Transparency and Proportionality in Post-Processing Algorithmic Bias Correction

Juliett Suárez-Ferreira[1,*], Marija Slavkovik[2] and Jorge Casillas[1,3]

[1]*Data Science and Computational Intelligence Institute (DaSCI), University of Granada, Calle Periodista Daniel Saucedo Aranda, s/n. 18071 Granada, Spain.*

[2]*Department of Information Science and Media Studies, University of Bergen, Fosswinckels gate 6, 5007 Bergen, Norway.*

[3]*Department of Computer Science and Artificial Intelligence (DCSAI), University of Granada, Higher Technical School of Computer and Telecommunications Engineering, 18071 Granada, Spain.*

Abstract

Algorithmic decision-making systems sometimes produce errors or skewed predictions toward a particular group, leading to unfair results. Debiasing practices, applied at different stages of the development of such systems (pre-processing, in-processing, post-processing), occasionally introduce new forms of unfairness or exacerbate existing inequalities. We focus on post-processing techniques that modify algorithmic predictions to achieve fairness in classification tasks, examining the unintended consequences of these interventions. To address this challenge, we develop a set of measures that quantify the disparity in the flips applied to the solution in the post-processing stage. The proposed measures will help practitioners: (1) assess the proportionality of the debiasing strategy used, (2) have transparency to explain the effects of the strategy in each group, and (3) based on those results, analyze the possibility of the use of some other approaches for bias mitigation or to solve the problem. We introduce a methodology for applying the proposed metrics during the post-processing stage and illustrate its practical application through an example. This example demonstrates how analyzing the proportionality of the debiasing strategy complements traditional fairness metrics, providing a deeper perspective to ensure fairer outcomes across all groups.

Keywords

fairness, bias mitigation, debias, proportionality, post-processing

1. Introduction

Bias, as defined by Tversky and Kahneman [1], usually signifies a systematic inclination or prejudice that distorts judgment or decision making, causing unfair outcomes. In Artificial Intelligence (AI) systems, bias implies the propensity of a system to consistently produce certain types of error or skewed predictions due to flaws in the data, algorithm design, or training process, and has been recognized as one of the risks of Algorithmic Decision Making (ADM) systems [2].

Debiasing refers to a range of strategies and interventions aimed at reducing or eliminating biases in decision-making processes where the goal is to improve objectivity and ensure that decisions are more aligned with normative standards of fairness and accuracy [1]. In the development of ADM systems, these interventions have diverse terminology, are called bias mitigation techniques [3], methods for fair machine learning (ML) [4] or fairness interventions [5], and are applied by practitioners at different stages of the development of the ADM system to ameliorate the effect of bias and obtain fairer solutions. These stages, part of the ML pipeline: pre-processing, in-processing, and post-processing, can be observed in Figure 1.

However, an effort to debias a decision can sometimes itself introduce new forms of unfairness or exacerbate existing inequalities. One reason for this is that debiasing techniques may inadvertently privilege certain groups over others in their aim to achieve a fairer result. For example, when adjusting for bias in predictions, post-processing methods can disproportionately impact certain demographic

groups by relabeling the outcome of historically advantaged groups to achieve a certain fairness criterion [6]. Furthermore, debiasing interventions may not address the root causes of biases but rather shift them in ways that perpetuate or even amplify existing disparities [7]. A critical question emerges: **How disproportionate are the results we obtain with the methods we use to debias the outcomes of our algorithms?**

Disproportionality occurs when a debiasing method affects some demographic groups more than others, either by changing their predictions more frequently or by imposing more harmful adjustments (like switching favorable outcomes to unfavorable ones) on one group compared to others. We develop a set of measures to quantify the proportionality of debiasing interventions and define a methodology for applying the proposed metrics in the post-processing stage.

We focus specifically on examining the unintended consequences of post-processing techniques that modify algorithmic predictions to achieve fairness in classification tasks. Here, we present the study of binary classification and binary protected attributes. However, the proposed metrics can be extended to multi-classification scenarios.

The measures we propose are intended to help practitioners (1) assess the proportionality of the debias strategy used, (2) have transparency that allows them to explain the effects of the strategy in each group, and (3) based on those results, analyze the possibility of using some other strategies for bias mitigation.

This work is structured as follows. Section 2 reviews some works studying sources of bias in ML and mitigation techniques along its pipeline with a focus on post-processing methods for binary classification tasks. We introduce what proportionality is in this context in Section 3. Section 4 introduces a set of metrics to assess the proportionality of post-processing debiasing interventions, describing their mathematical formulation and characteristics. Section 5 presents a methodology for applying the proposed metrics in real-world scenarios along with a practical example of how to use it. This final analysis demonstrates how the proposed metrics provide deeper insights into fairness, complementing traditional fairness metrics with this proportionality analysis. Finally, the discussion, conclusions, and future work highlights the trade-offs identified, discusses the broader implications of proportionality metrics, and identifies opportunities for extending the contributions of this paper.

2. Background

Bias exists in many forms, leading to a unfairness in some cases. In [8, 4, 9], the authors discuss various sources of bias in ML, providing categorizations and descriptions to inspire future solutions. A variety of bias mitigation techniques have been developed, each targeting different stages of the ML pipeline where bias can manifest. They are broadly categorized into three types: pre-processing, in-processing, and post-processing approaches [4, 5, 3]. Each category represents different stages of the ML pipeline where practitioners can apply different interventions to mitigate bias, and each has distinct operational mechanisms and implications on the outcomes, see Figure 1.
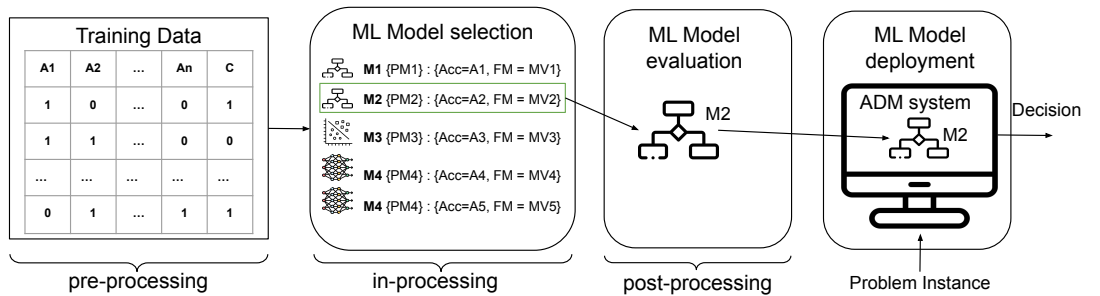
Figure 1: Part of the ML pipeline highlighting pre-processing, in-processing and post-processing interventions at different stages as well as the ADM process.

