk (Donini et al., 2018). The fairness-promoting learning algorithms that we evaluate are discussed in detail in Section 4.

Explainable AI (XAI) for fairness. Explanations have been used to improve user's perception and judgement of fairness (Shulner-Tal et al., 2022; Zhou et al., 2022). Balkir et al. (2022) give an overview of the *ACL literature where explainability is applied to detect and mitigate bias. They predominantly find work on uncovering and investigating bias for hate speech detection. Recently, also Ruder et al. (2022) call for more multi-dimensional NLP research where fairness, interpretability, multilinguality and efficiency are combined. The authors only find work by Vig et al. (2020) who use explainability to find specific parts of a model that are causally implicated in its behaviour. With this work, we want to extend this line of research from 'XAI for fairness' to 'XAI and Fairness'.

Post-hoc XAI. XAI methods commonly rely on saliency maps extracted post-hoc from a model using attention scores (Bahdanau et al., 2015; Abnar and Zuidema, 2020), gradients (Voita et al., 2019; Wallace et al., 2019; Ali et al., 2022), or perturbations (Ribeiro et al., 2016; Alvarez-Melis and Jaakkola, 2017; Murdoch et al., 2018) at inference time. These can be seen as an approximation of identifying which features (tokens) the model relied on to solve a given task for a specific example. Such methods do not necessarily lead to *faithful* explanations (Jacovi and Goldberg, 2020). Following DeYoung et al. (2020), faithfulness can be defined as the combination of *sufficiency*—tokens with the highest scores correspond to a sufficient selection to reliably predict the correct label—*and comprehensiveness*—all indicative tokens get attributed relatively high scores.

Rationale extraction by design. Unlike post-hoc explanations, *rationale extraction* frameworks *optimize* for rationales that support a given classification task and are faithful by design, *i.e.*, predictions are based on selected rationales by definition.

Lei et al. (2016) were the first to propose a framework to produce short coherent rationales that could replace the original full texts, while maintaining the model's predictive performance. The rationales are extracted by generating binary masks indicating which words should be selected; and two additional loss regularizers were introduced, which penalize long rationales and sparse masks that would select non-consecutive words.

Recently, several studies have proposed improved frameworks that rely mainly on complementary adversarial settings that aim to produce better (causal, complete) rationales and close the performance gap compared to models using the full input (Yu et al., 2019; Chang et al., 2019; Jain et al., 2020; Yu et al., 2021). The rationale extraction frameworks we evaluate are detailed in Section 4.

XAI *and* fairness. Mathew et al. (2021) release a benchmark for hate speech detection where human annotations are used as input to the model to improve performance and fairness across demographics. They evaluate both faithfulness of post-hoc explanations as well as performance disparity across communities affected by hate speech. He et al. (2022) propose a new debiasing framework that consists of two steps. First, they apply the rationale extraction framework (REF) from Lei et al. (2016) to detect tokens indicative of a given *bias* label, *e.g.*, gender. In a second step, those rationales are used to minimize bias in the task prediction.

With this work, we aim to complement prior work by systematically evaluating the impact of optimizing for fairness on explainability and vice versa, considering many different proposed techniques (Section 4). Moreover, we consider both post-hoc explanations extracted from standard Transformer-based classifiers, as well as rationale extraction frameworks evaluating model-based explanations (rationales) in terms of faithfulness and alignment with human-annotated rationales.

3 Datasets

BIOS. The BIOS dataset (De-Arteaga et al., 2019) comprises biographies labeled with occupations and binary gender in English. This is an occupation classification task, where bias with respect to gender can be studied. In our work, we consider a subset of 10,000 (8K train / 1K validation / 1K test) biographies targeting 5 medical occupations (*psychologist*, *surgeon*, *nurse*, *dentist*, *physician*) published by Eberle et al. (2023). For these occupations, as shown in Table 1, there is a clear gender imbalance, *e.g.*, 91% of the nurses are female, while 85% of the surgeons are male. We also compare with human rationales provided for a subset of 100 biographies.

For control experiments on the effect of bias mitigation methods, we also create a synthetic extremely unbalanced (*biased*) version of the train and validation split of BIOS, we call this version $\text{BIOS}_{\texttt{biased}}$. Here, our aim is to amplify gender-based spurious correlations in the training subset by keeping only the biographies where all psychologists and nurses are female; and all surgeons, dentists, and physicians are male. Similarly, we create a synthetic balanced (*debiased*) version of the dataset which we call $\text{BIOS}_{\texttt{balanced}}$. Here, our objective is to mitigate gender-based spurious correlations by down-sampling the majority group per medical profession. By doing so, in $\text{BIOS}_{\texttt{balanced}}$, both genders are equally represented per profession.

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.

ECtHR. The ECtHR dataset (Chalkidis et al., 2021) contains 11K cases from the European Court of Human Rights (ECHR) written in English. The Court hears allegations that a European state has breached human rights provisions of the European Convention of Human Rights (ECHR). For each case, the dataset provides a list of *factual* paragraphs (facts) from the case description. Each case is mapped to *articles* of the ECHR that were violated (if any). The dataset also provides silver (automatically extracted) paragraph-level rationales. We consider a subset of 4,500 (3.5K train / 500 validation / 500 test) single-labeled cases for five well-supported ECHR articles and the *defendant state* attribute. In practice, we use a binary categorization of the defendant states—Eastern European vs. the Rest—to assess fairness, similar to Chalkidis et al. (2022b). Table 1 shows a clear defendant state imbalance across most of the ECHR articles except for Article 8.

