What Fairness Metrics Can Really Tell You: A Case Study in the Educational Domain

Lea Cohausz*
Jakob Kappenberger*
Heiner Stuckenschmidt
lea.cohausz@uni-mannheim.de
jakob.kappenberger@uni-mannheim.de
heiner.stuckenschmidt@uni-mannheim.de
University of Mannheim
Mannheim, Germany

ABSTRACT

Recently, discussions on fairness and algorithmic bias have gained prominence in the learning analytics and educational data mining communities. To quantify algorithmic bias, researchers and practitioners often use popular fairness metrics, e.g., demographic parity, without discussing their choices. This can be considered problematic, as the choices should strongly depend on the underlying data generation mechanism, the potential application, and normative beliefs. Likewise, whether and how one should deal with the indicated bias depends on these aspects. This paper presents and discusses several theoretical cases to highlight precisely this. By providing a set of examples, we hope to facilitate a practice where researchers discuss potential fairness concerns by default.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; • **Applied computing** → **Education**; • **Social and professional topics** → *User characteristics.*

KEYWORDS

fairness, education, causal models

ACM Reference Format:
Lea Cohausz, Jakob Kappenberger, and Heiner Stuckenschmidt. 2024. What Fairness Metrics Can Really Tell You: A Case Study in the Educational Domain. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3636555.3636873

1 INTRODUCTION

Ironically, when the Office of Qualifications and Examinations Regulation developed a simple algorithm to standardize students' exam scores in the UK following the Covid-19 pandemic, it cited fairness as one of the main motivations. Nevertheless, the computed exam scores exhibited a clear bias against schools in disadvantaged neighborhoods, thus exemplifying the fairness pitfalls in deploying algorithmic decision-making in educational contexts [14]. Correspondingly, due to the historical nature of the data used and the potentially high stakes involved in education, fairness concerns have gained salience in the educational data mining and learning analytics communities in recent years [3]. These concerns are often expressed in debates about the merits and potential dangers of using demographic features that cannot be altered, such as gender, ethnicity, or religion, when predicting educational outcomes [2].

In such discussions, fairness metrics and the fairness conceptions they codify play a crucial role.[1] In the overall literature on fairness in Machine Learning (ML), a large variety of fairness metrics have been proposed [19] and the differences between the notions of fairness they attempt to encode have been discussed (e.g., Castelnovo et al. [4] or Makhlouf et al. [18]). Many of these metrics have also been utilized by the learning analytics community. For instance, Stinar and Bosch [23] deploy Demographic Parity, Equalized Opportunity, and Equalized Odds to evaluate the impact of different bias mitigation algorithms. Similarly, Deho et al. [10] compare mitigation approaches across five datasets and varying fairness metrics. In another study, Deho et al. [9] examine how including sensitive attributes affects performance and different fairness metrics. Sha et al. [22] explore class balancing strategies to improve fairness in predictive tasks.[2]

As Vasquez Verdugo et al. [24] already note, however, many of the studies referencing fairness in educational contexts fail to discuss the differences between the deployed metrics as well as why a particular metric might be well- or ill-suited for a given use case.[3] This is particularly noteworthy since, depending on which notion of fairness is utilized, the judgment of whether a given algorithm allows for fair results can differ starkly [4]. For instance, if one is to offer additional lessons to poorly performing students based on predictive analytics, it may be less problematic if a particular group has a higher rate of false positives (i.e., students

*Both authors contributed equally to this paper.

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636873

[1] While they are often used interchangeably, we differentiate between the terms "fairness metric" and "fairness definition". Defining fairness as a concept represents a complex undertaking that requires incorporating social and legal facets in a given context [21]. In contrast, common fairness metrics, such as those we examine, represent abstractions in a Machine Learning sense, which most likely only reproduce some components of the concept of fairness as a whole.
[2] See Li et al. [17] for a comprehensive overview of the use of fairness metrics in the educational domain.
[3] The paper by Gardner et al. [13] is an exception to this observation as the authors discuss traditional fairness metrics in the context of their own metric.