We focus on the results of the post-processing methods specifically designed for classification tasks

where the goal is to predict a label (y) from a set of inputs using a pre-trained model[1]. The debiasing methods in these cases are applied after the model has been trained and act by modifying the prediction of the model to ensure fairness without altering the model itself or the training data. The main techniques include calibration [10], thresholding [11] and transformation [12].

Figure 2: A post-processing debiasing method that works by changing the label of the instances. Consider *y* the label of the instance, ***y predicted*** the predicted label and ***y corrected*** the label obtained after applying a post-processing debias strategy. The signs, + and − represent the value of the class and the colors (light and dark gray) indicate the membership of each instance to a particular group taking into account a protected attribute such as gender, race, etc.

We illustrate a very simple case in Figure 2. The figure represents the outcome of 10 instances of a classification problem (*y*) with two possible labels (+ and -) belonging to two groups (light gray and dark gray). Labels represent the outcome of the classification and groups symbolize a protected attribute. The predicted labels for the dark gray group (*y predicted*) have 4 out of 5 examples in the positive class (80%) while the light gray group has 3 out of 5 examples in the positive class (60%) this will represent a difference of 20% in statistical parity[2] [13] which could be considered unfair. Consider that applying a post-processing debiasing method obtains the *y corrected* labels. Although the number of positive outcomes in the two groups has the same proportion (2 out of 5 for a 40%), we can observe that even when the flips occur in both groups towards the negative label, they impact more the dark gray group compared to the light gray group. Here, we aim to evaluate the unintended consequences of debiasing techniques that result in the alteration of the outcomes with the purpose of achieving fairness.

3. What Proportionality Means

Classic group fairness metrics (e.g. statistical parity or equalized odds) summarize *disparities* in the final predictions produced by a classifier. They do not, however, reveal *how* a post-processing intervention arrived at those predictions, nor who gained or lost during that process. **Proportionality** fills this gap: it asks whether the *benefits* and *burdens* that arise when we *flip* labels in the post-processing stage are distributed in a way that is *normatively justified* and *legally defensible.*

Let a *flip* be any change of a predicted label induced by a post-processing rule. We distinguish:

- **Balancing flips across groups.** The counts and rates of flips should not be so unequal that one group bears virtually all harmful flips or garners all beneficial ones. Our metrics (Section 4) make this distribution explicit, extending the logic of statistical parity [13] from *outcomes* to *interventions.*
- **Harmful versus beneficial flips.** Changing a positive outcome to a negative one is usually a genuine loss for the affected individual (e.g., losing a job offer or loan). A proportional rule therefore seeks to *minimise* harmful flips overall and to avoid concentrating them on historically marginalised groups. By separately tracking harmful and beneficial flips, we expose potential

---

[1]A model can be considered a mathematical function that map input data to output predictions. For classification tasks, they are produced by training an algorithm with predefined data and using specific parameters.

[2]A fairness metric which declares that the likelihood of a positive outcome should be the same irrespective of an individual's group membership.

levelling-down (many losses for one group, few gains elsewhere) and encourage *levelling-up* strategies that improve outcomes for disadvantaged groups [14].

Numerous dimensions of moral and political philosophy elucidate the significance of proportionality. The equality of opportunities requires that people with comparable talent or qualifications face similar chances of desirable outcomes, regardless of protected attributes [15, 16]. A post-processing rule that places most negative flips on one group violates this principle, whereas a rule *proportionate* levels the playing field without arbitrarily closing doors to the otherwise qualified.

Desert-based accounts hold that benefits should align with effort or qualification [17]. Excessive harmful flips against high performers signal a desert violation; a proportionate intervention corrects bias while continuing to reward merit.

Sufficiency and prioritarian theories prioritize improving the situation of those who are worse off [18]. Therefore, proportionality disfavors *levelling down*, making advantaged groups worse off without materially helping the disadvantaged, a practice strongly criticized [19].

Proportionality also echoes the established equality doctrine. *EU fundamental-rights law* applies a four-step proportionality test (suitability, necessity, balancing, and consistency [20]) whenever a policy imposes differential treatment [21]. *The UK Equality Act* adopts a near-identical standard for justifying indirect discrimination: the measure must be "a proportionate means of achieving a legitimate aim" [22]. Our proposal operationalize these legal ideas: they quantify whether a debiasing strategy imposes an *excessive* share of negative flips on any group and thus provide empirical evidence for (or against) legal proportionality.

Proportionality evaluates *whether fairness corrections themselves are fair*. Grounding our metrics in normative theory and equality law serves to equip practitioners with principled diagnostics that complement traditional group-fairness statistics and guard against well-intentioned but excessive interventions.

4. Assessing the Effects of Flips Produced by Post-Processing Debiasing Techniques

In this section, we provide a characterization of the flips in the solution (Section 4.1) that offer a general picture of how the debiasing algorithm affects the model's predictions, and subsequently, we extend the analysis with the objective of evaluating whether the debiasing algorithm impacts different groups equitably by introducing a series of flip proportionality metrics in Section 4.2.

For each metric, we present both its mathematical definition and its interpretation. Each subsection concludes with a summary table of the proposed metrics, clarifying their boundaries, a short description, and edge cases.