4 Methodology

We fine-tune classifiers optimizing for either fairness (Section 4.1), explainability (Section 4.2), or none, alongside the main classification task objective (Figure 2). The baseline classifier uses an n-way classification head on top of the Transformer-based text encoder (Vaswani et al., 2017), and it is

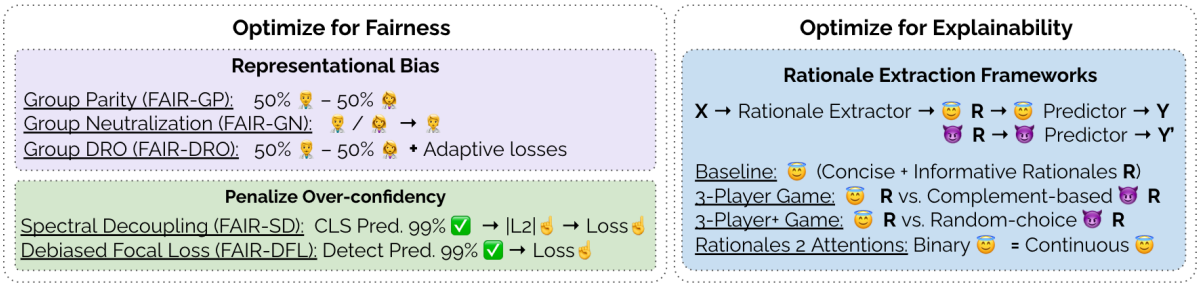

Figure 2: A short description / depiction of the *fairness-promoting* (Section 4.1) and *explainability-promoting* (Section 4.2) examined methods. The emojis represent male/female/neutral, and main, and adversarial modules.

optimized using the cross-entropy loss against the gold labels (Devlin et al., 2019).

4.1 Optimizing for Fairness

We use a diverse set of 5 fairness-promoting algorithms that are connected to two different approaches: (a) mitigating *representational bias* (FAIR-GP, FAIR-GN, FAIR-DRO), and (b) penalizing *overconfident predictions* (FAIR-SD, FAIR-DFL).

Representational bias *Representational bias* (*e.g.*, more data points for male vs. female surgeons) is considered a critical factor that may lead to performance disparity across demographic groups, as a model may rely on the protected attribute (*e.g.*, gender) as an indicator for predicting the output class (*e.g.*, occupation). We consider three methods to mitigate such effects:

i) **Group Parity** (FAIR-GP) where we over-sample the minority group examples per class up to the same level as the majority ones (Sun et al., 2009). For instance, by up-sampling biographies of male nurses and female surgeons in the BIOS dataset.

ii) **Group Neutralization** (FAIR-GN), where we replace (normalize) attribute-related information. For instance, for gender in BIOS, we replace gendered pronouns (*e.g.*'he/him', 'she/her'), and titles (*e.g.*'Mr', 'Mrs'), with gender-neutral equivalents, such as 'they/them' and 'Mx' (Brandl et al., 2022a), while also replacing personal names with a placeholder name (Maudslay et al., 2019), such as 'Sarah Williams' with 'Joe Doe'.

iii) **Group Robust Optimization** (FAIR-DRO) where we use GroupDRO as proposed by Sagawa et al. (2020). In this case, we apply group parity (up-sampling) on the training set to have group-balanced batches, while the optimization loss during training accounts for group-wise performance disparities using adaptive group-wise weights.

Penalizing overconfidency *Overconfident* model predictions are considered an indication of bias based on the intuition that all simple feature correlations—leading to high confidence—are spurious (Gardner et al., 2021). We consider two methods from this line of work:

iv) **Spectral Decoupling** (FAIR-SD) where the L_2 norm of the classification logits is used as a regularization penalty. The premise for this approach is that overconfidence reflects over-reliance to a limited number of relevant features, which leads to gradient starvation (Pezeshki et al., 2021).

v) **Debiased Focal Loss** (FAIR-DFL) where an additional task-agnostic classifier estimates if the model's prediction is going to be successful or not, and penalizes the model via focal loss (Karimi Mahabadi et al., 2020) in case a successful outcome is highly predictable (Orgad and Belinkov, 2023).

The first group of methods (representational bias) relies on demographic information, while the second group (penalizing overconfidency) is agnostic of demographic information, thus more easily applicable to different settings.

4.2 Optimizing for Explainability

We consider three alternative rationale extraction frameworks (REFs), where the models generate *rationales*; *i.e.*, a subset of the original tokens to predict the classification label. In these settings, the explanations (rationales) are *faithful* by design, since the classifier (predictor) encodes only the rationales and has no access to the full text input, thus soley relies on those rationales at inference.

i) **Baseline** (REF-BASE) The baseline rationale extraction framework of Lei et al. (2016) relies on two sub-networks (Eqs. 1-4): the *rationale selector* that selects relevant input tokens to predict the correct label (Eq. 1-2), and the *predictor* (Eq. 3-4) that predicts the classification task outcome based on the rationale provided by the first module.

ii) **3-Player** (REF-3P) Yu et al. (2019) improved the aforementioned framework introducing a 3-player adversarial minimax game between the main predictor, the one using the rationale, and a newly introduced predictor using the complement of the rationale tokens. They found that this method improves classification performance, and the predicted rationales are more complete (*i.e.*, they include a higher portion of the relevant information to solve the task) than the baseline framework.

iii) **Rationale to Attention** (REF-R2A) More recently, Yu et al. (2021) introduced a new 3-player version where, during training, they minimize the performance disparity between the main predictor (the one using the rationales) and a second one using soft attention scores. They find this to result in rationales that better align with human rationales.

