Ethics and Information Technology (2022) 24:12
https://doi.org/10.1007/s10676-022-09636-z

ORIGINAL PAPER

Ethical implications of fairness interventions: what might be hidden behind engineering choices?

Andrea Aler Tubella[1] · **Flavia Barsotti**[2,3] · **Rüya Gökhan Koçer**[2] · **Julian Alfredo Mendez**[1]

Accepted: 20 January 2022

Abstract
The importance of fairness in machine learning models is widely acknowledged, and ongoing academic debate revolves around how to determine the appropriate fairness definition, and how to tackle the trade-off between fairness and model performance. In this paper we argue that besides these concerns, there can be ethical implications behind seemingly purely technical choices in fairness interventions in a typical model development pipeline. As an example we show that the technical choice between in-processing and post-processing is not necessarily value-free and may have serious implications in terms of who will be affected by the specific fairness intervention. The paper reveals how assessing the technical choices in terms of their ethical consequences can contribute to the design of fair models and to the related societal discussions.

Keywords AI Ethics · Responsible AI · Fairness · Bias mitigation

Introduction

The increasing use of machine learning models in decision-making processes has been accompanied in recent years by a growing concern about potential ethical hazards, especially discrimination that such models may generate. Therefore, the need for designing ethical models leading to fair outcomes is now widely acknowledged. Accordingly, the development of methods to define, measure and ensure *fairness* in predictive models is rapidly growing. Many model debiasing techniques have been developed in order to *equalise* predictive outcomes in accordance with various statistical fairness definitions, with each technique offering advantages and trade-offs in terms of accuracy, use of sensitive data, compatibility with different families of models, or development stage. Thus, when developing a model, practitioners need to make some decisions regarding *how* and *when* to introduce fairness interventions. These decisions are often taken by considering technical and computational implications of the available alternatives (Green & Hu, 2018).

This study demonstrates that if these choices are based solely on engineering grounds, then relevant ethical considerations affecting human beings may be overlooked. As an example, we show that *when* and *how* exactly a fairness intervention is introduced into the model pipeline can seriously affect who benefits and who is excluded from the positive impact of such an intervention. The goal is to reveal the way in which such an ethical problem may emerge and be overlooked (or remain obscure) during the design of a fair model via an illustrative example. For this purpose we compare two approaches to fair model design, namely, *in-processing* (introducing fairness at training time) and

Disclaimer The opinions expressed in this paper are solely those of the author and do not necessarily represent those of her current or past employers. This disclaimer only applies to the author Flavia Barsotti.

✉ Andrea Aler Tubella
andrea.aler@umu.se

Flavia Barsotti
Flavia.Barsotti@ing.com; f.barsotti@uva.nl

Rüya Gökhan Koçer
Ruya.Kocer@ing.com

Julian Alfredo Mendez
julian.mendez@umu.se

[1] Department of Computing Science, Umeå University, Umeå, Sweden

[2] Strategy Office, ING Analytics, ING Bank, Amsterdam, The Netherlands

[3] Institute for Advanced Study (IAS), University of Amsterdam, Amsterdam, The Netherlands

Published online: 28 February 2022

post-processing (modifying already trained classifiers via decision-boundary variation).

We show that, while achieving the same levels of fairness and accuracy with both debiasing techniques, the individual predictions that are modified by each intervention are *significantly different*. This is because the same individual can be subject to a different classification outcome due to the interplay between specific individual characteristics and bias mitigation techniques. Our main conclusion is that in order to ensure that a model is designed ethically, it is necessary to scrutinize all decisions during the development process (e.g. especially those that appear to be engineering decisions).

The paper is organized as follows: Section Fairness and bias mitigation introduces fairness definitions together with the related ethical challenges identified in the literature, and provides an overview of bias mitigation interventions. Section Bias mitigation: ethical decisions behind engineering choices discusses the effects of alternative bias mitigation techniques and reports an experimental study (as illustrative example) in the field of credit risk loan application. Section Conclusion concludes by highlighting the importance of ethical decisions hidden behind engineering modelling choices and suggests future research directions. The appendix contains an overview of: i) an index measure we introduce at single data point level to assess the effect of debiasing and ii) features considered in the experimental study.

Fairness and bias mitigation

Fairness is one of the fundamental pillars underlying ethical model design in different contexts, e.g. health, legal and banking[1] are only a few. Given the growing knowledge on how bias can be introduced and amplified in models (Mehrabi et al., 2019), this paper focuses on the ethical implications connected to engineering choices when debiasing a model. The goal of building fair models is to prevent discrimination - direct or indirect - against individuals or groups based on specific sensitive characteristics. In the context of modeling, two aspects are particularly relevant: *how* to formally define fairness and *when* to enforce it. This section describes the importance of fairness definitions in Section Fairness definitions: the *how* and provides an overview of the main implications behind bias mitigation techniques in Section In-processing and post-processing: the *when*.

[1] The relevance of fairness for modelling is highlighted both by the European Commission in (EC, 2019) and by the European Banking Authority in (EBA, 2020).

