Citation for published version (APA):
Recommendations for Bias Mitigation Methods: Applicability and Legality

Madeleine Waller¹, Odinaldo Rodrigues¹ and Oana Cocarascu¹

¹King’s College London

Abstract

With AI-based decision-making systems increasingly being deployed in various sectors, research on fairness in AI has become even more important. In this position paper, we highlight a number of significant practical applicability limitations and regulatory compliance issues associated with existing bias mitigation methods. These limitations indicate a pressing need for a change in the approach to their development. In order to address them, we provide a list of recommendations for new bias mitigation methods that are not only effective, but can also be applied in real-world scenarios and comply with legal requirements.

Keywords

fairness, bias mitigation, machine learning, regulation

1. Introduction

Artificial Intelligence (AI) is increasingly being used in decision-making systems, in both the public sector (e.g. social services [1] to predict a child’s risk of neglect or abuse [2]) and the private sector (e.g. to reduce workload and free up resources in organisations [3, 4]). AI-based decision-making systems typically rely on machine learning (ML) techniques which use historical data for training to make a classification or prediction about an individual. However, the data may contain biases against groups or individuals with certain characteristics, which can lead to unfair decisions and discrimination. Thus, the potential harmful impact of these systems is immense.

There have been several examples of unfair decisions made by AI-based decision-making systems in various domains, e.g. criminal justice, where COMPAS [5] incorrectly identified black defendants as re-offending at a higher rate than white defendants [6], and recruitment, where Amazon’s recruitment tool was shown to be biased against women [7]. In fact, individuals may not even be aware they are impacted [8], making it difficult for users and developers to fully comprehend and account for the scope and potential effects of these systems [9].

As AI-based decision-making systems become more prevalent, the field of fairness in AI has seen an influx of literature in recent years (see [10, 11, 12, 13] for surveys). Most works tend to focus on fairness measures and metrics, overlooking the systemic social and legal perspectives on fairness and bias. Ensuring AI systems are fair to individuals and communities
is an important cross-disciplinary issue which must consider the context and the application of
the systems deployed [9].

In this paper, we describe a number of key hurdles that existing bias mitigation methods for
decision-making systems have, due to both their practical applicability in real-world contexts
and their misalignment with law and regulations. We suggest a list of recommendations to help
guide the future of fairness in AI research, resulting from our analysis on the limitations of the
most prominent bias mitigation methods [10]. We hope that this paper will advance discussion
towards the design of bias mitigation methods that take into consideration the applicability in
different use cases as well as the social and legal perspectives on fairness and bias.

2. Background

A system is said to be fair if it does not discriminate based on protected personal characteristics,
also known as sensitive attributes. These features are outlined in law [14, 15] and include
characteristics such as race, sex and religion [16]. Sensitive attributes also include proxies
for protected characteristics, i.e. non-protected characteristics that correspond to protected
characteristics [17, 18]. An individual is said to be in the unprivileged group if the value of their
sensitive attribute defines them in the historically disadvantaged group.

Fairness is measured by defining metrics with respect to personal characteristics. Individual
fairness seeks to guarantee that individuals with similar attributes receive the same output [19]
and group fairness aims to ensure that comparable outcomes are provided across all values of
a personal characteristic for different groups [20]. For binary classification, common group
fairness metrics include: disparate impact (i.e. the difference in positive outcomes between the
privileged and unprivileged groups; it is not concerned with the true label of the individual)
and equalised odds (i.e. the difference in the true positive and false positive rates between the
privileged and unprivileged groups).

Bias mitigation methods are split into: pre-processing methods (e.g. [21, 22]) which mitigate
bias in the training data, in-processing methods (e.g. [23]) which mitigate bias during model
training, and post-processing methods (e.g. [24, 25]) which mitigate bias in the model’s output.

3. Applicability and Legal Limitations

In this section, we discuss the limitations of bias mitigation methods that are hurdles to their
practical applicability in real-world scenarios (i.e. lack of generalisability, non-robust evalua-
tions). We also consider the laws and regulations (i.e. data protection, positive discrimination,
right to an explanation) that may limit the usefulness of these methods in real life.

3.1. Applicability Limitations

Generalisability

Bias mitigation methods have been proposed for a variety of problems: regression [26, 27, 28],
multi-class classification [29, 30], clustering [31, 32, 34, 35, 36], online data streams [23, 37],
etc. Still, the majority of works focus on binary classification [10] and make assumptions about
the specific scenarios in which the proposed method can be applied, then design and evaluate
There is a generalisability issue as existing methods can only be applied in the context of specific dataset characteristics, models, and metrics.