For all examined rationale extraction frameworks, we use the latest implementations provided by Yu et al. (2021), which use a top-k token selector, instead of sparsity regularization (Lei et al., 2016):

$$S = W^{H\times 1} * \text{TokenScorer}(X) \quad (1)$$
$$Z = \text{TopK}(X, S, k) \quad (2)$$
$$R = Z * X \quad (3)$$
$$L = \text{Predictor}(R) \quad (4)$$

where TokenScorer and Predictor are Transformer-based language models (encoders), $X = [x_1, x_2, \cdots, x_n]$ are the input tokens, S are the token importance scores based on the TokenScorer contextualized token representations, Z is a binary mask representing which input tokens are the top-k scored vs. the rest, R is the rationale (masked version of the input tokens), and L are the label logits. During training, the TopK operator is detached—since it is not differentiable—and gradients pass *straight-through* (Bengio et al., 2013) to the TokenScorer to be updated. For REF-3P, there is an additional adversarial Predictor (Eq. 4) which is fed with adversarial rationales (R_{adv}) based on the complement (REF-3P) of the original ones (R), while for REF-R2A, the adversarial predictor weighs the input tokens (X) given softmax-normalized scores (S).

5 Experiments

5.1 Experimental Setup

We fine-tune classifiers based on RoBERTa-base (Liu et al., 2019) for all examined methods. In the case of the ECtHR dataset, which consists of long documents, we build hierarchical RoBERTa-based classifiers similar to Chalkidis et al. (2022a).[2] We perform a hyperparameter search over the learning rate $\in [1e-5, 3e-5, 5e-5]$ with an initial warm-up of 10%, followed by cosine decay, using AdamW (Loshchilov and Hutter, 2019). We use a fixed batch size of 32 examples and fine-tune models up to 30 epochs with early stopping based on the validation performance. We fine-tune models with 5 different seeds and select the top-3 models (seeds) with the best overall validation performance (mF1) to report averaged results for all metrics.

For methods optimizing for fairness, we rely on the LRP framework (Ali et al., 2022) to extract post-hoc explanations, similar to Eberle et al. (2023).

Evaluation metrics. Our main performance metric is macro-F1 (mF1); *i.e.*, the F1 score macro-averaged across all classes, which better reflects the overall performance across classes regardless of their training support (robust to class imbalance) than accuracy.

Regarding *empirical fairness* metrics, we report group-wise performances (*e.g.*, male and female mF1 in BIOS, and E.E. and the Rest in ECtHR) and their absolute difference (group disparity). Ideally, a fair(er) model will improve the worst-case performance, i.e., the lower performance across both groups, while reducing the group disparity.

For *explainability*, we report Area Over the Perturbation Curve (AOPC) for *sufficiency* (DeYoung et al., 2020) as a proxy to *faithfulness* (Jacovi and Goldberg, 2020); *i.e.*, how much explanations reflect the true reasoning—as reflected by importance scores—of a model. We compute sufficiency for all models using as a reference (classifier) a large RoBERTa model to have a fair common ground. We also report token-level recall at human level (R@k), similar to Chalkidis et al. (2021), considering the top-k tokens, where k is the number of tokens annotated by humans,[3] as a metric of alignment (agreement) between model-based explanations and human rationales.

For estimating *bias*, we report the L_2 norm of the classification logits, which is used as a regularization penalty by Spectral Decoupling (Pezeshki et al., 2021) as a proxy for confidence. We also

[2] Similarly, rationales (Eq. 1-3) are computed based on paragraph-level, not token-level, representations.

[3] In this case, all models are compared in a fair manner using the number of the selected tokens in the human rationale as an oracle.

report gender accuracy, as a proxy for bias, by fine-tuning probing classifiers on the protected attribute examined (*e.g.*, gender classifiers for BIOS) initialized by the models previously fine-tuned on the downstream task (Section 5.4)

5.2 Results on Synthetic Data

In Table 2, we present results for all fairness-promoting methods in the artificially unbalanced (biased) and balanced (debiased) versions of the BIOS dataset: $\text{BIOS}_{\texttt{biased}}$ and $\text{BIOS}_{\texttt{balanced}}$, described in Section 3. These can be seen as control experiments, to assess methods in edge cases.

Fairness methods rely on biases in data. When training on $\text{BIOS}_{\texttt{biased}}$, we observe that all fairness-promoting methods outperform the baseline method in terms of our empirical fairness metrics: worst-group, i.e., female, performance and group disparity (difference in performance for male and female). We further see that almost all methods have mF1 scores of 0 when it comes to *male nurses* and very low scores ($15 - 49$) for *female surgeons*. For both classes (*nurse* and *surgeon*), there were only their female and male counterpart, respectively, in the training dataset of $\text{BIOS}_{\texttt{biased}}$. This result suggests that all but one fairness-promoting methods (namely FAIR-GN) heavily rely on gender information to solve the task when such a spurious correlation is present. Only FAIR-GN, where gender information is neutralized, is able to solve the task reliably, including almost no group disparity and mF1 scores > 60 for male nurses and female surgeons. In Table 8 in the Appendix, we present the top-attributed words for both occupations per gender which support this finding. All methods, except FAIR-GN, attribute gendered words a high (positive or negative) score following the imbalance in training. Words such as 'she', 'mrs.', and 'her' are positively attributed for females nurses, while 'he' is negatively attributed for male nurses; and symmetrically the opposite for surgeons (Table 8). The only exception is FAIR-GN, in which case gender information has been neutralized during training and testing and the model can no longer exploit such superficial spurious correlations, *e.g.*, that females can only be nurses or psychologists. Concluding, all fairness-promoting methods *improve* empirical fairness compared to the baseline, but in such extreme scenarios only a direct manual intervention on the data as in FAIR-GN reaches meaningful performance.