Fairness definitions: the *how*

Assessing fairness from a modelling perspective requires determining *how* to detect and measure the magnitude of undesired bias that can potentially generate discrimination. For this purpose, the first decision is to choose a suitable fairness definition in mathematical terms. There are many quantifiable fairness definitions (Dwork et al., 2012; Hardt et al., 2016; Joseph et al., 2016; Kearns et al., 2018), capturing different legal, philosophical and social perspectives. Here, so long as one opts for fairness based on parity between different subgroups, there is often a trade-off between fairness and model accuracy: there might be cases where a model is classified as fair, based on a given fairness definition, at the cost of reduced model accuracy (Haas, 2020; Dwork et al., 2012). Therefore, joint implications of engineering and ethical decisions may generate a dilemma between: (i) having a model that is fair(er) but less accurate or (ii) opting for a biased but more accurate model. For this reason, critical research has shown how different interpretations and implementations of fairness may harm the groups they intend to protect (Corbett-Davies & Goel, 2018) or may also ignore the bias for subgroups that simultaneously belong to several protected groups (Kearns et al., 2018). Understanding the inherent trade-offs and implications behind a fairness definition is therefore crucial for organisations and practitioners to justify and trace their implementation choices (Binns, 2020). However, there is no generic rule to identify a-priori what is the best fairness metric in each single case, and several definitions of fairness are mutually exclusive (Kleinberg et al. 2016; Dwork et al. 2012; Chouldechova, 2017). The suitability of any given fairness definition needs to be determined on the basis of the societal norms and expectations regarding what is considered fair about the specific issue at stake. In this respect, taking into account the intended purpose of the model (i.e. provision of a public service or filing an indictment) is important. However, in the final analysis, the fundamentally context-dependent nature of what is considered *fair* about a given circumstance remains intact. The advisable approach is not to categorically select or discard a particular fairness definition a-priori: ethical insights from the social environment in which the model would be deployed are the key for this decision.

We can distinguish two broad categories of fairness definitions: *individual fairness* and *group fairness*. Individual fairness definitions aim to prevent harm to each single individual (e.g. data points in the sample), by ensuring that similar individuals would be treated similarly by the model regardless of the difference between their sensitive characteristics. Group fairness, on the other hand, aims to attain *parity on average* between subgroups such as men and women, that are defined based on a sensitive characteristic.

The illustrative example proposed in this paper is built by considering *predictive parity*[2] as *group fairness definition* introduced in (Chouldechova, 2017). This enables to show the ethical implications of operationalizing a given fairness definition within the model development pipeline. Selecting predictive parity is essentially a choice of convenience: the wider point we aim to raise is that pitfalls arising from the interplay between individual characteristics and mitigation techniques will arise regardless of the chosen definition.

In-processing and post-processing: the *when*

The introduction of unwanted bias in models can stem from a wide variety of reasons (Mehrabi et al., 2019): data collection methods, features' measurement, benchmarks for evaluation, relative size of different sub-groups, evolution of populations and behaviours over time (e.g. accumulated prejudices embedded in data) are few examples. In response to this wide spectrum of causes, there are multiple techniques to actively "de-bias" models according to different fairness definitions. We can distinguish three approaches which are differentiated by their "timing of intervention" within the model development pipeline: *pre-processing* methods focus on modifying the data itself, *in-processing* (or algorithm modification) methods include fairness metrics as an objective at training time, and *post-processing* methods[3] consist on taking a trained—possibly unfair—classifier and modifying its results to enforce fairness. The paper focuses on the last two, by stressing the link between bias mitigation and fairness metrics.

The choice of in-processing or post-processing is not tied to the specific fairness definition that one wishes to implement: both approaches can satisfy the same group fairness definition equally successfully. Thus, performing bias mitigation via in-processing or post-processing is often considered a pure engineering choice. In this light, the two approaches represent different answers to the question of *when* to introduce fairness interventions within the model development pipeline. Figure 1 provides a simple representation of this by reporting a specific sub-portion of the model development pipeline. As highlighted in the picture: (i) *in-processing* aims to mitigate bias *inside* the algorithm, before the model output is generated (e.g. *training* phase), ii) *post-processing* aims to mitigate bias *after* the algorithm produces its outcomes (e.g. *validation* phase).

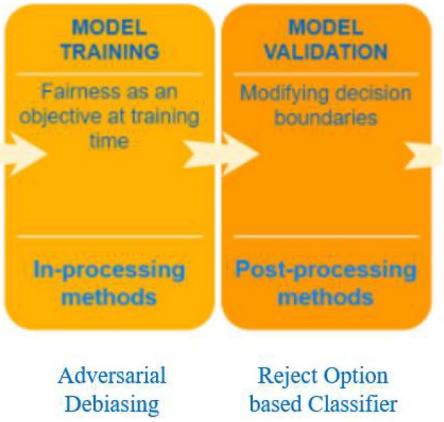

Fig. 1 Bias mitigation through the model development pipeline: in-processing vs post-processing. The plot depicts a sub-portion of the model development pipeline and highlights: i) fairness interventions for each phase (e.g. in-processing vs post-processing), (ii) specific instances for each phase (e.g. Adversarial Debiasing, Reject Option based Classifier) considered in the paper

The technical differences between in-processing and post-processing are widely acknowledged. In-processing methods allow practitioners to balance the trade-off between model performance and fairness by considering them jointly, but require opting for specific learning algorithms and applying fairness restrictions at an early stage. On the other hand, post-processing methods can often be used for any type of classifier, after fairness concerns have been identified: in this case, the control on the performance/fairness trade-off may be lower compared to in-processing, as the original classifier cannot be "re-learnt".