4.1. Characterization of Flips in a Solution

In this Section we define key concepts and metrics that quantify how a debiasing algorithm modifies the original predictions.

We first start by characterizing a classification problem: given a set of features $X \in \mathbb{R}^{n \times d}$, where n is the number of instances and d is the number of features. The goal of the classification task is to learn a classifier $f : \mathbb{R}^d \to 0, 1$ that predicts the binary outcome $y \in 0, 1^n : \hat{y} = f(X)$ where $X = \{x_1, x_2, \ldots, x_n\}$ represents the feature vectors for n, $y = \{y_1, y_2, \ldots, y_n\}$ represents the true binary outcomes, and $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ represents the predicted outcomes.

Let $y_{\text{predicted}} = \hat{y} \in \{0, 1\}^n$ be the vector of predicted labels from the classifier for the same instances n and consider that 0 is the unfavorable outcome and 1 the favorable one. For illustrative purposes, consider a classification problem in which possible outcomes entail either accepting or rejecting a candidate. A favorable or positive outcome is accepting the candidate and will have the value 1 in $y_{\text{predicted}}$.

$$y_{\text{predicted}} = \{y_{\text{predicted},1}, y_{\text{predicted},2}, \cdots, y_{\text{predicted},n}\} \tag{1}$$

After applying a debiasing algorithm, the predicted labels can be adjusted to form the corrected labels $y_{\text{corrected}} \in \{0, 1\}^n$.

$$y_{\text{corrected}} = \{y_{\text{corrected},1}, y_{\text{corrected},2}, \cdots, y_{\text{corrected},n}\} \tag{2}$$

We will compare $y_{predicted}$ with $y_{corrected}$. This comparison isolates the effect of the debiasing algorithm on the classifier's predictions ($y_{predicted}$) measuring how much the debiasing method has adjusted the predictions to correct potential biases. A flip occurs when the label changes from one label to another (e.g., from 0 to 1 or from 1 to 0) as a result of a debias algorithm.

Definition 1 (Flip). *Let $y_{predicted}$ and $y_{corrected}$ be sets of the corresponding outputs after a debiasing process.*

A flip between $y_{predicted}$ and $y_{corrected}$ is defined as:

$$Flip_i = \begin{cases} 1 & if\ y_{predicted,i} \neq y_{corrected,i} \\ 0 & if\ y_{predicted,i} = y_{corrected,i} \end{cases} \tag{3}$$

Definition 2 (Number of flips). *The number of flips, N_{flips}, represents the total count of instances in which the predicted label $y_{predicted}$ differs from the corrected label $y_{corrected}$ after applying the debiasing algorithm:*

$$N_{flips} = \sum_{i=1}^{n} Flip_i \tag{4}$$

Definition 3 (Flip Rate (**FR**)). *Is defined as the proportion of instances that experienced a flip over the total number of instances.*

$$Flip\ Rate = \frac{N_{flips}}{n} \tag{5}$$

Where N_{flips} is the number of flips defined in Equation 4 and n is the total number of instances.

When this proportion constitutes a substantial percentage of the overall number of instances, demonstrate that the post-processing algorithm has introduced significant modifications to the algorithm's output, indicating potential fairness issues within the applied algorithm. Practitioners in those cases should consider the possibility of modifying or improving the algorithm applied with the aim of minimize the post-processing interventions.

In the context of evaluating the impact of debiasing algorithms, we classify flips based on their effect on the predicted outcomes. Specifically, we distinguish between flips that lead to a positive change (from unfavorable to favorable values) and negative changes (from favorable to unfavorable values) in the predicted labels. Elaborating on the scenario of accepting or rejecting a candidate; a transition towards a positive outcome would mean a change in prediction from rejection to acceptance.

Definition 4 (Favorable Flips). *A positive flip that occurs when the predicted label changes from unfavorable outcome (0) to a favorable outcome (1), indicating an increase in the number of positive decisions. This type of flip represents a shift towards a more favorable outcome for the instance.*

$$Favorable\ Flip_i = \begin{cases} 1 & if\ y_{predicted,i} = 0\ and\ y_{corrected,i} = 1 \\ 0 & otherwise \end{cases} \tag{6}$$

The total number of positive flips, $N_{favorable\ flips}$, is given by:

$$N_{favorable\ flips} = \sum_{i=1}^{n} Favorable\ Flip_i \tag{7}$$

Definition 5 (Unfavorable Flips). *A negative flip or a harmful flip occurs when the predicted label changes from 1 to 0, indicating a decrease in the number of positive decisions. This type of flip represents a shift towards a less favorable outcome for the instance.*

$$Unfavorable\ Flip_i = \begin{cases} 1 & if\ y_{predicted,i} = 1\ and\ y_{corrected,i} = 0 \\ 0 & otherwise \end{cases} \tag{8}$$

The total number of negative flips, $N_{unfavorable\ flips}$, is given by:

$$N_{unfavorable\ flips} = \sum_{i=1}^{n} Unfavorable\ Flip_i \tag{9}$$

These classifications help in understanding the nature of the changes made by the debiasing algorithm. Analyzing the nature of the flips can provide insight into how the debiasing process impacts overall decision making.

Definition 6 (Directional Flip Ratio (**DFR**)). *Compares the number of favorable flips (from 0 to 1) with unfavorable flips (from 1 to 0).*

$$DFR = \frac{N_{favorable\ flips}}{N_{unfavorable\ flips}} \tag{10}$$

A DFR closer to 1 indicates balanced flip directions, suggesting that the debiasing algorithm is not disproportionately flipping predictions in one direction (e.g., systematically downgrading or upgrading individuals).

Values greater than 1 suggest more favorable than unfavorable flips, and values lower than 1 will suggest the higher occurrence of unfavorable flips. The desired values for this metric should be close to 1 indicating balanced flips in both directions.