Method	Empirical Fairness (mF1)		
	M ↑ / F ↑ / Diff. ↓	**Nurse (M) ↑**	**Surgeon (F) ↑**
$\text{BIOS}_{\texttt{biased}}$ *(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
$\text{BIOS}_{\texttt{balanced}}$ *(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.

Data debiasing improves fairness methods. After downsampling the data to reach an equal number of males and females for all five professions for $\text{BIOS}_{\texttt{balanced}}$, we see almost equal performance across genders for BASELINE, FAIR-GN and FAIR-DRO (*lower* part of Table 2). Moreover, the performance for FAIR-GN and FAIR-DRO is both higher and more equal across M and F than for BASELINE. Overall, the models show an mF1 score of around 3% lower than in the main results in Table 3, which is probably caused by down-sampling (fewer training samples), and to a smaller degree from not relying on gender bias.

5.3 Main Results on Real Data

In Table 3, we present results for all examined methods for both datasets, BIOS and ECtHR.

Overall performance. In the case of BIOS, we observe a drop in performance, in particular when optimizing for explainability where mF1 scores decrease from 88% down to 85% in comparison to the BASELINE. We also see an increase in group disparity for 3 out of 5 fairness-promoting methods and 2 out of 3 explainability-promoting methods. This is further supported by the results in Figure 3, where we show F1 scores for the two classes *surgeon* and *nurse* from the BIOS dataset (see Figure 4 in Appendix for results across all classes and metrics). We see a severe drop in performance for the two most underrepresented demographics, female surgeons and male nurses, of up to 25% in comparison to the overrepresented counterpart. In contrast, in the case of ECtHR, fairness-promoting (bias mitigation) methods, have a beneficial effect, especially

Method	BIOS – Occupation Classification				ECtHR – ECHR Violation Prediction			
	mF1	Empirical Fairness mF1 (M / F / Diff.)	Explainability AOPC	R@k	mF1	Empirical Fairness mF1 (EE / R / Diff.)	Explainability AOPC	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**. We present detailed validation and test results including standard deviations in Tables 5- 7.

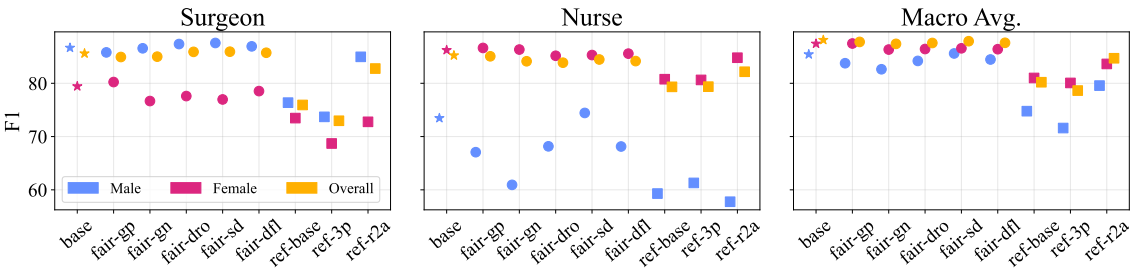

Figure 3: F1 and macro-F1 scores for the classes *surgeon* and *nurse* from the BIOS dataset for all methods per gender. Baseline is marked as ⋆, fairness-promoting methods as ○, and REFs as □. We see a severe drop in performance for the underrepresented class (female surgeons and male nurses).

in the case of confidence-related methods FAIR-SD and FAIR-DFL where overall task performance increase by $0.8-1.4\%$ with respect to the BASELINE. We suspect that the positive impact in the case of ECtHR is partly a side-effect of a higher class imbalance (label-wise disparity), e.g., there are many more cases tagged with Article 6 compared to the rest of the labels, as demonstrated in Table 1 (lower part), similar to the findings of Chalkidis and Søgaard (2022) who showed that FAIR-SD works particularly well for high class imbalance.

Fairness-promoting methods. In the case of BIOS, we observe that only FAIR-SD can slightly improve empirical fairness, reflected through lower group disparity at the cost of a lower group performance for FEMALE (F), while the remaining fairness-promoting methods lead to a more or similar unfair performance. We observe similar results for ECtHR, where only two out of four methods (FAIR-SD, FAIR-DFL) are able to improve the performance for both groups (EE, R), while increasing the group disparity, as all other methods.[4] Concluding, we find that bias mitigation algorithms do not always lead to fairer models which is in line with Cabello et al. (2023). Considering explainability-related metrics—faithfulness and human-model alignment as measured by R@k—for the fairness-promoting (bias mitigation) methods, we observe that improved empirical fairness does not lead to *better* model explanations, neither for faithfulness (AOPC) nor for plausibility (R@k) when comparing FAIR-SD and FAIR-DFL with the BASELINE.

Rationale Extraction Frameworks (REFs). Considering the results for the rationale extraction frameworks (REFs, see Section 4.2) presented in the lower part of Table 3, we observe that the overall performance (mF1) decreases by 2-3% in the

[4]We do not consider FAIR-GN in ECtHR, since there is no straightforward way to anonymize (neutralize) information relevant to the defendant state, which is potentially presented in the form of mentions to locations, organization, etc..

case of BIOS, and by 0.5-2% for ECtHR, since the models' predictor only considers a subset of the original input, the rationale. All REFs that aim to improve explainability compromise empirical fairness (*i.e.*, performance disparity) in both datasets.

When it comes to explainability, the results are less clear. For BIOS, both scores—faithfulness and human-model alignment—, drop in comparison to the baseline, while all REF methods substantially improve human–model alignment (by 2-8%) in the case of ECtHR. For REFs, we also observe that an improvement in empirical fairness does not correlate with an improvement in explainability.