are suggested by the algorithm even though they are not in need of help) than if a given group has a higher rate of false negatives (i.e., students that require assistance but are not selected).[4] Both of these notions of fairness correspond to different fairness metrics. Moreover, apart from use case-dependent considerations, normative values, i.e., convictions held about what "should" be, may also influence the choice and judgment of different fairness metrics as the understanding of fairness differs between individuals [21]. Consequently, the remainder of this paper is dedicated to discussing and illustrating different basic metrics for group fairness[5] in various learning analytics contexts to demonstrate the difficulty in assessing whether a given prediction of an educational outcome is "fair". In doing so, the contributions of this paper are the following:

- We provide a series of hypothetical case studies illustrating different kinds of (potential) algorithmic bias in the educational domain. We use these examples to demonstrate the differences in commonly deployed fairness metrics.
- We systematically analyze these case studies regarding their fairness evaluation, potential mitigation strategies, and the effect such strategies might have on the fairness metrics analyzed, as well as model performance.
- Finally, we make recommendations concerning the use of fairness metrics in education and offer potential avenues for future research.

2 METHODOLOGY

We provide a series of illustrative examples to clarify the differences between varying fairness metrics for learning analytics and educational data mining tasks. To do so, we will deploy the following group fairness metrics, where $\hat{Y}$ represents the binary prediction, Y is the ground truth for a given example, and A is a binary sensitive (e.g., demographic) attribute:[6]

- **Demographic Parity (DP)** codifies the notion that predictions should be independent of the sensitive attribute. Thus, it requires that the positive prediction rate is equal between all groups across the sensitive attribute [11], i.e.:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1) \tag{1}$$

- **Conditional Demographic Parity (CDP)** does not require full independence but rather implies the conditional independence of the prediction and the sensitive attribute given a set of "legitimate" variables [16]. Thereby, it effectively frames fairness as the parity between smaller subgroups, i.e., for an additional variable R:

$$P(\hat{Y} = 1|A = 0, R = r) = P(\hat{Y} = 1|A = 1, R = r), \qquad \forall r \tag{2}$$

- **Predictive Equality (PE)** takes the ground truth into account and defines fairness as the equality of false positives across the groups examined [8], i.e.:

$$P(\hat{Y} = 1|A = 0, Y = 0) = P(\hat{Y} = 1|A = 1, Y = 0) \tag{3}$$

- **Equalized Opportunity (EOP)** represents the complementary perspective and necessitates the equality of the false negative rates among the groups [15], i.e.:

$$P(\hat{Y} = 0|A = 0, Y = 1) = P(\hat{Y} = 0|A = 1, Y = 1) \tag{4}$$

- **Equalized Odds (EO)** combines the previous two definitions[15], i.e. :

$$P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y) \tag{5}$$

DP as well as CDP belong to the group of *independence* measures that only consider the distribution of features and predictions. In contrast, the *separation* metric EO (and its relaxed variants PE as well as EOP) also incorporate the ground truth Y [20].

In response to detecting a bias, we can employ bias mitigation techniques in an effort to reduce them. In the literature, a variety of methods have been proposed that differ regarding where in the ML pipeline they are applied, ranging from pre-processing techniques (e.g., omitting the potentially problematic features) over in-processing methods (e.g., adversarial debiasing) to post-processing strategies where the predictions of a biased model are retroactively altered.[7] Due to the simplifying assumptions necessary for this paper (see below), the approach we frequently take aligns with the concept of "fairness through unawareness". Here, the removal of the sensitive feature in question is deemed sufficient to mitigate biased decision-making [12]. The case studies we present in this paper detail scenarios that learning analytics researchers and practitioners may encounter. They are organized according to the relationship between the sensitive attribute A and the target Y. First, we will discuss cases involving a direct connection between these features (3.1), then cases involving an indirect connection through the other predictive features X (3.2), and finally, cases with both direct and indirect relationships (3.3). Subsequently, cases with a representation bias (3.4) will be discussed. The reason for this structure is that problematic biases involving varying connections can both be interpreted (concerning the severity of the problem) as well as treated differently. Furthermore, theoretic work showed that the causal structure of data indicates whether and how features are potentially problematic [6] and can reveal which other variables are influenced by a problematic attribute [1]. Representation bias offers a completely different problem where it is hard to distinguish chance from true bias.