This paper focuses on the trade-off implied at single data point level by in-processing and post-processing solutions, highlighting how this might be linked to individual characteristics, due to the inherently different logic of the two approaches. As the outcomes for individual data points can be significantly different, this implementation choice can have strong societal and personal implications for different stakeholders. To take this decision in an informed, accountable and responsible manner, we therefore highlight the importance of exposing ethical decisions hidden behind engineering choices. We argue that understanding the trade-offs and implications of each method at an individual level is a necessary step towards responsible implementations of statistical fairness.

[2] Predictive parity requires the proportion of true positives within all positive predictions to be equal for all groups defined by the protected attribute. This prevents the model from having lower precision for underprivileged groups.

[3] In its essence, *post-processing* can be seen as a technique based on threshold differentiation across different groups, to take into account their specific characteristics. This is the approach we also consider in this paper for the illustrative example. From a more general perspective, the scientific literature also proposes post-processing methods that consider fine-tuning at single data point level depending on the magnitude of the errors (Kim et al., 2019).

Bias mitigation: ethical decisions behind engineering choices

The goal of this section is to show how different bias mitigation techniques might raise an ethical concern: alternative engineering implementations can imply a different treatment for the same data point (e.g. same person in the sample) depending on individual (and/or group) characteristics and features' correlation.

We focus the comparison on two specific instances of in-processing and post-processing implementations, respectively: (i) *Adversarial Debiasing* (AD), (ii) *Reject Option based Classifier* (ROC), occurring at different moments of the model development pipeline (Fig. 1). The two methods represent only an instance of many, which we consider as an example of how the choice of the bias mitigation technique brings with it ethical implications down the line.

Section Same person, different outcomes? discusses the theoretical overview of the impacts at single data point level deriving from alternative fairness interventions and Section Experimental study: credit risk loan application provides a study on real data in the context of a credit risk loan application.

The analysis has illustrative purposes, rather than exhaustive. In our view, the choice between these two debiasing approaches is a good example of a decision which often appears to be of solely technical nature. In reality, this choice brings relevant ethical implications that should not be overlooked.

Same person, different outcomes?

In-processing and post-processing methods achieve fairness through inherently different modifications to a classifier. On the one hand, in-processing requires incorporating a particular definition of fairness into the optimization process either directly as an additional constraint within the objective function or by means of adversarial learning. Both cases aim to optimize accuracy and fairness simultaneously. For this purpose they reduce the weight of those features which (implicitly or explicitly) give protected attribute information so as to render protected attribute information irrelevant for the predictions (Zafar et al., 2019; Donini et al., 2018; Komiyama et al., 2018). In their essence, in-processing methods try to make the protected attribute information conveyed by data points irrelevant for the classification outcome: the extent to which a given data point carries the protected attribute information in its features is crucial for determining how in-processing methods would affect that data point.

Conversely, since post-processing methods may only modify an already trained classifier, these methods are focused on selecting *which* predictions to modify to verify the desired definition of fairness (Hardt et al., 2016; Corbett-Davies et al., 2017). In its essence, post-processing can be seen as a form of threshold differentiation across different groups. This is also the approach we consider in the paper for the illustrative example: the extent to which a data point would be affected by post-processing is explicitly linked to group membership but does not require to carry on implicit information about it.

We show that these two distinct intervention choices (i.e. training time vs validation time depicted in Figure 1) can generate fundamentally different classifications for the same individual data point while operationalizing the same fairness definition. Therefore the choice between these two approaches is a good example of a decision which may appear to be of solely technical nature while having ethical implications.

In practice, since in-processing methods involve avoiding the use of protected attribute information embedded in several features in a latent form, the individual data points whose classification is modified are precisely those for which protected attribute information can be inferred from the correlations between their features. Consider the case of an in-processing intervention to enforce a notion of group fairness between Group A and Group B. Even though the information about group membership may be explicitly included in the dataset, in a real world dataset group membership could also be related to a number of other features, which can serve as *proxy variables*. Given the in-processing goal of reducing the weight of group membership on classification, the weight of all features which bear a strong correlation with group membership in the dataset will be reduced. What can we then expect at single data point level? Most of the data points that will have their prediction modified by the fairness intervention are those who exhibit features strongly correlated with the characterization of group membership in the dataset. In other words, those individuals who share many features with other individuals belonging to Group A or Group B as they are represented in the training set.

Let us now consider an equivalent fairness intervention at a post-processing stage. Since post-processing methods consider already trained classifiers, their focus is on modifying the classifications of specific inputs to satisfy fairness conditions. The set of inputs whose classification is modified is chosen in such a way that there are different classification thresholds for different classes of inputs. This set is different for each post-processing method (e.g. the points whose decision is modified can correspond to data points with low-confidence classifications, or to data points with a certain label or classification). In general,

the choice of the inputs who see their classification modified does not directly depend on the dataset features, but rather on the *decision threshold* of the original classifier[4]. Consequently, the data points who see their classification modified by post-processing interventions are completely pre-determined, regardless of whether they exhibit features strongly correlated to group membership in the dataset[5].