5.4 Bias Mitigation $\neq$ Empirical Fairness

Based on our findings in Section 5.3, we investigate the dynamics between bias mitigation and empirical fairness further. We examine the fairness-promoting methods on both datasets considering two indicators of bias: (a) the $L2$ norm of the classification logits as a proxy for the model's overconfidency (also used as a penalty term by FAIR-SD), and (b) the accuracy of a probing classifier for predicting the attribute (gender/nationality). This probing classifier relies on a frozen encoder that was previously fine-tuned on the occupation/article classification task with a newly trained classification head. To avoid conflating bias with features learned for the main classification tasks, e.g., medical occupation classification for BIOS, we use new datasets, excluding documents with the original labeling, e.g., for BIOS we use 3K biographies for 5 non-medical professions to train the gender classifier. With this analysis, we want to quantify to what degree we can extract information on gender/nationality, from the biographies/court cases.

In Table 4, we report empirical fairness metrics and the bias indicators (proxies) for all examined methods together with F1 scores for *worst-case-scenario* (WC) across all classes and the difference in mF1 between the two groups from Table 3. First of all, with respect to BIOS, we observe that all fairness-promoting algorithms, except FAIR-GN, show a high accuracy for gender classification ($> 95\%$), thus, are more biased compared to the baseline with respect to gender classification accuracy. Furthermore, the least biased classifier (FAIR-GN), is outperformed by all other fairness-promoting methods in both empirical fairness metrics. In the case of ECtHR, we observe that 3 out of 4 fairness-promoting methods decrease bias, shown by lower group accuracy and lower confidency scores (L2 norm) for FAIR-SD and FAIR-DFL. This does not lead to improvements in empirical fairness, with the exception of worst-case performance for FAIR-SD and FAIR-DFL.

Method	Fairness (mF1)		Bias Proxies	
	WC ↑	**Diff.** ↓	$\lvert L2\rvert$ ↓	**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	<u>65.4</u>
FAIR-DRO	84.2	2.2	21.2	98.2
FAIR-SD	<u>85.6</u>	<u>1.0</u>	<u>00.7</u>	96.0
FAIR-DFL	84.5	1.9	06.5	96.2
ECtHR – ECHR Violation Prediction				
BASELINE	83.1	<u>0.2</u>	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	<u>84.2</u>	2.9	<u>00.4</u>	72.4
FAIR-DFL	83.6	0.5	04.5	<u>63.0</u>

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.

6 Conclusion

We systematically investigated the interplay between empirical fairness and explainability, two key desired properties required for trustworthy NLP systems. We did so by considering five fairness-promoting methods, and three rationale extraction frameworks, across two datasets for multi-class classification (BIOS and ECtHR). Based on our results, we see that improving either empirical fairness or explainability does *not* improve the other. Interestingly, many fairness-promoting methods do not mitigate bias, nor promote fairness as intended, while we find that these two dimensions are also orthogonal (Figure 1). Furthermore, we see that (i) gender information is encoded to a high amount in the occupation classification task, and (ii) the only successful strategy to prevent this seems to be the normalization across gender during training. We conclude that trustworthiness, reflected through empirical fairness and explainability, is still an open challenge. With this work, we hope to encourage more efforts that target a holistic investigation and the development of new algorithms that promote both crucial qualities.

Limitations

Our analysis is limited to English text classification datasets. In order to make general conclusions about the interplay between fairness and explainability, one need to extend this analysis to other languages, downstream tasks and more datasets.

Datasets that provide both annotations for demographics and rationales are very rare. We consider the two out of three that we found available, excluding the one in (Thorn Jakobsen et al., 2023) because the demographic annotations were referring to the annotators and not to groups affected by the task per se. We hope that our work motivates future benchmarks that aim at assessing both fairness and explainability at larger scales.

We do neither consider generative models nor generative explanations for this work as fairness and explainability methods are not fully developed at the point of carrying out this analysis. We leave this for future work.

Furthermore, we argue within the limited scope of specific definitions of fairness, bias and explainability for binary attributes. This analysis could be applied to further attributes. Also, we have not included human evaluation with respect to explainability, i.e., humans evaluating the rationales for usability and plausibility, see Brandl et al. (2022b); Yin and Neubig (2022).

Acknowledgements

We thank our colleagues at the CoAStaL NLP group for fruitful discussions in the beginning of the project. In particular, we would like to thank Daniel Hershcovich, Desmond Elliott, Laura Cabello and Rita Ramos for valuable comments on the manuscript. EB is supported by the funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199. IC is funded by the Novo Nordisk Foundation (grant NNF 20SA0066568). SB receives funding from the European Union under the Grant Agreement no. 10106555, FairER. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor REA can be held responsible for them.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

US Algorithmic Accountability Act. 2022. Algorithmic Accountability Act (US AAA). Discussed by the US Congress.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. 2022. XAI for transformers: Better explanations through conservative propagation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 435–451. PMLR.

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

EU Artificial Intelligence Act. 2022. Artificial Intelligence Act (EU AIA). Proposed by the European Commission.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. 2022. Challenges in applying explainability methods to improve the fairness of NLP models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 80–92, Seattle, U.S.A. Association for Computational Linguistics.

Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022a. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

Stephanie Brandl, Daniel Hershcovich, and Anders Søgaard. 2022b. Evaluating deep taylor decomposition for reliability assessment in the wild. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media*.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 370–378.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022b. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis and Anders Søgaard. 2022. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, Dublin, Ireland. Association for Computational Linguistics.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S. Jaakkola. 2019. *A Game Theoretic Approach to Class-Wise Selective Rationalization*. Curran Associates Inc., Red Hook, NY, USA.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Oliver Eberle, Ilias Chalkidis, Laura Cabello, and Stephanie Brandl. 2023. Rather a nurse than a physician - contrastive explanations under investigation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6907–6920, Singapore. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew Peters, Alexis Ross, Sameer Singh, and Noah A. Smith. 2021. Competency problems: On finding and removing artifacts in language data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938, Stockholmsmässan, Stockholm Sweden. PMLR.