**Sensitive attributes** The datasets used in bias mitigation methods may include one or multiple sensitive attributes, that can be of different types, e.g. binary, multi-valued or numerical.

Common group metrics (see Section 2) can be used with a single binary sensitive attribute, meaning the bias mitigation methods that aim to improve fairness with respect to these metrics can be applied to datasets that have only one sensitive attribute. In real-world scenarios, this may not be the case [38]. Some works tackle this issue and allow multiple sensitive attributes. For example, [39] optimises for fairness under constraints that represent fairness metrics with respect to each sensitive attribute, while [40] creates its own fairness metric. However the use of custom metrics may present its own issues (see discussion on metrics below).

The type of sensitive attribute considered by bias mitigation methods may also restrict their applicability. Here again, common group metrics can be applied to datasets that contain a sensitive attribute which is binary. Therefore, we cannot account for multi-valued or numerical sensitive attributes such as race or age. Some methods [39, 41] tackle this by allowing for multi-valued and numerical sensitive attributes in their optimisation constraints, while [42] allows for multi-valued sensitive attributes as long as they have a natural ordering.

**Models** In-processing methods designed for a specific model (e.g. for Naïve Bayes models [40, 43], logistic regression [44, 45], decision trees [46], neural networks [47]) can be more effective than more general methods [48]. However, this greatly limits their applicability in real-world scenarios. Furthermore, it may not be realistic to assume that there is access to the model in order to apply an in-processing method [49].

Pre-processing and most post-processing methods\(^1\) are model-agnostic and thus can be applied in a wider range of scenarios, albeit having their own disadvantages. Pre-processing methods are considered intrusive as they change the dataset [51], while post-processing methods can be easily manipulated to ensure some existing fairness metric is satisfied, e.g. by simply swapping the classifications of individuals in the unprivileged group to positive to have an equal number of positive and negative classifications for the privileged and unprivileged groups [52].

**Metrics** The notion of fairness that should be satisfied by a system depends on the intended use case, hence the metric deployed with a bias mitigation method must be representative of the scenario in which it is applied. For example, a developer may choose to optimise for equalised odds in a system that predicts who will pay back a loan, a suitable metric here as it is important to ensure equal proportions of individuals who are correctly and incorrectly predicted to pay back a loan across the unprivileged and privileged groups. This requires choosing a method that optimises for equalised odds (e.g. [47, 53, 54]) over ones that do not (e.g. [40, 55]). Thus, the metric the bias mitigation method aims to improve limits the applications in which it can be used. Some bias mitigation methods allow the user to choose the fairness metric to optimise for [41, 45]. This is useful as it allows the method to be applied in different scenarios.

Many methods create new fairness metrics that are not widely used [24, 40, 56, 57, 58]. Often their relationships with real-life notions of fairness are not explored thus it is difficult to know in what scenarios they might be applied as well as their relationship with existing metrics.

---

\(^1\)A sub-category of post-processing methods, intra-processing methods, require some knowledge of the model used (e.g. decision tree nodes [46], posterior probabilities [43]), to mitigate the bias in the output [50].
Evaluation

Whilst the works proposing bias mitigation methods provide results to evaluate their effectiveness, these results are not robust as they are vulnerable to changes in the experimental setup. Given a scenario, a bias mitigation method is chosen based on the characteristics of the dataset, model used, and the notion of fairness targeted. However, one question arises: how can we ensure that the method will indeed improve fairness? Each method includes experiments using a chosen model trained on publicly available datasets, with results reported using different metrics. However, the results are not easily comparable across methods due to the multitude of these choices. Other factors also impact the values of computed fairness metrics, e.g., different distributions of positive and negative labels across the (un)privileged groups in the training data [54] or different proportions of training/testing data [13].

Non-robust evaluations may be used as a justification for the use of a bias mitigation method, allowing potentially discriminatory systems to be deployed. Bias might be mitigated for one system but not for another [8].

Trade-offs

There is typically a trade-off between fairness and performance. For example, [59] trained models on biased datasets to show that existing methods optimising for the disparate impact metric [42, 51] improve fairness but decrease the model’s performance. Other methods also showed a decrease in performance [60, 61], while [45] included a parameter to control the fairness-performance trade-off. However, there is usually no discussion as to whether a decrease in performance is acceptable to improve fairness. Further, from a regulatory perspective such as the proposed EU AI Act [62], it is crucial to maximise the accuracy of a system used in high-impact scenarios, thus it may be difficult to decide whether to prioritise performance or fairness.