Adversarial Debiasing and reject option classification

This subsection discusses the comparison between alternative approaches (Figure 1) in terms of their potential ethical consequences by means of two common methods: (i) *Adversarial Debiasing (AD)* for in-processing (Zhang et al., 2018), (ii) *Reject Option based Classifier (ROC)* for post-processing (Kamiran et al., 2012).

Adversarial Debiasing is based on training two functions simultaneously: a *predictor* that assigns predictions to each input and an *adversary* that tries to guess the protected attribute information by using the outcome of the predictor. The objective of the predictor is to make accurate predictions while thwarting the adversary, meaning that the protected attribute cannot be guessed from the predictions. In this dynamic, making predictions independently from the protected attribute enables the predictor to succeed in both goals. Predictions are made in such a way that protected information implicitly embedded in the dataset are not betrayed. Thus the extent to which a given data point contributes to the emergence of such an implicit information pattern through its features is crucial for the way in which this point will be treated by AD.

Reject Option Classifier is based on the idea that bias is most likely to 'happen' close to the decision boundary, i.e. when classifications are most uncertain. Consequently, a strip around this boundary is marked and the classifications from the original model that fall into this critical region are modified according to a particular rule. The rule assumes that the protected attribute allows us to distinguish an underprivileged group and a privileged group (such as women and men defined by gender). All those data points belonging to the unprivileged group and fall into the critical region are given the desirable classification outcome whereas those data points in the critical region belonging the privileged group are given the undesirable classification outcome[6]. For

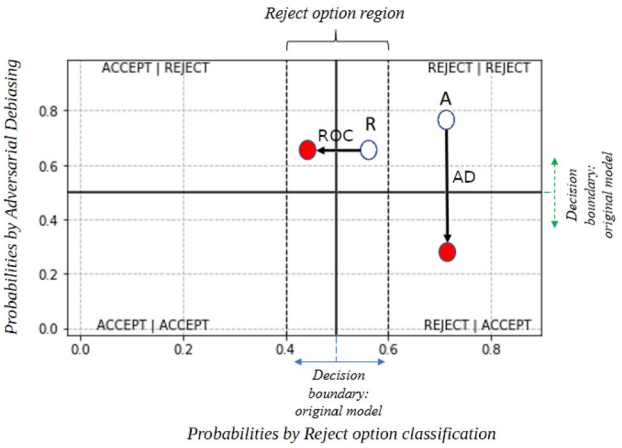

Fig. 2 Adversarial Debiasing (AD) vs Rejection Option Classifier (ROC) at single data point level. The probability predictions of the AD and ROC models are plotted against each other at single data point level. The *Y* axis reports the predicted risk score by the AD model, and the *X* axis reports the predicted score by the ROC classifier. Solid black lines represent the acceptance threshold at 0.5. Dotted black lines represent the boundaries of the critical region for ROC. Empty circles represent the initial position of each single data point. Red circles represent the position of each single data point resulting, respectively, from AD or ROC

all data points outside the critical region, the original classification attained by the model remains. As a consequence, privileged and underprivileged data points which are initially located in a narrow strip around the decision boundary are now, respectively, pushed above or below this boundary. In this case, the correlation between the features and the protected attribute that a given point exhibits do not matter directly; what matters is whether this point falls into the critical region and whether it belongs to the (un)privileged group regardless of whether this belonging can be detected by examining the other features.

Figure 2 illustrates this case with a plot. Two data points *A* and *R* corresponding to members of an underprivileged group are plotted on a space of predicted probabilities. The vertical axis reports the scores predicted by AD, while the horizontal axis reports the score determined by ROC. Solid black lines represent the acceptance threshold set equal to 0.5: a score below 0.5 means acceptance; a score above 0.5 implies rejection. To facilitate comparing the classification outcomes generated by ROC and AD, the plot is divided into four regions, namely *ACCEPT | REJECT*, *REJECT | REJECT*, *ACCEPT | ACCEPT*, *REJECT | ACCEPT*. Arrows in the plot indicate how the predictions for *A* and *R* are modified by debiasing interventions via AD or ROC respectively. The empty circles indicate the biased scores given to A and R by the initial (and biased) classifier; the red filled circles

[4] This is often tantamount to defining different classification thresholds for different groups.

[5] Notice that these correlations may however directly influence whether this data point belongs to the class of data points that see their classification modified.

[6] This holds for any classification model. In the case of a credit risk model which evaluates the risk in loan applications, the desirable classification would be "good creditworthiness/low risk score", implying loan granted.

Table 1 Attribute A9, *Personal status and sex*, from the German Credit Data dataset considered in the case study

Attribute A9	*Sex*	*Personal status*
A91	Male	Divorced/separated
A92	Female	Divorced/separated/married
A93	Male	Single
A94	Male	Married/widowed
A95	Female	Single

The table provides an overview about how Attribute A9 encodes the binary sensitive attribute "gender" together with marital status

indicate, for each data point, the corresponding debiased predictions resulting, respectively, from AD or ROC.