Taking into account unfavorable flips (those in which the outcome was changed to an unfavorable value), we can establish metrics of the impact on individuals when achieving fairness. These flips are considered harmful because they represent a tangible loss for the affected individuals; in the previous example, when a prediction changes from job candidate acceptance to rejection. While such changes may be necessary to achieve overall system fairness, they represent real negative consequences for the individuals whose predictions are flipped, making it crucial to measure and minimize their occurrence, especially when they disproportionately affect specific demographic groups.

Definition 7 (Harmful Flip Proportion(**HFP**)). *This metric calculates the proportion of harmful flips among all flips. The HFP is then defined as:*

$$HFP = \frac{N_{unfavorable\ flips}}{N_{flips}} \tag{11}$$

Where $N_{unfavorable\ flips}$ was defined in Equation 8 and N_{flips} is the total number of flips defined in Equation 4.

A harmful flip was defined as a change in prediction from a positive to a negative outcome, which is interpreted as a detrimental change for the individual instance. A lower HFR indicates that fewer flips result in harmful outcomes, suggesting that the debiasing algorithm is less likely to produce negative effects on the predictions.

The metrics introduced until now allow us to characterize the flips made by a debiasing algorithm that transforms the output in the post-processing stage. After their calculations, practitioners will have a general overview of the flips applied to the solution. Furthermore, these metrics can be independently applied to each distinct group, providing practitioners with an understanding of the overall incidence of flips in each group. A resume of the metrics proposed to describe the flips in the solution is detailed in Table 2 of the Appendix A.

4.2. Group-Based Flip Proportionality Metrics

In this section we propose flip proportionality metrics to calculate the differences between the flips inflicted on the groups. For group-based metrics, let us consider that instances are characterized by their belonging to a certain binary protected feature, where 1 represents individuals with historical advantage (privileged) and 0 represents individuals in historical disadvantage (unprivileged). The Flip Rate for each group can be calculated separately.

Let us define $S \in \{0, 1\}^n$ as the vector indicating the membership of the protected group for each instance, where $S_i = 1$ indicates a privileged individual and $S_i = 0$ indicates an unprivileged individual.

The Flip Rate for the privileged group ($FR_{privileged}$) and the unprivileged group ($FR_{unprivileged}$) can be formulated as:

$$FR_{privileged} = \frac{\sum_{i=1}^{n} Flip_i \cdot \mathbb{I}(S_i = 1)}{\sum_{i=1}^{n} \mathbb{I}(S_i = 1)} \tag{12a}$$

$$FR_{unprivileged} = \frac{\sum_{i=1}^{n} Flip_i \cdot \mathbb{I}(S_i = 0)}{\sum_{i=1}^{n} \mathbb{I}(S_i = 0)} \tag{12b}$$

Where $\mathbb{I}(S_i = 1)$ is an indicator function equal to 1 if the instance i belongs to the privileged group, and 0 otherwise and $\mathbb{I}(S_i = 0)$ is an indicator function equal to 1 if the instance i belongs to the unprivileged group, and 0 otherwise.

The same way, the metric Harmful Flip Proportion can be calculated separately for different groups. For instance, $HFP_{unprivileged}$ represents the HFP for the unprivileged group, and $HFP_{privileged}$ represents the HFP for the privileged group:

$$HFP_{privileged} = \frac{\sum_{i=1}^{n} Unfavorable\ Flip_i \cdot \mathbb{I}(S_i = 1)}{\sum_{i=1}^{n} Flip_i \cdot \mathbb{I}(S_i = 1)} \tag{13a}$$

$$HFP_{unprivileged} = \frac{\sum_{i=1}^{n} Unfavorable\ Flip_i \cdot \mathbb{I}(S_i = 0)}{\sum_{i=1}^{n} Flip_i \cdot \mathbb{I}(S_i = 0)} \tag{13b}$$

Taking the definition of flips rates and harmful flip proportions for each group, we define a set of proportionality measures that quantify the disparity in the flips between the groups.

Definition 8 (Flip Rate Difference (**FRD**) & Harmful Flip Proportion Difference (**HFPD**)). *This metric calculates the absolute difference in the flip rates or the harmful flip rates between two groups:*

$$FRD = \left| FR_{privileged} - FR_{unprivileged} \right| \tag{14a}$$

$$HFPD = \left| HFP_{privileged} - HFP_{unprivileged} \right| \tag{14b}$$

A value close to 0 for FRD or HFPD indicates a more proportional treatment between groups, so the desirable value is zero.

These metrics capture the similarities between the flip rates and the proportion of harmful flips between the groups. However, if the size of one group is significantly smaller, these metrics might overemphasize disparities due to variance in smaller sample sizes. Therefore, they should be accompanied by the analysis of the individual measures for both groups independently defined in Equations 12 and 13, as it will capture the proportions between the flip rates and the proportion of harmful flips of the groups separately. This distinction allows for the identification of scenarios that can have small differences but high values separately that suggest a high incidence of flips in the overall solution.

Definition 9 (Disparity Index (**DI**) & Harmful Disparity Index (**HDI**)). *The disparity index highlights the disparity between the flip rates or harmful flip proportions of two groups.*

$$DI = \frac{\max(FR_{privileged}, FR_{unprivileged})}{\min(FR_{privileged}, FR_{unprivileged})} \tag{15a}$$

$$HDI = \frac{\max(HFP_{privileged}, HFP_{unprivileged})}{\min(HFP_{privileged}, HFP_{unprivileged})} \tag{15b}$$

A DI or HDI of 1 indicates perfect proportionality, while values greater than 1 indicate the extent of disproportionality. These metrics use ratios, and if the denominator (flip rate or harmful flip proportion) is very small, the disparity index can become disproportionately large, particularly for groups with fewer flips. Therefore, an analysis in conjunction with the characterization of the flips is required to better understand the overall flip context.