Zexue He, Yu Wang, Julian McAuley, and Bodhisattwa Prasad Majumder. 2022. Controlling bias exposure for fair interpretable predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5854–5866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the applicability of machine learning fairness notions. *SIGKDD Explor. Newsl.*, 23(1):14–23.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Chinese Measures on Generative AI. 2023. Measures on Generative AI. Released by the Cyberspace Administration of China.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Hadas Orgad and Yonatan Belinkov. 2023. BLIND: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821, Toronto, Canada. Association for Computational Linguistics.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. 2021. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*.

Candice Schumann, Jeffrey Foster, Nicholas Mattei, and John Dickerson. 2020. We need fairness and explainability in algorithmic hiring. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95, Online only. Association for Computational Linguistics.

Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology*, 24(1):2.

Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. 2009. Classification of imbalanced data: a review. *Int. J. Pattern Recognit. Artif. Intell.*, 23:687–719.

Terne Sasha Thorn Jakobsen, Laura Cabello, and Anders Søgaard. 2023. Being right for whose right reasons? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1033–1054, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Mo Yu, Yang Zhang, Shiyu Chang, and Tommi Jaakkola. 2021. Understanding interlocking dynamics of cooperative rationalization. *Advances in Neural Information Processing Systems*, 34.

Jianlong Zhou, Fang Chen, and Andreas Holzinger. 2022. Towards explainability for ai fairness. In *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 375–386. Springer.

A More results

In Table 5- 6, we present validation, and test results across all methods for both examined datasets.

In Table 8, we present the list of words that were assigned the highest importance scores (positive and negative) for the 5 fairness-promoting methods and the baseline on the BIOS dataset. Additionally, we show class-wise F1 scores, separated by gender, for the BIOS dataset in Figure 4.

Method	BIOS – Occupation Classification				ECtHR – ECHR Violation Prediction			
	Avg.	Empirical Fairness			Avg.	Empirical Fairness		
		M	F	Diff.		EE	R	Diff.
BASELINE	89.7 ± 0.1	**90.9** ± 0.2	86.2 ± 0.7	4.7 ± 0.7	87.2 ± 0.2	87.4 ± 0.7	84.3 ± 3.4	3.1 ± 1.7
Optimizing for Fairness								
FAIR-GP	89.9 ± 0.1	89.7 ± 1.0	86.9 ± 0.1	2.8 ± 1.1	86.3 ± 0.4	87.0 ± 0.4	81.4 ± 0.8	5.6 ± 0.5
FAIR-GN	89.1 ± 0.2	86.7 ± 1.3	85.7 ± 1.0	1.0 ± 1.4	———— Not Applicable (N/A) ————			
FAIR-DRO	89.7 ± 0.3	90.5 ± 1.0	86.4 ± 0.8	4.1 ± 1.7	86.9 ± 0.9	87.6 ± 0.7	82.5 ± 2.4	5.1 ± 1.8
FAIR-SD	**90.3** ± 0.0	90.2 ± 0.9	87.7 ± 0.3	2.5 ± 0.6	87.6 ± 1.1	**88.5** ± 1.0	82.9 ± 1.9	5.6 ± 1.0
FAIR-DFL	90.0 ± 0.1	88.5 ± 0.6	**88.0** ± 0.4	**0.5** ± 0.7	**88.1** ± 0.7	88.4 ± 0.8	**85.8** ± 2.9	**2.6** ± 2.9
Optimizing for Explainability								
REF-BASE	87.2 ± 0.2	88.5 ± 0.2	82.7 ± 1.2	5.8 ± 1.1	87.1 ± 0.2	87.5 ± 0.2	85.1 ± 2.5	3.1 ± 1.8
REF-3P	86.8 ± 0.6	87.1 ± 2.1	81.1 ± 0.9	6.0 ± 1.4	86.9 ± 0.5	87.7 ± 0.3	83.7 ± 1.9	4.1 ± 2.0
REF-R2A	87.5 ± 0.4	88.5 ± 1.5	83.7 ± 1.3	4.8 ± 1.9	88.0 ± 0.9	88.4 ± 0.8	85.8 ± 0.9	2.6 ± 0.3

Table 5: Validation Results (mF1) with standard deviations (±) for all examined methods in the examined datasets. We report the overall (Avg.) macro-F1 (mF1), alongside fairness-related metrics: macro-F1 (mF1) per group and their absolute difference (Diff.), also referred to as group disparity. The best scores across all models in the same group (FAIR-, REF-) are underlined, and the best scores overall are in **bold**.