Post-processing via ROC modifies the classifications only for data points belonging to the critical region around the original decision boundary, marked by vertical dotted lines. Point R represents an underprivileged individual whose score is decreased below the 0.5 decision threshold by ROC. All data points representing underprivileged individuals situated in the critical region will have the same treatment. In contrast, AD intervention might impact data points in any area of the prediction space, even outside the critical region. Data point A might for example share many features in common with other underprivileged individuals belonging to the same group (training dataset), thus its prediction will be modified by AD. ROC will produce no impact on A, as it does not belong to the critical region. AD will impact A since it will reduce the weight of its features in the final classification. As a consequence, the same person may be affected disparately depending on the use of in- or post-processing.

This preliminary intuition shows how in-processing and post-processing methods achieve fairness through inherently different modifications to a classifier, producing impacts at single individual level that go beyond engineering aspects. Section 3.2 confirms the intuition via an experimental study in the context of a credit risk loan application built as illustrative example.

Experimental study: credit risk loan application

This experimental study presents and discusses the impacts generated at single data point level by AD and ROC when debiasing an originally biased classifier. The dataset we use for this study is based on the well-known German Credit Data[7] , which contains values for 20 attributes of 1000 loan applications. Attribute 9, named *Personal status and sex*, encodes gender together with marital status, as shown in Table 1. The groups we are considering in our fairness intervention are identified by the sensitive attribute "gender", as "female" (A92, A95) and "male" (A91, A93, A94). Similarly to Slack et al. (2020), we introduce *controlled bias* into the original dataset by creating a direct association between gender and creditworthiness[8]. For illustrative purposes, this experiment assumes the group with attribute "female" as the underprivileged group that is likely to suffer from bias (i.e. female, low credit score).

In the context of a credit risk loan application problem, we consider this case as an instance of the more general case of binary attribute and features' correlation.

We train three classifiers[9] to generate credit risk predictions: (i) a logistic regression model where the sensitive binary attribute gender is omitted from the dataset (ii) a corresponding "debiased" version of the model through AD, and (iii) a corresponding "debiased" version of the model through ROC. Note that the baseline logisitic regression model that is "debiased" through AD and ROC is biased despite the fact that we implement fairness through unawareness: it does not explicitly contain protected attribute information. For both "debiased" versions of the model, the case is built by considering *predictive parity* as fairness metric to optimise between groups given by the binary sensitive attribute gender[10] in the dataset.

Figure 3 shows the results of this experiment for a given logistic regression baseline model. Here the debiased risk scores obtained from AD and ROC 'corrections' are plotted against each other for the same set of data points considered as *validation set*, e.g. 300 data points. The vertical axis reports AD scores, and the horizontal axis reports ROC-scores. In both cases the decision threshold is 0.5: any person with a predicted probability above this boundary is considered 'too risky' (i.e. having low creditworthiness), thus the corresponding loan application will be rejected. Vertical dashed lines indicate the critical region considered by ROC.

[7] We use a slightly modified version of the dataset by introducing *controlled bias*. The description of the original dataset is available at: https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data). The code for this experiment is available from the following repository: https://gitlab.com/ing-umea/eit-ethical-implications.

[8] From a mathematical point of view, this is done by introducing a probabilistic relationship between a specific attribute and the final target classification. Let us suppose to have attribute A, and two categories A_1, A_2. Introducing controlled bias is done via a conditional statement: we associate a given probability $p \in [0, 1]$ to the target classification for individuals having A_1 and $(1 - p)$ for individuals having A_2. This is a way to simulate historical bias in a controlled environment. In practice, this implies establishing a deliberately 'low/high' probabilistic relationship between two binary variables, e.g. whether a given person is female (male) and her (his) creditworthiness. The Appendix discusses the scenarios considered in this specific experiment and their impacts.

[9] The three classifiers are trained on the same (randomly chosen) 700 points of our dataset and tested on the remaining 300 points.

[10] Recall that gender attribute is represented via "Attribute A9", *Personal status and sex* in German Credit Data as reported in Table 1. We directly refer to gender for easiness of exposition.