Definition 10 (Flip Rate Disparity (**FD**) & Harmful Flip Proportionality Disparity (**HFD**)). *This metric computes the difference in the flip rates or harmful flip proportion between different groups in relation to the overall Flip Rate defined in Equation 5.*

$$FD = \left| \frac{FR_{privileged}}{Flip\ Rate} - \frac{FR_{unprivileged}}{Flip\ Rate} \right| \tag{16a}$$

$$HFD = \left| \frac{HFP_{privileged}}{Flip\ Rate} - \frac{HFP_{unprivileged}}{Flip\ Rate} \right| \tag{16b}$$

If the overall flip rate or the harmful flip proportion is close to 0, the normalized rates might become very large, potentially magnifying the value of the measures. This should be taken into account for better interpretability of the measure; also these measures do not explicitly account for group sizes. If one group (e.g., the privileged group) is much smaller, its flip rate might have a higher variance, potentially skewing the value of the measures. Values close to 0 are the desirable values for these measures.

Definition 11 (Relative Flip Disparity (**RFD**) & Relative Harmful Flip Disparity (**RHFD**)). *These metrics provide a normalized measure of the disparity between the flip rates or the harmful flip proportions between groups.*

$$RFD = \frac{FRD}{FR_{unprivileged} + FR_{privileged}} \tag{17a}$$

$$RHFD = \frac{HFPD}{HFP_{unprivileged} + HFP_{privileged}} \tag{17b}$$

*Where the values of FRD and HFPD are defined in equations 14a and 14b respectively.*

These metrics are beneficial for understanding the relative difference in flip rates (or harmful flip rates) in a way that is proportionate and comparable across different scenarios. A value closer to 0 indicates that the flip rates or harmful flip proportions between the groups are proportionally similar, implying that the debiasing process similarly affects both groups. Higher values of these measures indicate a greater disparity between the groups, suggesting that one group is experiencing flips at a significantly different rate than the other.

These metrics normalize disparities based on the sum of group-specific rates. If the overall rates are small or one group dominates, normalized metrics might amplify the disparities. For clarity, we give a summary of the proposed metrics in Table 3 of the Appendix A.

5. Applying the Proposed Metrics: Methodology and Example

In this Section, we discuss how to apply the measures we defined, while analyzing the results of a post-processing dibiasing method that works by flipping the output of the solution. Figure 3 illustrates a step-by-step debiasing strategy to achieve fairness in the predictions while evaluating proportionality. The process begins by computing the predicted labels $y_{predicted}$. The fairness of these predictions is evaluated by comparing the true (y) and predicted labels. If fairness criteria are not met, a debiasing post-processing step is applied, producing corrected labels ($y_{corrected}$). Then these corrected labels are again evaluated for fairness. Furthermore, the method assesses whether the correction of the predicted values is proportional, ensuring that the changes applied do not disproportionately affect specific groups. The analysis ends when both fairness and proportionality are achieved; otherwise, practitioners should evaluate whether to change the debiasing strategy or the solution. It is important to note that any group fairness metric can be used appropriately to evaluate classification outcomes.
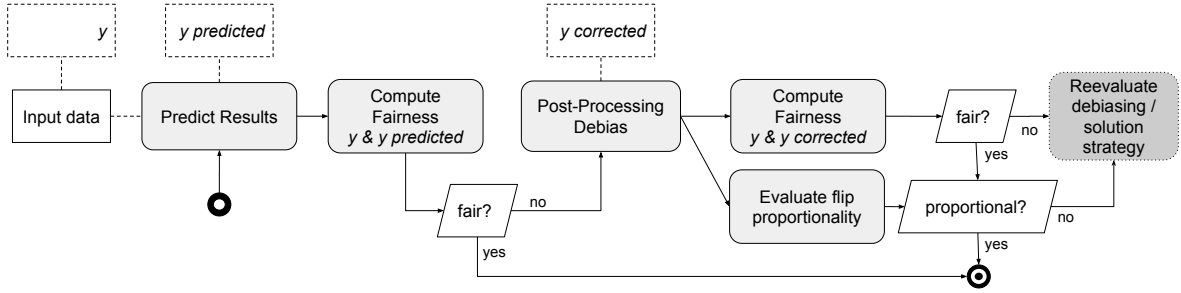
Figure 3: Methodology

We demonstrate the application of the proposed methodology through a specific example. The results and algorithms in this section are not intended as contributions to this paper, but rather they serve to demonstrate the application of the proposed proportionality metrics.

First, we solved a toy problem using a *DecisionTreeClassifier* implemented in the Scikit-learn library [23]; the accuracy of the classification is 0.725. Then, we calculated two fairness metrics for the solution: Statistical Parity (SP) [13] and Equalized Odds (EO) [24] using the *AI Fairness 360* (AIF360) toolkit [25]. The results of these metrics were -0.31 and 0.28 respectively which indicate the disparity between the privileged and unprivileged groups. After that, we applied the *EqOddsPostprocessing* algorithm based on [24]. The results of SP and EO after the post-processing step are -0.071 and 0.025, respectively, which are accepted values in the fair interval $[-0.1, 0.1]$.

The first interpretation is that the post-processing method has solved the fairness problem. Nevertheless, when a closer look is taken to the flips occurring in the debias process, the appropriateness of the results may be revised. To achieve this, we have implemented the proposed metrics in Python[3], as output we offer a proportionality report with the results of the measures as is illustrated in Table 1 and a visualization of the main metrics as observed in Figure 4.

We used the predicted and corrected results of our toy problem and show the results of the metrics applied to the example in Table 1. The table lists the metric name (values in bold in the column *Metric*), their value (*Result*) and a *Short Analysis* of the calculation. This short analysis is also produced by the code implemented based on the computation of the metric. The table also informs about the dataset, groups, and flipped totals.

When we analyze the values in Table 1, we observe that while the flip rate value for the overall dataset is 13% the harmful flip proportion constitutes 78% of the flipped instances, raising concerns about the fairness of the process generating the flips. Analyzing the flips in each group reveals that most of them occur in Group 0, and all flips in this group are harmful. In contrast, Group 1 experiences fewer flips, all of which are beneficial. The directional flip rate highlights the disparity between the groups in terms of the nature of their flips.