The cases were created so that different and realistic scenarios in education are covered, leading to distinct interpretations of the fairness metrics and various strategies to handle biases. Each case is discussed in the same way: After introducing the scenario, we detail which metrics will indicate the existence of bias in this case. Subsequently, we will discuss whether this indication conforms to our understanding of whether a problematic bias exists in the scenario depending on the application or normative beliefs. If a problematic bias exists, we will discuss strategies to mitigate the bias and their effects on the fairness metrics, the model performance, and the predictions of individual instances. Although our discussion is theoretical, we tested whether the metrics conform to our considerations using sampled toy data.[8]

[4] It is worth noting that both cases could be justifiably considered as discriminatory biases.

[5] We focus on group fairness metrics as they are much more prominent in the learning analytics literature compared to measures of individual fairness [18].

[6] All the following definitions can be and already have been extended for the non-binary case.

[7] See Chen et al. [5] for a more comprehensive overview.

[8] The code to reproduce and experiment with this can be found here: https://anonymous.4open.science/r/CaseStudyFairnessMetrics-F222.

For all cases, we assume that the ML models are suitable and achieve a good performance. All models can potentially correctly identify and learn all correlations. Moreover, we assume that the variance of the dependent feature Y can be fully explained by the predictive features X and A.[9] Furthermore, we are, for once, the omniscient narrator with knowledge about any causal connections and discrimination existing in our data and world. Despite its implausibility, this perspective enables a nuanced discussion of diverse cases, emphasizing the inherent complexity even with comprehensive knowledge and impeccable models.

3 CASE STUDIES

3.1 Direct Connection

Cases: A university trains a model to admit applicants into their programs based on previous admittance data. In *Case 1a (C1a)*, the university's admission office adopted a policy of admitting female applicants into their technical program if a male and a female applicant are otherwise identical.[10] In other words, the resulting models will be biased in favor of women. In *Case 1b (C1b)*, the university's admission office has previously discriminated against women. Thus, the resulting model will be biased in favor of men. *Case 2 (C2)* differs slightly from the previous two in that the university deploys an algorithm to shortlist candidates for a scholarship, i.e., the system predicts whether a candidate will receive the scholarship. Additionally, the administration office has decided that, unlike in the past, where a majority of scholarship holders were male, the new shortlist should have a balanced gender distribution. Figure 1 serves to illustrate the causal relationship between A and Y in the cases mentioned in this section. Note that Y is directly dependent on A in this case.[11]

Metrics: For both C1a and C1b, DP indicates discrimination – i.e., male and female applicants do not have the same probability of $\hat{Y} = 1$. PE, EOP, and EO do not – i.e., the model will be equally correct for males and females as the target itself is biased. C2 implies DP due to the policy of balancing the candidate list by gender. While PE will likely also indicate some bias, EOP and, to a lesser extent, EO will definitely, as men are more likely to receive false negatives, which may represent more relevant fairness metrics in this instance since the example concerns distributing a resource (the scholarship).

Problem Assessment: Although the metrics are identical in C1a and C1b, indicating a fairness problem in one but not the other metrics, most people's intuitive assessment would likely be that the model in *C1a* is not problematic as long as female applicants are not favored even if they are less well suited, whereas *C1b* is problematic. Hence, for *C1b*, the bias has to be reduced. For C2, the fairness evaluation depends on how one judges the historical gender disparity in scholarship receivers (i.e., is it due to other, more structural biases or "merited"?).

Bias Mitigation Strategy: Given the direct connection between the sensitive feature gender and the target, simply removing this feature as well as features potentially leaking information, so-called proxy-variables [2], such as, e.g., an applicant's school ("ABC School for Girls"), should be enough to mitigate the bias for the first two cases. If one draws the normative conclusion that the bias detected by the separation metrics for C2 is problematic, one will have to alter the parity policy in place.[12]

Effect on Metrics: For C1b, DP will probably no longer indicate a bias but all other metrics (PE, EOP, and EO) will as some women will falsely receive positive predictions and some actually accepted men will not.