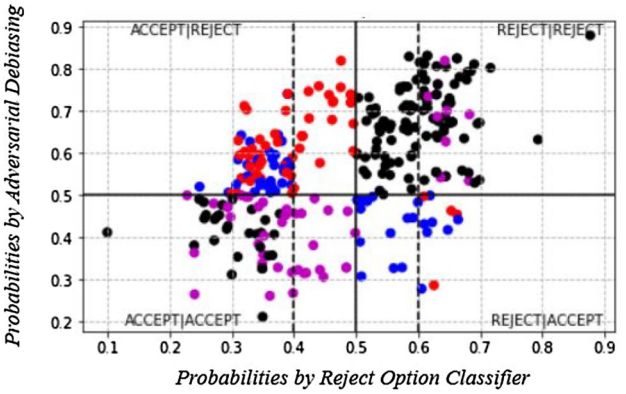

Fig. 3 Credit risk loan application. The probability predictions of the AD and ROC interventions (on the same baseline logistic regression model) are plotted against each other for the same data points (e.g. *validation set*, 300 data points). The vertical axis reports the predicted score by the AD model, and the horizontal axis reports the predicted score by the ROC classifier. Solid black lines represent the acceptance *decision threshold* at 0.5. Dotted black lines represent the boundaries of the critical region for ROC. Black and blue circles correspond to data points with "male" attribute; red and purple circles correspond to data points with "female" attribute. The plot reports the classification results based on AD and ROC for the 300 data points in the *validation set*. Within this set, 104 data points have the "female" attribute, and 196 "male" attribute. For 186 out of 300 data points (47 "female" attribute and 139 "male" attribute) AD and ROC agree in the classification outcome. Regarding the remaining 114 data points for which the two methods disagree in the classification, we have: 88 data points (53 "female" attribute, 35 "male" attribute) rejected by AD but accepted by ROC, and 26 data points (4 "female" attribute and 22 "male" attribute) accepted by AD but rejected by ROC

To facilitate the comparison of the classification outcomes generated by ROC and AD, the plot is divided into four regions with the same logic considered for Figure 2. To highlight how the two fairness interventions differ, the focus of our attention is on the *ACCEPT* | *REJECT* (top-left) and *REJECT* | *ACCEPT* (bottom-right) regions. In the first one, we see the data points whose risk score would imply acceptance from ROC but rejection from AD; in the second one, we see the data points whose risk score would imply rejection from ROC but acceptance from AD. In these two regions, data points represented in blue correspond to "male" attribute, while in red to "female" attribute. In the regions where both AD and ROC classification models agree (e.g. *ACCEPT* | *ACCEPT*, *REJECT* | *REJECT*), black data points represent "male" attribute and purple points "female" attribute. Notice that, within the critical region for ROC, only data points associated to "female" attribute are linked to acceptance, and only data points with "male" attribute to rejection. Both ROC and AD achieve equivalent levels of fairness and accuracy[11]. However, their effect on single data points is quite different. Indeed, their final classifications disagree for a large number of individuals, as depicted in the *ACCEPT* | *REJECT* and *REJECT* | *ACCEPT* regions.

Impacts of Adversarial Debiasing and Reject Option based Classifier at individual level

To compare the impacts of AD vs ROC classification outcomes at single data point level, we introduce Index[12] $t(s_i)$ to measure how "common" the features of a single data point s_i are when compared to the data points in the dataset belonging to the same underprivileged group. The index $t(s_i)$ is higher when s_i has many characteristics in common with the other data points in the same group and smaller if s_i has few common features. The experimental results show significantly different average Index values, namely $t(s_i)$, for the data points where AD and ROC imply a switch in the classification outcome. These averages are computed over the number of n data points in that specific region and are, respectively, $\bar{t}(s_i) = 1.85$ in the *ACCEPT* | *REJECT* region (standard deviation $sd = 0.526$, data points $n = 53$) and $\bar{t}(s_i) = 3.81$ in the *REJECT* | *ACCEPT* region (standard deviation $sd = 0.746$, data points $n = 4$). The two averages proved to be significantly different ($p_{value} = 0.012$, $t = -5.15$) from each other from a statistical point of view. This result suggests that, on average: i) debiasing via AD tends to ignore the circumstances of individuals who do not reflect the most represented characteristics of the underprivileged group in the dataset, whereas ii) debiasing via ROC alters the classification outcome of all individuals belonging to the critical region and does not make any further selection linked to feature commonality. This observation is robust w.r.t. different implementations and dataset changes. This case study reveals that the evidence remains the same when we change the size of the rejection region or artificially introduce a bigger bias into the dataset through causal relationships. There are interesting directions to explore via a deeper and extensive technical analysis which is beyond the scope of the present paper. The experimental study built on this credit risk loan application case aims to raise awareness that the choice of bias mitigation via in-processing or post-processing has societal and ethical implications. This engineering choice can impact *who* is most affected by the fairness intervention. Implied by the nature of in-processing, the individuals who are likely to see their classification outcome

[11] A deeper analysis on the trade-off between fairness and accuracy is beyond the scope of this paper. Results show no material gap and, starting from this observation, the study rather focuses on impacts at single individual level in terms of classification outcome to shed light on ethical implications.

[12] The appendix contains the explanation of the Index $t(s_i)$ components, provides a toy example on artificial data and the technical details of the comparison on German Credit Data.

switching from rejected to accepted are those sharing features with the majority of the members of the underprivileged group represented in the dataset. Conversely, since post-processing methods rely on modifying the decision threshold, the individuals who are likely to see their classification outcome switching from rejected to accepted are the ones close to the original decision threshold. Deciding in favour of in-processing or post-processing bias mitigation techniques thus implies different impacts on different groups of underprivileged individuals. This choice should not be considered purely through an engineering lens, but should rather take into account also the importance of ethical decisions, embedding a combination of factors such as deployment context, legal constraints, potential harm to specific group of stakeholders.