Method	BIOS – Occupation Classification				ECtHR – ECHR Violation Prediction			
	Avg.	Empirical Fairness			Avg.	Empirical Fairness		
		M	F	Diff.		EE	R	Diff.
BASELINE	**88.1** ± 0.3	85.5 ± 1.4	**87.5** ± 0.9	2.0 ± 1.2	83.5 ± 0.6	83.1 ± 0.7	83.3 ± 0.8	0.2 ± 0.7
Optimizing for Fairness								
FAIR-GP	87.8 ± 0.4	83.8 ± 1.6	**87.5** ± 0.3	3.7 ± 1.2	83.9 ± 0.2	83.5 ± 0.2	81.8 ± 2.2	2.5 ± 1.3
FAIR-GN	87.8 ± 0.2	82.5 ± 0.6	86.8 ± 0.6	4.2 ± 1.1	———— Not Applicable (N/A) ————			
FAIR-DRO	87.6 ± 0.6	84.2 ± 0.4	86.4 ± 1.2	2.2 ± 1.3	83.9 ± 0.5	83.6 ± 0.5	80.6 ± 2.0	3.0 ± 1.7
FAIR-SD	87.9 ± 0.1	**85.6** ± 0.3	86.6 ± 0.2	1.0 ± 0.4	**84.9** ± 0.2	**84.2** ± 0.2	**87.1** ± 2.9	2.9 ± 3.1
FAIR-DFL	87.6 ± 0.6	84.5 ± 0.8	86.4 ± 0.6	1.9 ± 0.9	84.3 ± 1.0	84.1 ± 0.6	83.6 ± 4.2	0.5 ± 1.8
Optimizing for Explainability								
REF-BASE	85.3 ± 0.9	82.2 ± 1.1	83.9 ± 0.9	1.7 ± 1.0	81.8 ± 1.8	81.9 ± 2.1	81.3 ± 3.5	0.6 ± 0.9
REF-3P	86.4 ± 0.7	81.8 ± 1.0	85.0 ± 1.4	3.1 ± 1.4	83.1 ± 0.3	83.3 ± 0.6	80.8 ± 2.2	2.5 ± 1.8
REF-R2A	86.1 ± 0.6	82.4 ± 0.4	85.4 ± 1.0	3.0 ± 1.0	82.8 ± 0.6	82.6 ± 0.5	83.4 ± 2.6	0.8 ± 0.8

Table 6: Test Results (mF1) with standard deviations (±) for all examined methods in the examined datasets. We report the overall (Avg.) macro-F1 (mF1), alongside fairness-related metrics: macro-F1 (mF1) per group and their absolute difference (Diff.), also referred to as group disparity. The best scores across all models in the same group (FAIR-, REF-) are underlined, and the best scores overall are in **bold**.

Method	BIOS – Occupation Classification		ECtHR – ECHR Violation Prediction	
	Explainability		Explainability	
	AOPC	**R@k**	**AOPC**	**R@k**
BASELINE	**88.5** ± 0.0	**52.0** ± 1.7	77.4 ± 0.8	48.8 ± 0.2
Optimizing for Fairness				
FAIR-GP	88.0 ± 0.0	47.8 ± 2.5	77.0 ± 0.7	50.5 ± 0.4
FAIR-GN	88.0 ± 0.0	48.7 ± 2.3	—— Not Applicable (N/A) ——-	
FAIR-DRO	88.4 ± 0.0	48.8 ± 0.9	77.9 ± 0.2	49.8 ± 0.8
FAIR-SD	**88.5** ± 0.0	49.4 ± 3.2	**78.8** ± 0.8	49.9 ± 0.3
FAIR-DFL	87.3 ± 0.0	45.5 ± 2.4	78.2 ± 0.7	49.2 ± 1.6
Optimizing for Explainability				
REF-BASE	78.1 ± 0.0	45.7 ± 4.0	73.2 ± 1.4	**57.1** ± 0.7
REF-3P	79.6 ± 0.0	44.3 ± 2.9	73.3 ± 0.5	54.0 ± 1.0
FAIR-R2A	82.9 ± 0.0	50.7 ± 7.4	74.9 ± 1.0	50.9 ± 0.3

Table 7: Test Results for all examined methods. We report explainability-related scores with standard deviations (±): 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**.

Method	NURSE				SURGEON			
	POSITIVE		NEGATIVE		POSITIVE		NEGATIVE	
	M	F	M	F	M	F	M	F
BASELINE	(nursing, 0.2) (nurse, 0.2) - - -	(**mrs.**, 0.4) (nurses, 0.4) (nursing, 0.3) (**she**, 0.3) (nurse, 0.2)	(**he**, -0.3) - - - -	(research, -0.2) (inc, -0.2) (no, -0.1) (**elizabeth**, -0.1) (mental, -0.1)	(surgeon, 0.4) (surgery, 0.4) (surgical, 0.3) (surgeons, 0.3) (neurosurgery, 0.2)	(surgery, 0.5) (practice, 0.1) (dr., 0.1) (treatment, 0.1) -	(working, -0.2) (care, -0.2) (interests, -0.2) (health, -0.1) (md, -0.1)	(**she**, -0.3) (**her**, -0.1) (health, -0.1) - -
FAIR-GN	(nurse, 0.5) (nursing, 0.4) - - -	(nurse, 0.6) (nursing, 0.4) (nurses, 0.4) (rn, 0.3) (diabetes, 0.1)	- - - - -	(research, -0.2) (dr., -0.1) (practice, -0.1) (work, -0.1) -	(surgeon, 0.5) (neurosurgery, 0.4) (surgery, 0.4) (surgeons, 0.4) (surgical, 0.3)	(surgery, 0.6) (dr., 0.1) - - -	(working, -0.2) (group, -0.2) (over, -0.1) (health, -0.1) (general, -0.1)	(health, -0.2) (center, -0.1) - - -
FAIR-DRO	(nurse, 0.1) (nursing, 0.1) - - -	(**mrs.**, 0.4) (nursing, 0.3) (**she**, 0.3) (nurses, 0.2) (**ms.**, 0.2)	(**he**, -0.2) - - - -	(research, -0.3) (mental, -0.2) (affiliates, -0.1) (no, -0.1) (without, -0.1)	(surgeon, 0.4) (surgeons, 0.4) (surgery, 0.3) (neurosurgery, 0.3) (surgical, 0.3)	(surgery, 0.5) - - - -	(care, -0.2) (group, -0.2) (5, -0.2) (areas, -0.1) (experience, -0.1)	(**she**, -0.3) (**her**, -0.2) (health, -0.2) - -
FAIR-SD	(nursing, 0.1) - - - -	(**mrs.**, 0.2) (**she**, 0.1) (nursing, 0.1) (nurses, 0.1) (**ms.**, 0.1)	(**he**, -0.1) - - - -	(mental, -0.2) (research, -0.1) (dr., -0.1) (via, -0.1) (who, -0.1)	(surgeon, 0.3) (surgery, 0.3) (surgeons, 0.2) (surgical, 0.2) (surgeries, 0.2)	(surgery, 0.3) (practice, 0.3) (âĝļs, 0.1) - -	(group, -0.1) (general, -0.1) (supports, -0.1) (health, -0.1) (clinic, -0.1)	(**she**, -0.1) - - - -
FAIR-DFL	- - - - -	(**she**, 0.2) (mrs., 0.1) (**ms.**, 0.1) (**her**, 0.1) -	(**he**, -0.2) (medical, -0.1) - - -	(doctors, -0.1) (:, -0.1) (other, -0.1) (groups, -0.1) (., -0.1)	(surgeon, 0.6) (neurosurgery, 0.5) (surgery, 0.3) (surgeons, 0.2) (surgical, 0.2)	(surgery, 0.4) (shield, 0.1) (dr., 0.1) - -	(each, -0.2) (working, -0.1) (care, -0.1) (general, -0.1) ((, -0.1)	(**she**, -0.2) (**her**, -0.1) - - -