Effect on Model Performance and Predictions: The model will be less accurate now for C1b. The more biased the data is, the stronger the decrease in model performance. Accordingly, the predictions for some applicants will change.

3.2 Indirect Connection

Cases: A university trains a model to predict the grades of students' first academic year for all of their programs and uses historical data to do so. We present two cases that differ regarding the relationship between the features and how we perceive this relationship. In *Case 3 (C3)*, there is an indirect connection between gender and grades through study program: Female students tend to pick study programs that award better grades on average. In *Case 4 (C4)*, there is an indirect connection between socio-economic status (SES) and grades through having to work: Students with a lower SES tend to work a lot, working a lot leads to less study time, and this leads to poorer grades. The training data contains all of the mentioned features (working, study time).[13] Figure 2 shows the causal graph for the described cases. Note that A and Y are conditionally independent given X. This means that as long as we know X, additionally knowing A does not add new information as this is already encoded in X. It does not mean, however, that ML models do not exploit this correlation [7].

Metrics: DP will indicate a bias for all of the above-described cases because women have a higher chance of receiving a good grade (C3), and students with a low SES have a lower chance of receiving a good grade (C4). To compute CDP, we could condition on study program for C3 and working and study time for C4. Then, CDP would not indicate a problem for either case. The other fairness metrics (PE, EOP, and EO) will not, or to a smaller degree than for DP, indicate a bias, however, as the true and false negatives of the different groups are likely similar. They might show one in cases where the correlation between the sensitive attribute and the target is utilized, as women might disproportionally be incorrectly predicted to do well, and low SES students might disproportionally be incorrectly predicted to do poorly.

Problem Assessment: Here, the cases differ. For C3, many would probably not see a huge problem. While we would not want a model that predicts that women get better grades due to them being women, we probably have no problem with a model that correctly identifies that students of a certain study program get better grades.

[9] Note that although we usually look at one feature for X, this holds just as well for cases where X consists of multiple features.

[10] The decision to code gender as binary in these examples was made purely for simplifying purposes.

[11] For this reason, we do not apply CDP here.

[12] If one removes this policy, C2 resembles C1b.

[13] Note that if we only had SES as a feature, this case would – due to the omission of additional features – be the same as C1b, i.e., it would become a de facto direct connection.

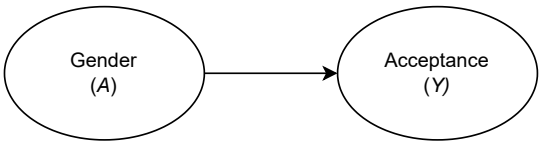

Figure 1: Causal Graph for *Case 1a (C1a)*, *Case 1b (C1b)* and *Case 2 (C2)*.

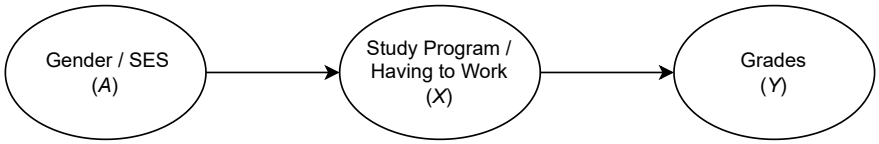

Figure 2: Causal Graph for *Case 3 (C3)* and *Case 4 (C4)*.

If we ensure that the correlation between study program and grades is utilized by removing gender from the training data, then we can probably call C3 fair. Similarly, for the other cases, we obviously do not want models to predict poor grades because someone has a low SES. This might happen for C4. Here, similar to C3, the model can either utilize the correlation between SES and the target or working/study time and the target. Apart from not wanting to have a direct effect of SES on the target (which we can again deal with by removing the sensitive feature), we might also argue that even the indirect effect of SES through working/study time is not something that we want. Working is forced upon the students with a low SES, which appears different from the relationship between gender and study program. Hence, depending on our normative choice and the application scenario, we may validly argue that using study time as a predictor of success is fair or that it is not, prompting us to remove the effect of SES influencing study time. For now, we will assume that we want to remove this effect as well.

