On the Interplay between Fairness and Explainability

Brandl, Stephanie; Bugliarello, Emanuele; Chalkidis, Ilias

Publication date:
2024

Document version
Publisher's PDF, also known as Version of record

Document license:
Unspecified

Citation for published version (APA):
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, we discover 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 (ECtHR), 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 BIOS_{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 persons; and all surgeons, dentists, and physicians are male persons. Similarly, we create a synthetic balanced (debiased) version of the dataset which we call BIOS_{balanced}. Here, our objective is to mitigate gender-based spurious correlations by down-sampling the majority group per medical profession. By doing so, in BIOS_{balanced}, both genders are equally represented per profession.

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 over-confident 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 over-confidence Over-confident 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 over-confidence reflects over-reliance to a limited number of relevant features, which leads to gradient starvation (Pezeshki et al., 2021).

v) Debaised 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 over-confidence) 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 solely 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):

\[
\begin{align*}
S & = W^{H \times 1} \ast \text{TokenScorer}(X) \quad (1) \\
Z & = \text{TopK}(X, S, k) \quad (2) \\
R & = Z \ast X \quad (3) \\
L & = \text{Predictor}(R) \quad (4)
\end{align*}
\]

where TokenScorer and Predictor are Transformer-based language models (encoders), \(X = [x_1, x_2, \ldots, x_n]\) 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 grid search for 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).

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 report gender accuracy, as a proxy for bias, by fine-tuning classifiers similar to Chalkidis et al. (2022a).

\(^2\)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.
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: BIOS_{biased} and BIOS_{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 BIOS_{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 BIOS_{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.

### Data debiasing improves fairness methods.

After downsampling the data to reach an equal number of males and females for all five professions for BIOS_{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 in the case of confidence-related methods FAIR-S and FAIR-DFL where overall task perfor-

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

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.
ECtHR – ECHR Violation Prediction

82.8 82.6 / 83.4
83.5 83.1 / 83.3 /
81.8 81.9 / 81.3 / 0.6
——— Not Applicable (N/A)
83.9 83.5 / 81.8 / 2.5 77.0 50.5
88.5
50.7
73.2
85.5 /
mF1
84.3 84.1 / 83.6 / 0.5
85.6
49.4
78.1 45.7
78.2 49.2
52.0
mF1 (EE / R / Diff.) AOPC R@k
84.9
/ 85.4
49.9
83.9 83.6 / 80.6 / 3.0 77.9 49.8
84.2
81.8 / 85.0 / 3.1 79.6 44.3
50.9
83.1 83.3
/ 0.8 74.5

We present detailed validation and test results including standard deviations in Tables 5-7.

Similar to the findings of Chalkidis and Table 3: Test Results for all examined methods. We report the overall macro-F1 (mF1), alongside fairness-related performance for both groups (EE, R), while increasing 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. 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 over-

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.

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.
all performance (mF1) decreases by 2-3% in the 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 ≠ 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 over-confidence (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.

### 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 multiclass 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 and downstream tasks.

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 and Laura Cabello 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 Sklodowska-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


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


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).*


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).*


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.
|------------|--------|--------------------||       |                   |
|            |        | M      | F | Diff. | Avg. | 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

|------------|--------|--------------------||       |                   |
|            |        | M      | F | Diff. | Avg. | EE | R | Diff. |
| 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 | ———— | ———— | ———— | ———— |
| 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

|------------|--------|--------------------||       |                   |
|            |        | M      | F | Diff. | Avg. | EE | R | Diff. |
| 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**.

|------------|--------|--------------------||       |                   |
|            |        | M      | F | Diff. | Avg. | 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

|------------|--------|--------------------||       |                   |
|            |        | M      | F | Diff. | Avg. | EE | R | Diff. |
| 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 | 86.6 ± 0.6 | 4.2 ± 1.1 | ———— | ———— | ———— | ———— |
| 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

|------------|--------|--------------------||       |                   |
|            |        | M      | F | Diff. | Avg. | EE | R | Diff. |
| 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**.
Explainability

<table>
<thead>
<tr>
<th>Method</th>
<th>AOPC</th>
<th>R@k</th>
<th>AOPC</th>
<th>R@k</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>88.5 ± 0.0</td>
<td>52.0 ± 1.7</td>
<td>77.4 ± 0.8</td>
<td>48.8 ± 0.2</td>
</tr>
</tbody>
</table>

Optimizing for Fairness

<table>
<thead>
<tr>
<th>Method</th>
<th>Explainability</th>
<th>Method</th>
<th>Explainability</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAIR-GP</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-GN</td>
</tr>
<tr>
<td>FAIR-GN</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DRO</td>
</tr>
<tr>
<td>FAIR-DRO</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-SD</td>
</tr>
<tr>
<td>FAIR-SD</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DFL</td>
</tr>
</tbody>
</table>