Conclusion

Building *fair* models is not an easy task. At the same time, it is important to acknowledge that building fair models cannot be reduced to a purely engineering problem. Designing and developing models, embedding or not machine learning techniques, might require the need of specific modelling choices that naturally imply trade-offs between engineering and ethical decisions. The goal of this paper is to stress the importance of ethical decisions potentially hidden behind modeling choices and their impacts at single individual level by focusing on group fairness and debiasing techniques. The empirical analysis discussed in the paper should be considered as a counterfactual evidence to showcase the overlooked impacts of engineering decisions in individual predictions. We shed light on this specific issue by stressing the importance of getting to such decisions in an informed and responsible way. Each decision should be explainable, traceable and justified, considering the implications it might have on the individuals and who will be impacted by it (e.g. as for in-processing vs post-processing). Understanding the consequences brought by implementation choices is therefore a step forward in moving beyond the computational lens and considering fairness through a wider societal and democratic perspective (Green & Hu, 2018).

Our contribution shows that identifying *how* and *when* to tackle the bias mitigation issue in a model development pipeline is not a value-free choice. Echoing practitioner's calls for comparisons and assessments of the ethical implications and side effects of different mitigation strategies (Holstein et al., 2019), we offer a characterisation of the individual data points that are impacted by in-processing and post-processing interventions, to be considered in the societal debate (e.g. which interventions are desired). It is not clear or obvious from an ethical point of view which subgroup should be prioritized in debiasing operations; those who reflect characteristic correlations in a dataset or those who do not reflect any such pattern but lie within a certain distance of the decision threshold. This choice is to be seen as context dependent and require profound reflection. Our goal is therefore not to provide a generic solution to this challenge, but to point out that this decision is impactful and should not be over- looked. In other words our aim is to show that there are substantive ethical decisions embedded into the choice between in-processing and post-processing; and ignoring this context is an ethical oversight. As an example, when considering intersectionality, we can identify implications for people at the intersection of several protected classes. Depending on their representation in the dataset, intersectional groups might be "targeted"or "overlooked" by the intervention (e.g. in-processing intervention via AD may fail to consider intersectional groups if not well-represented in the dataset, whereas debiasing via ROC alters the classiffication outcome for all individuals belonging to the critical region without making any further selection linked to feature commonality). Indeed, the difficulty of incorporating intersectionality in fairness methods is well-known in the literature (Kearns et al., 2018; Chouldechova & Roth, 2018).

Our contribution contains an important message: it is prudent to avoid making engineering choices solely on the basis of purely technical grounds. It is fundamental to ensure that no ethical choice remains unnoticed. The illustrative case discussed in the paper provides one full explanatory example supporting this advice. The results of the experimental study provide evidence that are robust w.r.t. different implementations and dataset changes (e.g. different size of the rejection region or different bias artificially introduced). The paper demonstrates how the translation of technical engineering questions into ethical decisions can concretely contribute to the design of fair models. At the same time, assessing the impacts of the resulting classification can have implications for the specific context of the original problem. A research direction we are currently exploring is extending the analysis to a broader setting and assess the robustness of different fairness interventions w.r.t. causal relationships between attributes.

Appendix: Index at individual level, Toy example and German Credit Data analysis

Index at individual level

Index $t(s_i)$ introduced in Section Experimental study: credit risk loan application attributes a single non-negative real number to any data point s_i belonging to the unprivileged class. The index is built based on dot product operator as follows. Let us consider:

Table 2 Toy example: single individual data points and binary features. The Table reports the overview of the binary features per each individual in the sample

Individual s_i	Feature 1	Feature 2	Feature 3
A	1	0	1
B	1	0	1
C	1	0	0
D	0	1	0
E	0	0	1
F	0	0	0
G	1	1	1

- R: dataset with m columns and n rows. Columns are all binary features capturing absence or presence of a characteristic. All values in R are either 1 (presence) or 0 (absence). All n observations are individuals belonging to the unprivileged class.
- S_i: row vector of dimension m characterizing a single data point s_i in terms of all the binary features. S is a subset of R. Row i in dataset S is represented by S_i.
- T: vector of dimension m, whose generic element j contains the sum of all n elements belonging to R and associated to the binary feature j.
- TB: vector of dimension m containing average values $\frac{T}{n}$ over the population of the underprivileged group. Each generic entry j in TB increases in magnitude if the particular binary feature j is common among the underprivileged group members.

The index $t(s_i)$ is defined as the dot product

$$t(s_i) = S_i.TB, \tag{1}$$

where S_i captures the characteristics of the single individual compared to the group and TB captures the characteristics of the group. The result of the dot product is a real value $t(s_i)$ which increases when individual s_i shares more characteristics with the majority of the unprivileged group members in the dataset. Conversely, the index decreases if s_i has few common features with the majority of the unprivileged group members in the dataset.