Bias Mitigation Strategy: Like before, we can remove the sensitive attributes. This solves the problems we had with C3 but not with C4. We can consider working as a sensitive feature in C4, but this still does not solve the problem, as studying time is still impacted by SES. As a remedy, we could use bias mitigation techniques such as adversarial networks or thresholding.[14]

Effect on Metrics: When we only remove the sensitive attributes, DP might decrease for C3 and C4 if the correlation utilized before was the one between these features and the target – but it will still indicate a bias. If the models did not utilize this relationship, DP will remain the same. For both cases, the other metrics will either not or only slightly change depending on the correlation used before. Once we use bias mitigation techniques in C4, we can expect DP to decrease and the other metrics to increase.

Effect on Model Performance and Predictions: Some predictions may change in both cases depending on which correlation ($A \rightarrow Y$ or $X \rightarrow Y$) was used. Using bias mitigation techniques in C4 might also lead to a decrease in model performance to some extent and to changing predictions. C3 will not experience a reduction in performance metrics.

3.3 Indirect and Direct Connection

Cases: A university trains a model to predict students' grades in a particular course of a study program in the second year. They use previous performance data (i.e., the grades previously achieved in other courses) and demographic data to do so. Note that the underlying causal relationships, as portrayed in Figure 3, include the unobserved variable Motivation, which is actually the common cause of previous grades and the target grade. In Machine Learning contexts, it is common to have latent variables and to use other variables, such as previous grades, to approximate them. Coming back to the scenario, some university professors are biased against students of color, resulting in them having poorer grades on average. In *Case 5a (C5a)*, the professor teaching the course that is to be predicted is not discriminating against people of color, making the target (both historically and currently) unbiased. The resulting model can use an interaction of ethnicity and previous grades, which can "correct" the discrimination (approximating the latent variable motivation) and lead to a generally unbiased model. This requires a method allowing interaction effects[15], however. In *Case 5b (C5b)*, the course grades to be predicted are biased as well, with students of color receiving poorer grades on average.

Metrics: For C5a, DP will not indicate a bias as students of either group will have the same probability of receiving a good grade.[16] However, when we condition on previous grades, CDP will indicate a bias. This is due to the imbalance in the previous grades: If we, e.g., compare students with poor previous grades, students of color will disproportionally be predicted to have good grades. For C5b, DP will indicate a bias, as the groups have a different probability of receiving a good grade. If we condition on previous grades here, CDP will not indicate a bias or to a lesser extent (depending on whether the amount of discrimination is at the same level). For C5a and C5b, PE, EOP, and EO do not indicate a bias as the models are

[14]However, we have to be careful with these. Using a decorrelation technique, decorrelating the other predictive features from SES, might not be something that we want as we might want to remove the effect of SES on study time through working instead of the effect of SES on study time.

[15]Crudely speaking, an interaction effect occurs if two (or more) features have a combined effect. This effect should be naturally learned in Neural Networks but must be enforced in linear/logistic regression by, e.g., multiplying the features.

[16]This assumes that there is no other effect of the discrimination, such as, e.g., a lack of motivation as a response to the discrimination.

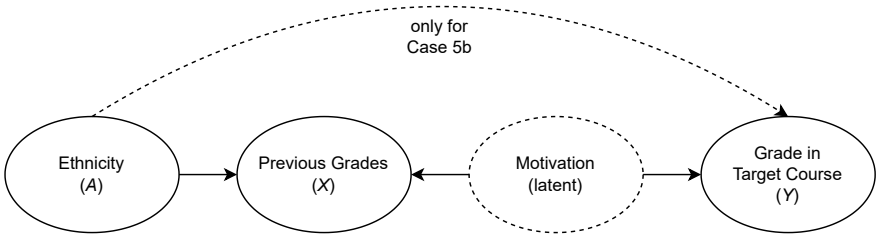

Figure 3: Causal Graph for *Case 5a (C5a)* and *Case 5b (C5b).*

roughly equally correct for all groups.[17]

Problem Assessment: Although the previous discrimination is clearly problematic, the model using interaction effects in C5a is not flagged by the fairness metrics, which might appear to be surprising considering the highly biased training data. The reason for this, as already mentioned, is that the resulting model can use an interaction of ethnicity and previous grades, which can "correct" the discrimination (and estimate the latent variable Motivation), leading to a generally unbiased model. C5b is definitely problematic.