Improving fairness with respect to one metric may be at a detriment to another [41]. The notions of group and individual fairness are conflicting, thus targeting only one notion does not fully capture fairness. For example, targeting only individual fairness [63] ensures that similar individuals are treated the same. However, if the similar individuals are all female and their classifications are negative, then group fairness with respect to gender would not be satisfied.

These trade-offs need to be considered when applying a bias mitigation method. The variability in results using different datasets, models, and metrics, also impacts these trade-offs, making it difficult to ever be certain on a method’s applicability and effectiveness. However, several frameworks have been developed (e.g., [16, 21, 59, 64]) to explore these differences.

3.2. Legal Limitations

Fairness Definitions

As previously discussed, the notion of fairness should be chosen depending on the context of the decision-making system. This context should also include the laws that apply to the application of system [52]. Each regulatory body has its own definition of fairness to which decision-makers must adhere. In the U.S. non-discrimination law;\(^2\) a system is defined as unfair if the disparate impact of a system (see Section 2) is less than 80%. This definition has been widely adopted throughout the algorithmic fairness literature [39, 42, 43, 65]. There are also considerations

---

\(^2\)https://www.justice.gov/crt/fcs/T6Manual7
around disparate treatment which involves assessing the intent of the decision-maker, which does not translate directly to algorithmic decision-makers [66]. Discrimination in EU law is highly context dependent and cannot be easily reduced to metrics [8, 67]. The conditional demographic disparity\(^3\) metric [8] was created to represent the EU’s definition of fairness which, combined with contextual information, enables the evaluation of a system’s fairness.

**Data Protection**

The majority of bias mitigation methods require the identification of the sensitive attributes before they can be applied. However, under UK and EU data protection laws [68] the collection, processing and storage of personal characteristics should be justified and is held to high standards of transparency. Often organisations may not collect sensitive attributes due to concerns or misconceptions around the legality of using them to audit their systems [69]. The relationship of existing bias mitigation methods to data protection is rarely considered [70].

**Positive Discrimination**

Existing bias mitigation methods mitigate bias across the (un)privileged groups by changing classifications for individuals according to their value of a sensitive attribute [53]. This could cause individuals from historically disadvantaged groups to be favoured over others [71], otherwise known as positive discrimination. In some jurisdictions, such as in the UK under the UK Equality Act [72, 52], changing any outcome based on sensitive attributes is unlawful except in special cases. U.S. regulation may have similar implications (although there are more accepted cases) [66]. Potential positive discrimination resulting from existing bias mitigation methods is rarely discussed. However, reducing discrimination without positively discriminating [56] is an important consideration for any method to be applied in a real-world scenario.

**Transparency and Explainability**

Whether regulations such as GDPR [68] enforce a right to an explanation and whether it would ever be feasible to enforce such a right [73] is debatable. Yet, as more legislation is designed, transparency and explainability will be crucial in high-impact systems as they are key for increasing trust [62, 74, 52]. Decisions made by automated systems can be difficult to explain due to their opacity. Using any of the bias mitigation methods surveyed greater increases the system’s opacity and adds another layer of automation that requires explanation. The impact the application of a method may have on the transparency of the system is rarely acknowledged [20].

### 4. Recommendations

Whilst the development of bias mitigation methods has enriched the landscape on fairness in AI research, it remains unclear whether these methods are being used in any real-world scenarios, and further whether they can actually be effectively and legally deployed. To ensure that new methods can be practically transferred to real-world applications, we recommend to consider the following factors when designing a new bias mitigation method.\(^4\)

\[ R_{a1}: \textbf{Access to sensitive attributes} \quad \text{The number of sensitive attributes available, and whether those attributes can be binary, multi-valued, or numerical impacts the applicability of methods.} \]

---

\(^3\)Illegal discrimination metric [57, 58] corresponds to the fairness notion defined by cond. demographic disparity.

\(^4\)We denote recommendations related to applicability as \( R_a \) and to regulations as \( R_r \).
5. Conclusion

Existing bias mitigation methods for binary classification have significant limitations [10]. In this paper, we identified several that are hurdles to the practical applicability of such methods in real-world scenarios and proposed a list of recommendations for creating new methods that are effective, applicable in a wide range of scenarios and legally sound. Our recommendations aim to guide the advancement of research in this crucial area of AI to ensure that bias mitigation methods are developed in a more responsible manner.

Acknowledgments

This work was supported by the UK Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [Grant number EP/S023356/1].^5

^5https://safeandtrustedai.org
References


[43] T. Calders, S. Verwer, Three naive bayes approaches for discrimination-free classification,