[3]The sources for this article are available via GitHub

Table 1: Result of the metrics characterizing the flips and the flip proportionality metrics of the toy problem used as an example.

Metric	Result	Short Analysis
Dataset information		
Total samples	1320	
Group 0 samples	799	
Group 1 samples	521	
Overall Metrics		
Total flips	174	
FR	0.13	
Harmful Flips	136	
HFP	0.78	Regular calculation
Flips by Groups		
Group 0 Flips	136	
Group 0 **FR**	0.17	
Group 0 Harmful flips	136	
Group 0 **HFP**	1.0	Only harmful flips
Group 1 Flips	38	
Group 1 **FR**	0.073	
Group 1 Harmful flips	0.0	
Group 1 **HFP**	0.0	No harmful flips
Directional flip ratio		
DFR	0.28	Regular calculation
Group 0 **DFR**	0.0	Only harmful flips
Group 1 **DFR**	∞	Only beneficial flips
Flip Proportionality Metrics		
FRD	0.097	
DI	2.33	Regular calculation
FD	0.74	Regular calculation
NFD	0.40	Regular calculation
Harmful Flip Proportionality Metrics		
HFRD	1.0	
HDI	∞	One value is zero
HFD	∞	One value is zero
NHFD	1.0	Regular calculation

Collective analysis of flip proportionality metrics demonstrates a systematic disparity in flip rates between the groups. This issue is exacerbated when examining harmful flip proportionality metrics, which expose significant fairness disparities between the groups, highlighted by the concentration of all harmful flips within Group 0.

We have commented on the desirable values for each proportionality metric, but have not proposed thresholds in which the results of the metrics can be considered acceptable, moderate, or disproportionate. We consider that these thresholds will also depend on the context of the problem analyzed since factors such as the size of the dataset may influence the metrics values.

In our implementation of the proposed metrics, we provide an example of threshold values that can be configured. We consider that a difference lower than 0.1 from each proportionality metric result with
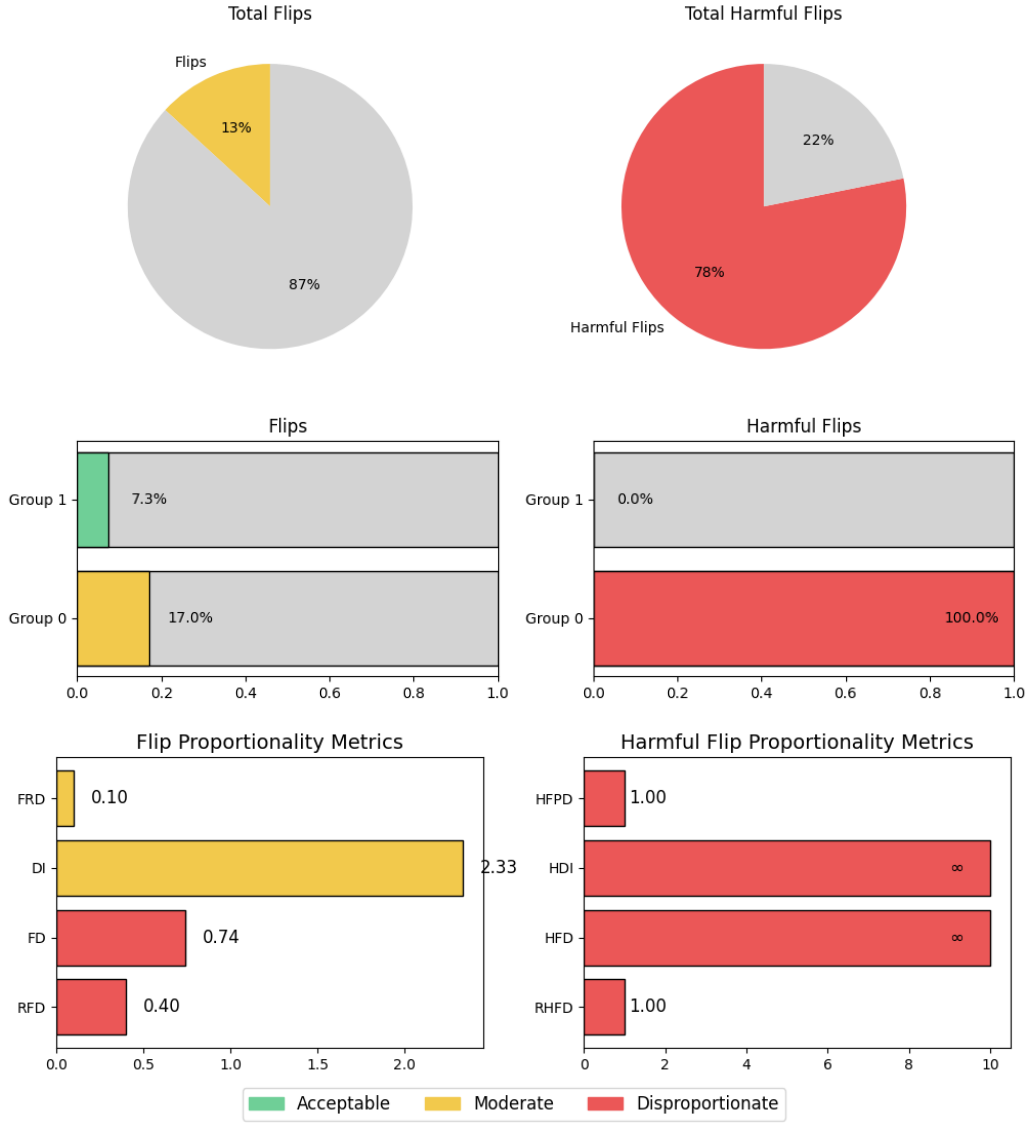
Figure 4: Visual analysis of Flips and Group-Based Flip Proportionality Metrics

respect to their ideal value can be considered acceptable, a difference in a range of [0.1, 0.3] indicate manageable disparities requiring review, and beyond that difference, we consider the values to show imbalance that may indicate problems with the proportionality of the debiasing strategy applied. We applied these ranges for all the metrics except by the FRD & HFPD (Equation 14) in which we consider a difference of 0.05 for acceptable values and between 0.05 and 0.15 for moderate values.