Bias Mitigation Strategy: As C5a is not problematic, we do not need to change anything. However, we want to highlight that removing the sensitive feature in this scenario produces a biased model, as previous performance cannot be adjusted. This results in a model more often erroneously predicting students of color to do poorly and other students to do well, and affects all fairness metrics. This might appear counterintuitive at first. A slightly different scenario would evolve if the students of color noticed the discrimination occurring, which demotivates them from studying, leading to poorer grades and an additional correlation between ethnicity and the target. In this case, we would need to remove the sensitive attribute so that it cannot have an effect on its own but prepare the data so that ethnicity and previous performance are combined. This ensures that an interaction effect correcting for the bias is learned. Note that this is not a strategy typically considered for improving fairness and that we typically do not know enough about the relationships to even do this. For C5b, we can remove the sensitive attribute and use a bias mitigation strategy such as decorrelation. This will help, but, of course, the general problem of having a professor discriminating against students in this course will not be solved (and cannot be solved by any technological strategy).

Effect on Metrics: For C5b, DP and CDP will most likely no longer or to a lesser extent indicate a bias, but all other metrics now will.

Effect on Model Performance and Predictions: For C5b, the performance will drastically decrease as we now regularly predict students of color to do well, which did not. Logically, there will be instances receiving different predictions as well.

3.4 Representation Bias

Cases: Our final cases are again concerned with a model trained to predict whether applicants should be admitted based on historical records. One feature available to the model is the nationality of the applicant.[18] Certain nationalities (e.g., "Jupiter" and "Mars") are very rare, and all applicants from "Jupiter" are admitted and all from "Mars" rejected. In Case 6a (C6a), this happened due to chance; the applicants of "Jupiter" happened to be better. In Case 6b (C6b), applicants from "Jupiter" are preferred for some reason by the admittance office.

Metrics: It is very likely that all metrics will detect a bias for both cases. We have the largest bias possible for DP as all applicants from "Jupiter" are admitted and all from "Mars" rejected. It is also likely for both cases that the models will perform perfectly on these data instances, while some instances of other nationalities will most likely be falsely classified, which will lead to a slight bias indication via PE, EOP and EO.

Problem Assessment: It is generally problematic if the models learn to predict based on the sensitive feature. This has to happen for C6b as there is no other reason that the applicants are preferred.[19] For C6a, the model could also not use the sensitive attribute at all and, instead, predict admittance based on truly relevant factors (as applicants from "Jupiter" happen to be the better applicants by chance) – but this is not likely as the relationship between nationality and admittance is easier to learn.

Bias Mitigation Strategy: We can remove the sensitive feature, which fixes the problem that we do not want to predict based on the sensitive feature.

Effect on Metrics: Simply due to the very small sample size, DP might still indicate a problem for C6b. For C6a, DP will still indicate a fairness concern because, by chance, all applicants from one country have the same prediction. The other fairness metrics will most likely detect some unfairness in C6b due to the predictions no longer fitting the ground truth.

Effect on Model Performance and Predictions: The model performance will decrease, and some individual predictions will change for C6b. For C6a, this could only happen due to misclassification occurring now; otherwise, everything stays the same.

3.5 Main Findings

Our case studies lead to a variety of interesting observations summarized in Table 1 in the Appendix. First of all, and as already mentioned, different metrics are sensible in different scenarios. If a meaningful and legitimate correlation exists between a sensitive

[17] The result for C5a would be different if the model does not allow interaction effects, in which case there could be an indication for a bias as the model would likely just learn a positive effect on grades for students of color to "compensate" for the historical bias in the data set which could lead to them being more often falsely classified since the target is unbiased.

[18] As we focus on this feature, we do not apply CDP in this section.