Table 8: Top-attributed positive and negative words based on normalized LRP scores for the unbalanced (biased) version of BIOS. We normalize positive and negative independently using the softmax function and aggregate across all test examples.

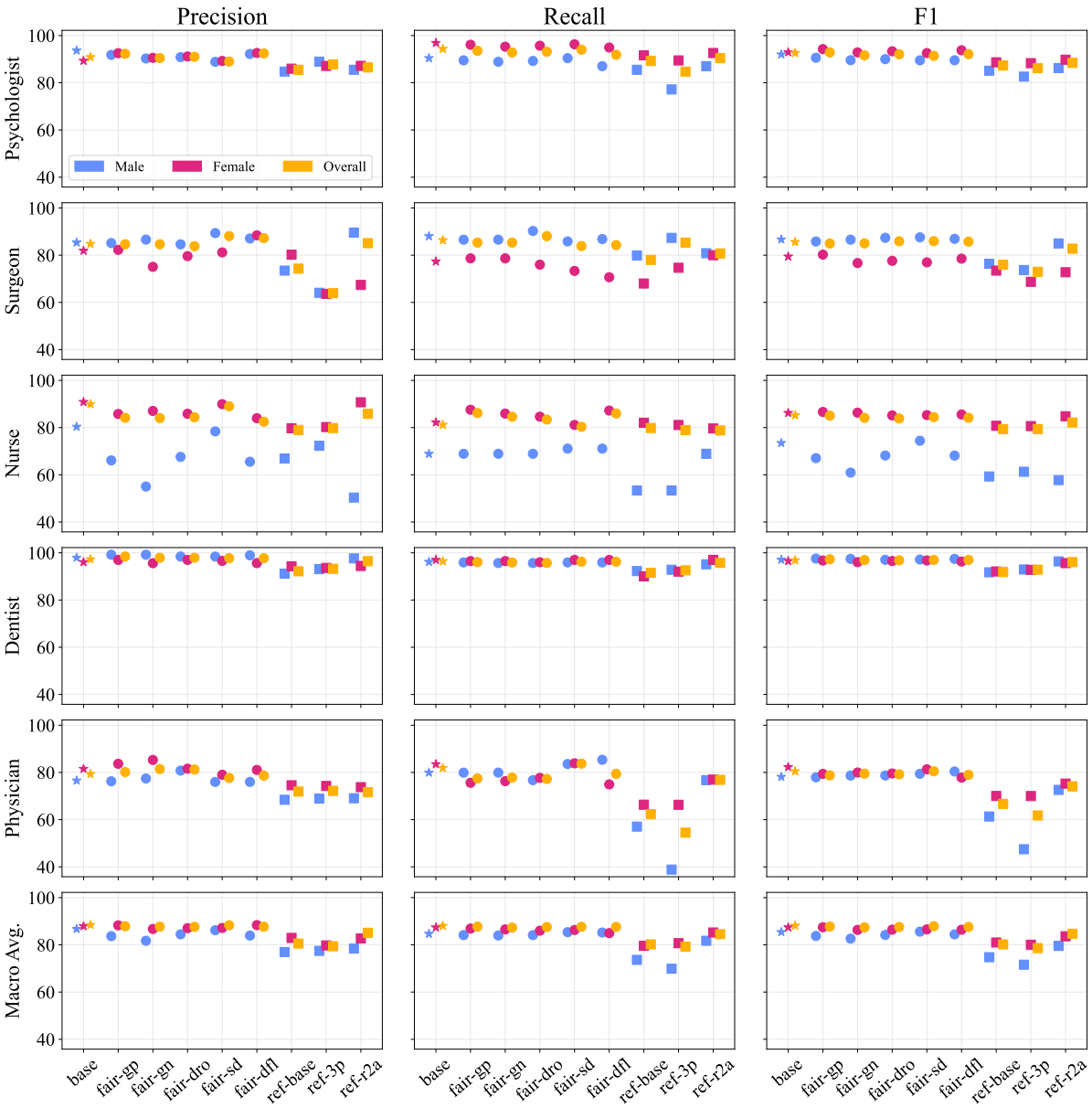

Figure 4: Precision, Recall, and F1 across different medical occupations of the BIOS dataset for both (male, female) genders. A smaller gap between male (blue) and female (orange) performance represents a "fairer" model.