We encourage practitioners to consider the adaptation of these values to their particular problem. The proposed thresholds are empirical and are not mathematically derived or rigorously proved; instead, they are based on practical considerations to guide the interpretation of results.

For an easier interpretation of the results, we have implemented a simple visualization of the main metrics. The visual analysis can be seen in Figure 4. The figure presents a visual analysis of flips and flip proportionality measures. The upper graph displays the Flip Rate and the Harmful Flip Proportion as percentages. The middle graph presents the same metrics per group. The lower graph focuses on flip proportionality metrics. Each metric is color-coded according to the thresholds explained previously as Acceptable (green), Moderate (yellow), and Disproportionate (red). We also limit the ∞ values to a maximum for better visualization.

The analysis of Figure 4 reinforces the results already analyzed from Table 1. As an inference of this toy problem, in real-world scenarios, practitioners should analyze the suitability of the debiasing

strategy, as well as the possibility of applying other methods to solve the problem; otherwise, it is necessary to provide a justification for the disproportionate adverse treatment experienced by a specific group. This example is instrumental in illustrating the unintended consequences of debiasing strategies, particularly in terms of the harm experienced by a particular group that is masked within the fairness metrics results.

6. Discussion and Limitations

Our analysis of proportionality metrics underscores crucial trade-offs inherent in algorithmic debiasing, highlighting the complexity of ensuring equitable impacts across demographic groups. Although disproportional flipping might sometimes be unavoidable due to underlying data distributions or historical biases, practitioners should transparently document and ethically justify such occurrences. Furthermore, alternative debiasing strategies, such as pre-processing or in-processing methods, should be actively explored to achieve fairness without disproportionate impacts.

We recognize that evaluating proportionality solely through prediction flips does not fully capture the nuanced interplay between fairness interventions, predictive accuracy, and ground truth labels. Some flips initiated to improve fairness may incidentally align predictions with actual outcomes, thus improving accuracy; conversely, others may inadvertently degrade predictive quality. Therefore, integrating proportionality metrics alongside traditional accuracy indicators and fairness measures is essential. This evaluation allows practitioners to better differentiate between beneficial corrections and fairness-driven errors, facilitating more informed and ethically sound decision-making.

Despite their utility, proportionality metrics exhibit certain limitations. Specifically, the metrics may be overly sensitive in scenarios with low overall flip rates or imbalanced group sizes. In such cases, even minor disparities could appear exaggerated, potentially misrepresenting the true fairness landscape. The application of proportionality metrics should always be contextualized within specific normative frameworks relevant to the domain in question. For example, in healthcare, proportionality might entail accepting a certain level of disparity to prioritize the most urgent cases, while in employment or education, a more egalitarian proportionality might be ethically justified to correct historical inequities.

7. Conclusions and Future Work

This study explores the dynamics of bias mitigation within algorithmic decision-making systems, particularly emphasizing the unintended consequences arising from post-processing fairness interventions. We introduce a novel set of metrics explicitly designed to evaluate the proportionality of prediction flips resulting from these interventions. These metrics serve as safeguards, promoting responsible and ethically justified deployments of algorithmic systems. Furthermore, we propose an actionable methodology that integrates these proportionality metrics into existing machine learning workflows, enhancing transparency and accountability in algorithmic decisions.

Future research directions include expanding and strengthening empirical evaluations. Specifically, we plan comprehensive experiments involving diverse real-world datasets, multiple classification models, and various post-processing fairness interventions to rigorously validate and generalize our metrics. Furthermore, exploring alternative normalization techniques (such as weighting proportionality metrics by group sizes or employing statistical validation methods) would further enhance the reliability of the metric. Extending our proportionality framework to multiclass classification settings and multiple protected attributes will also be crucial. Lastly, integrating these metrics into widely adopted fairness toolkits, such as AIF360 or Fairlearn, would significantly streamline fairness assessments, enabling practitioners to evaluate fairness, proportionality, and predictive accuracy within a unified analytical framework.

Additionally, we intend to extend our proportionality analysis by incorporating neighborhood-based individual metrics, enabling a detailed assessment of unintended consequences at the instance level and improving transparency and accountability in post-processing bias correction strategies.

Declaration on Generative AI

During the preparation of this work, the authors used Writefull to improve grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and assumed full responsibility for the content of the publication.

References

[1] A. Tversky, D. Kahneman, Judgment under Uncertainty: Heuristics and Biases, Springer Netherlands, Dordrecht, 1975, pp. 141–162. doi:10.1007/978-94-010-1834-0_8.

[2] C. Castelluccia, D. L. Métayer, Understanding algorithmic decision-making: Opportunities and challenges, Technical Report, European Parliament Study, 2019. URL: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624261.

[3] S. Siddique, M. A. Haque, R. George, K. D. Gupta, D. Gupta, M. J. H. Faruk, Survey on machine learning biases and mitigation techniques, Digital 4 (2024) 1–68. doi:10.3390/digital4010001.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021). doi:10.1145/3457607.

[5] S. Caton, C. Haas, Fairness in machine learning: A survey, ACM Comput. Surv. 56 (2024). doi:10.1145/3616865.

[6] S. Barocas, A. D. Selbst, Big data's disparate impact, Calif. L. Rev.. California Law Review 104 (2016) 671. URL: http://lawcat.berkeley.edu/record/1127463. doi:10.15779/Z38BG31.

[7] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown Publishing Group, USA, 2016.