[19] In principle, this case is equivalent to C1b. However, the small sample size in C6b renders the decision of whether a problematic bias exists difficult.

feature and another predictive feature, DP is not a helpful metric; otherwise, it is. CDP is only sensible in cases where we know what to condition on and consider the conditioning feature unproblematic. The separation metrics should only be used if the target itself is unbiased. Moreover, the choice for a specific metric is often normative. If the target is biased, then the metrics will start to display a bias once we try to deal with the bias detected by demographic parity. This shows that in most cases, the metrics are incompatible and, more importantly, that they should not be understood as an absolute measure of fairness but rather as an indicator of a specific notion of fairness in a specific case [20]. Second, some cases present themselves in a similar fashion but are, in fact, very different (e.g., C1a and C1b); others appear rather different but are similar. This has to be considered when performing fairness evaluations. Third, our assessment of whether a fairness concern exists or not strongly depends on the application we have in mind and our normative convictions. We can also observe that while fairness concerns evolving due to direct connections are somewhat more clearly problematic, concerns evolving from indirect connections are much more normative and application-dependent. Fourth, some patterns emerge as to how we can solve fairness concerns. If there is a direct connection, fairness concerns can be solved by simply removing sensitive and proxy features. The same mostly holds for issues stemming from representation biases. If there are indirect effects, which is likely, we might also need to employ bias mitigation techniques. If both are at work, we may have to find very specialized approaches.

When we consider the discussion of the cases, it seems like almost any kind of fairness concern can be solved (except for C5b). However, the cases are artificial in the sense that we know everything about the true data generation mechanism, and we have good models. Without these assumptions, we cannot know what metrics to look out for and what strategies to employ. Even removing demographic features – usually a very sensible strategy – can, at times, be the wrong decision (C5a). All in all, our observations clearly show no "one size fits all" strategy exists. We need to consider the data and potential fairness concerns carefully.

4 RECOMMENDATIONS AND FUTURE WORK

A starting point for this might be Vasquez Verdugo et al.'s work providing a framework to critically asses the specific case and decide on relevant fairness metrics [24].

Additionally, we believe that future research in the educational domain with potential for fairness concerns should facilitate discussion in this vein. Those that present new methods should highlight for which applications and data this might be problematic and what metrics might be indicative of it. Those publishing datasets should be specific regarding data selection, generation, and potential biases. Those who apply data to methods should try to analyze their results accordingly. We furthermore want to stress that there potentially exists a fallacy when trying to quantify unfairness: It may give a false sense of objectivity where, in reality, much is normative. This does not mean that the metrics should not be used but that they should be used consciously and that researchers should transparently explain their decisions. Both for researchers and practitioners, in many cases, it may be prudent not to deploy algorithmic decision-making in a given context and rather proactively attempt to alleviate the bias by, e.g., altering policies promoting said bias.

Finally, we would like to stress that future research should attempt to develop methods that reveal and quantify connections between demographic and other predictive features. We think that data-driven methods to uncover causal mechanisms should be employed and investigated in the area of fairness.

5 CONCLUSION

Even under the most unrealistic assumptions, thinking about fairness metrics and what they reveal is complex. In reality, it is even more complicated. This does not mean that we should not use the metrics or – worse yet – give up on fairness. We believe that our paper highlights the need to critically assess whether data, a method, or an application has potential fairness concerns, what metrics can detect them, and how we can deal with them. If we normalize the inclusion of such remarks in research, we facilitate discussions about it. Eventually, thinking about these issues will, at least, no longer feel quite as complicated, and this will help us produce fairer models for all involved in education. We see our paper as a first step towards this and hope it offers a starting point for practitioners and researchers alike to think critically about potential fairness concerns and how to handle them.

ACKNOWLEDGMENTS

Jakob Kappenberger is funded by the grant "Consequences of Artificial Intelligence for Urban Societies (CAIUS)," by Volkswagen Foundation. We want to thank our colleague Darshit Pandya for his helpful feedback on this article.

REFERENCES

[1] Carolyn Ashurst, Ryan Carey, Silvia Chiappa, and Tom Everitt. 2022. Why fair labels can yield unfair predictions: Graphical conditions for introduced unfairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9494–9503.