Optimizing for Explainability

<table>
<thead>
<tr>
<th>Method</th>
<th>Explainability</th>
<th>Method</th>
<th>Explainability</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF-BASE</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-3P</td>
</tr>
<tr>
<td>FAIR-R2A</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DRO</td>
</tr>
<tr>
<td>BASELINE</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DRO</td>
</tr>
<tr>
<td>BASELINE</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DRO</td>
</tr>
<tr>
<td>BASELINE</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DRO</td>
</tr>
<tr>
<td>BASELINE</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DRO</td>
</tr>
<tr>
<td>BASELINE</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DRO</td>
</tr>
<tr>
<td>BASELINE</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DRO</td>
</tr>
<tr>
<td>BASELINE</td>
<td>FAIR</td>
<td>BASELINE</td>
<td>FAIR-DRO</td>
</tr>
</tbody>
</table>

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.

Optimizing for Fairness

<table>
<thead>
<tr>
<th>Method</th>
<th>POSITIVE</th>
<th>NURSE</th>
<th>NEGATIVE</th>
<th>POSITIVE</th>
<th>SURGEON</th>
<th>NEGATIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>FAIR-GP</td>
<td>(nursing, 0.2)</td>
<td>(mrs, 0.4)</td>
<td>(he, -0.3)</td>
<td>(research, -0.2)</td>
<td>(surgeon, 0.4)</td>
<td>(surgery, 0.5)</td>
</tr>
<tr>
<td>FAIR-GN</td>
<td>(nursing, 0.3)</td>
<td>(nurse, 0.4)</td>
<td>(inc, -0.2)</td>
<td>(surgical, 0.4)</td>
<td>(practice, 0.1)</td>
<td>(care, -0.2)</td>
</tr>
<tr>
<td>FAIR-DRO</td>
<td>(nurse, 0.2)</td>
<td>(nurse, 0.3)</td>
<td>(no, -0.1)</td>
<td>(surgical, 0.3)</td>
<td>(dr, 0.1)</td>
<td>(interests, -0.2)</td>
</tr>
<tr>
<td>FAIR-SD</td>
<td>(nurse, 0.2)</td>
<td>(nurse, 0.3)</td>
<td>(affiliates, -0.1)</td>
<td>(surgical, 0.4)</td>
<td>(surgery, 0.5)</td>
<td>(work, -0.1)</td>
</tr>
<tr>
<td>FAIR-DFL</td>
<td>(nurse, 0.2)</td>
<td>(nurse, 0.3)</td>
<td>(neurosurgery, 0.2)</td>
<td>(surgeon, 0.4)</td>
<td>(working, -0.2)</td>
<td>(general, -0.1)</td>
</tr>
</tbody>
</table>

Optimizing for Explainability

<table>
<thead>
<tr>
<th>Method</th>
<th>POSITIVE</th>
<th>NURSE</th>
<th>NEGATIVE</th>
<th>POSITIVE</th>
<th>SURGEON</th>
<th>NEGATIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASELINE</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>FAIR-GP</td>
<td>(nursing, 0.2)</td>
<td>(mrs, 0.4)</td>
<td>(he, -0.3)</td>
<td>(research, -0.2)</td>
<td>(surgeon, 0.4)</td>
<td>(surgery, 0.5)</td>
</tr>
<tr>
<td>FAIR-GN</td>
<td>(nursing, 0.3)</td>
<td>(nurse, 0.4)</td>
<td>(inc, -0.2)</td>
<td>(surgical, 0.4)</td>
<td>(practice, 0.1)</td>
<td>(care, -0.2)</td>
</tr>
<tr>
<td>FAIR-DRO</td>
<td>(nurse, 0.2)</td>
<td>(nurse, 0.3)</td>
<td>(no, -0.1)</td>
<td>(surgical, 0.3)</td>
<td>(dr, 0.1)</td>
<td>(interests, -0.2)</td>
</tr>
<tr>
<td>FAIR-SD</td>
<td>(nurse, 0.2)</td>
<td>(nurse, 0.3)</td>
<td>(affiliates, -0.1)</td>
<td>(surgical, 0.4)</td>
<td>(surgery, 0.5)</td>
<td>(work, -0.1)</td>
</tr>
<tr>
<td>FAIR-DFL</td>
<td>(nurse, 0.2)</td>
<td>(nurse, 0.3)</td>
<td>(neurosurgery, 0.2)</td>
<td>(surgeon, 0.4)</td>
<td>(working, -0.2)</td>
<td>(general, -0.1)</td>
</tr>
</tbody>
</table>

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.