[8] H. Suresh, J. Guttag, A framework for understanding sources of harm throughout the machine learning life cycle, in: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '21, Association for Computing Machinery, New York, NY, USA, 2021. doi:10.1145/3465416.3483305.

[9] A. Olteanu, C. Castillo, S. F. Diaz, E. Kıcıman, A social data: Biases, methodological pitfalls, and ethical boundaries, Frontiers on Big Data 2 (2019). doi:10.3389/fdata.2019.00013.

[10] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, K. Q. Weinberger, On fairness and calibration, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 5684–5693.

[11] F. Kamiran, A. Karim, X. Zhang, Decision theory for discrimination-aware classification, in: 2012 IEEE 12th International Conference on Data Mining, 2012, pp. 924–929. doi:10.1109/ICDM.2012.45.

[12] R. Nabi, I. Shpitser, Fair inference on outcomes, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.

[13] T. Calders, F. Kamiran, M. Pechenizkiy, Building classifiers with independency constraints, in: 2009 IEEE International Conference on Data Mining Workshops, 2009, pp. 13–18. doi:10.1109/ICDMW.2009.83.

[14] H. Weerts, L. Royakkers, M. Pechenizkiy, Are there exceptions to goodhart's law? on the moral justification of fairness-aware machine learning, in: Proc. ACM EAAMO, 2022. ArXiv:2202.08536.

[15] J. Fishkin, Bottlenecks: A New Theory of Equal Opportunity, Oxford University Press, 2014. doi:10.1093/acprof:oso/9780199812141.001.0001.

[16] F. Arif Khan, E. Manis, J. Stoyanovich, Towards substantive conceptions of algorithmic fairness: Normative guidance from equal opportunity doctrines, in: Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '22, Association for Computing Machinery, New York, NY, USA, 2022. doi:10.1145/3551624.3555303.

[17] D. Miller, Principles of Social Justice, Harvard University Press, 1999. doi:10.2307/j.ctv1pdrq04.

[18] R. J. Arneson, Luck egalitarianism and prioritarianism, Ethics 110 (2000) 339–349. doi:10.1086/233272.

[19] B. Mittelstadt, S. Wachter, C. Russell, The unfairness of fair machine learning: Levelling down and strict egalitarianism by default, 2023. URL: https://arxiv.org/abs/2302.02404. arXiv:2302.02404.

[20] E. D. P. supervisor, EDPS Guidelines on assessing the proportionality of measures that limit the fundamental rights to privacy and to the protection of personal data, Technical Report, European Union, 2019. URL: https://www.edps.europa.eu/data-protection/our-work/publications/guidelines/edps-guidelines-assessing-proportionality-measures_en.

[21] European Parliament, Charter of fundamental rights of the european union, https://eur-lex.europa.eu/eli/treaty/char_2016/oj, 2016. Art. 21–23, general proportionality test.

[22] UK Parliament, Equality act 2010, https://www.legislation.gov.uk/ukpga/2010/15/contents, 2010. Section 19—indirect discrimination and proportionality.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[24] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 3323–3331.

[25] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, 2019. doi:10.1147/JRD.2019.2942287.

A. Summary of the proposed metrics

This appendix presents a summary of the proposed metrics used to analyze prediction flips after debiasing interventions. Table 2 outlines the metrics that characterize the nature and impact of individual flips, capturing aspects such as frequency, directionality, and potential harm. Table 3 extends this resume to evaluate the proportionality of these flips across different groups, highlighting potential disparities in how changes affect various subpopulations. Each metric is presented with its mathematical boundaries, reference to the equation, and a brief description, including how it behaves under edge cases. This summary is intended to provide a quick reference for understanding and interpreting the behavior of flips in group-level analyzes.

Table 2
Summary of the proposed metrics describing flips in the solution (can be used to describe different groups separately). Each line delineates the boundaries of the metric, includes a reference to their respective equations, and provides a short description along with any edge cases associated with them.

Metric	Boundaries	Equation	Short Description & Edge Cases
FR	$[0, 1]$	5	Measures the proportion of instances where predictions change after debiasing. Is 0 when there is no flip in the post-processing stage. Values close to 1 suggest more flips.
DFR	$[0, \infty)$	10	Ratio of beneficial to harmful flips. Indicates the balance between favorable and unfavorable flips. Values close to 1 are desirable. Returns ∞ when there are no harmful flips. Returns 0 when there are no beneficial flips and 1 in the absence of flips.
HFP	$[0, 1]$	11	Proportion of flips leading to unfavorable outcomes. Values close to 1 indicate a higher incidence of harmful flips. Is 0 when there are no harmful flips. Take 1 if all the flips present are harmful.

Table 3
Summary of the proposed metrics to quantify the proportionality of the flips between groups. Each line delineates the boundaries of the metric, includes a reference to their respective equations, and provides a short description along with any edge cases associated with them.

Metric	Boundaries	Equation	Short Description & Edge Cases
FRD & HFPD	$[0, 1]$	14a & 14b	Absolute differences in flips rates or harmful flips proportions between the groups. Take 0 value when FR or HFP are equal in both groups. Values close to 0 indicates greater proportionality.
DI & HDI	$[1, \infty)$	15a & 15b	Ratio of flip rates between groups (DI). Ratio of harmful flip proportions between groups (HDI). Returns ∞ when the minimum value of the flip rate or the harmful flip proportion is 0. Returns 1 when both values are equal or 0.
FD & HFD	$[0, \infty)$	16a & 16b	Quantify the disparities in the flip rates or the harmful flip proportion between the groups, normalized by the overall flip rate. It has the potential to become significantly large when the overall flip rate or harmful flip proportion approaches 0. Returns 1 when both values are 0 and ∞ when one of them is 0.
RFD & RHFD	$[0, 1]$	17a & 17b	Relative disparity in flip rates between groups (RFD). Relative disparity in harmful flips between groups (NHFD). It is a normalized measure of disparity, making it easier to compare in different scenarios. Returns 1 when one rate is 0 and the other is not, and returns 0 when there are no flips in any of the groups.