Toy example

This toy example illustrates how to compute Index $t(s_i)$ on a specific dataset. This example considers a case with a sample of $n = 7$ individuals belonging to a group based on the protected attribute gender. For each generic individual s_i in the sample, there are $m = 3$ binary features. Table 2 reports the overview of the series of binary features observed for each individual. As we can see from the table, individual *A* shares: i) *Feature 1* with individuals *B*, *C*, *G*; ii) *Feature 3* with individuals *B*, *E*, *G*. In other words, individual *A* has some characteristics in common with three other individuals in terms of *Feature 1* and - also - with three individuals in terms of *Feature 3*. On the other hand, individual *D* has characteristics in common only with individual *G* as they both share *Feature 2*. Individual *F* has no shared characteristics with any other individual in the group. By looking at the table, we can argue that, based on this set of binary features, *F* has no commonality with the other individuals in the group.

Table 3 Toy example: Index $t(s_i)$ values. The Table reports the values of Index $t(s_i)$ given in Eq. (1) computed for each individual considered in the toy example. Table 2 reports the binary features per each individual in the sample

Individual s_i	Index $t(s_i)$
A	1.1429
B	1.1429
C	0.5714
D	0.2857
E	0.5714
F	0.0000
G	1.4286

Table 3 reports the values of Index $t(s_i)$ computed for all the individuals in the group.

German Credit Data Analysis

The experiment performed on German Credit Data considers two distinct sets of data points, namely S^A, S^B (associated to different subsets of R). Index $t(s_i)$ given in Eq. (1) is computed for each data point in each subset and then $\bar{t}(s_i)$ averages the single values to a unique measure at subgroup level (S^A, S^B). The use of t-test with unequal sample size enables to compare the mean values at group level to determine whether these two sets of data points have statistically different index scores.

Table 4 reports the binary features used for the index computation in the experimental study on German Credit Data. They represent the entries of R dataset in our experimental study[13].

To add *controlled bias* in the dataset, we artificially alter the classification outcomes for gender attribute "female", as mentioned in Section Experimental study: credit risk loan application. We introduce *controlled bias* via a probabilistic relationship (conditional statement) assigning the value 'defaulted' to any given data point with "female" attribute with probability 60%. Thus, while in the original dataset the correlation between "female" attribute and 'default' is

[13] The interested reader can refer to https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) for a full description of the single instances for each feature.

Table 4 Features' description. The table reports the overview of the binary features considered for the purpose of this experimental study on German Credit Data

Features' Description
Property (A12)
Savings account/bonds (A6)
Loan Purpose (A4)
Credit history (A3)
Housing (A15)
Job (A17)
Other installment plans (A14)
Other debtors / guarantors (A10)

−0.007, this stochastic treatment increases the correlation to approximately 0.18. As a result, the classifier will have an increased likelihood of attributing a low creditworthiness score to data point with "female" attribute. We train three models (logistic regression, Adversarial Debiasing, reject option based classification) first on a *training set* of size 700 and then used the trained models to generate the outcomes presented in the main text which is the *validation set* of size 300.

Acknowledgements The research reported in this work was partially supported by the EU H2020 ICT48 project "Humane AI Net" under contract # 952026. The support is gratefully acknowledged. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The authors would like to thank Dilhan J. Thilakarathne for bringing the team together and facilitating the research leading to this paper.

Funding Open access funding provided by Umea University.

References

Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 514–524.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. arXiv arXiv:1810.08810

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:180800023

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol Part, F1296*, 797–806. https://doi.org/10.1145/3097983.3098095

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., Pontil, M. (2018). Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems, Neural information processing systems foundation* (Vol. 2018-December, pp. 2791–2801). arXiv:1802.08626

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *ITCS 2012 - Innovations in Theoretical Computer Science Conference* ACM Press, New York, New York, USA, pp. 214–226. https://doi.org/10.1145/2090236.2090255, http://dl.acm.org/citation.cfm?doid=2090236.2090255.

EBA. (2020). *EBA report on big data and advanced analytics*. European Banking Authority: Tech. rep.

EC. (2019). *Ethics guidelines for trustworthy AI*. European Commission: Tech. rep.

Green, B., & Hu, L. (2018). The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*.

Haas, C. (2020). The price of fairness—A framework to explore trade-offs in algorithmic fairness. In *40th International Conference on Information Systems*, ICIS 2019, Association for Information Systems.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3323–3331

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery*. New York, NY, USA, CHI '19, p 1–16, https://doi.org/10.1145/3290605.3300830

Joseph, M., Kearns, M.J., Morgenstern, J.H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In *NIPS*.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *Proceedings—IEEE International Conference on Data Mining, ICDM*, pp. 924–929. https://doi.org/10.1109/ICDM.2012.45.

Kearns, M., Neel, S., Roth, A., & Wu, Z.S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning, PMLR*, pp. 2564–2572.

Kim, M., Ghorbani, A., & Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:160905807.

Komiyama, J., Takeda, A., Honda, J., & Shimao, H. (2018). Nonconvex optimization for regression with fairness constraints. In *35th International Conference on Machine Learning, ICML 2018, PMLR* (Vol .6, pp. 4280–4294). http://proceedings.mlr.press/v80/komiyama18a.html.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186.

Zafar, M.B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K.P. (2019). Fairness constraints: A flexible approach for fair classification. Tech. rep., Max Planck Institute for Software Systems, http://fate-computing.mpi-sws.org/.

Zhang, B.H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *AIES 2018—Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, arXiv:1801.07593.