[2] Ryan S Baker, Lief Esbenshade, Jonathan Vitale, Shamya Karumbaiah, et al. 2023. Using Demographic Data as Predictor Variables: a Questionable Choice. *Journal of Educational Data Mining* 15, 2 (2023), 22–52.

[3] Ryan S Baker and Aaron Hawn. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* (2021), 1–41.

[4] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (March 2022), 4209. https://doi.org/10.1038/s41598-022-07939-1

[5] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Transactions on Software Engineering and Methodology* 32, 4 (Oct. 2023), 1–30. https://doi.org/10.1145/3583561

[6] Silvia Chiappa and William S Isaac. 2019. A causal Bayesian networks viewpoint on fairness. *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data: 13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Vienna, Austria, August 20-24, 2018, Revised Selected Papers 13* (2019), 3–20.

[7] Lea Cohausz, Andrej Tschalzev, Christian Bartelt, and Heiner Stuckenschmidt. 2023. Investigating the Importance of Demographic Features for EDM-Predictions. (2023).

[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. https://doi.org/10.1145/3097983.309809 arXiv:1701.08230 [cs, stat].

[9] Oscar Blessed Deho, Srecko Joksimovic, Jiuyong Li, Chen Zhan, Jixue Liu, and Lin Liu. 2022. Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why. *IEEE Transactions on Learning Technologies* (2022).

[10] Oscar Blessed Deho, Chen Zhan, Jiuyong Li, Jixue Liu, Lin Liu, and Thuc Duy Le. 2022. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology* 53, 4 (July 2022), 822–843. https://doi.org/10.1111/bjet.13217

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, Cambridge Massachusetts, 214–226. https://doi.org/10.1145/2090236.2090255

[12] Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2023. Measuring Fairness Under Unawareness of Sensitive Attributes: A Quantification-Based Approach. *Journal of Artificial Intelligence Research* 76 (April 2023), 1117–1180. https://doi.org/10.1613/jair.1.14033 arXiv:2109.08549 [cs].

[13] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, 225–234. https://doi.org/10.1145/3303772.3303791

[14] Karen Hao. 2020. The UK exam debacle reminds us that algorithms can't fix broken systems. *MIT Technology Review* (Aug. 2020). https://www.technologyreview.com/2020/08/20/1007502/uk-exam-algorithm-cant-fix-broken-system

[15] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[16] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems* 35, 3 (June 2013), 613–644. https://doi.org/10.1007/s10115-012-0584-8

[17] Lin Li, Lele Sha, Yuheng Li, Mladen Raković, Jia Rong, Srecko Joksimovic, Neil Selwyn, Dragan Gašević, and Guanliang Chen. 2023. Moral Machines or Tyranny of the Majority? A Systematic Review on Predictive Bias in Education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. ACM, Arlington TX USA, 499–508. https://doi.org/10.1145/3576050.3576119

[18] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management* 58, 5 (Sept. 2021), 102642. https://doi.org/10.1016/j.ipm.2021.102642

[19] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2022), 1–35. https://doi.org/10.1145/3457607

[20] Tim Räz. 2021. Group Fairness: Independence Revisited. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 129–137. https://doi.org/10.1145/3442188.3445876 arXiv:2101.02968 [cs].

[21] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 59–68. https://doi.org/10.1145/3287560.3287598

[22] Lele Sha, Mladen Rakovic, Angel Das, Dragan Gasevic, and Guanliang Chen. 2022. Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education. *IEEE Transactions on Learning Technologies* 15, 4 (Aug. 2022), 481–492. https://doi.org/10.1109/TLT.2022.3196278

[23] Frank Stinar and Nigel Bosch. 2022. Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results. (July 2022). https://doi.org/10.5281/ZENODO.6853135 Publisher: Zenodo.

[24] Jonathan Vasquez Verdugo, Xavier Gitiaux, Cesar Ortega, and Huzefa Rangwala. 2022. FairEd: A Systematic Fairness Analysis Approach Applied in a Higher Educational Context. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. ACM, Online USA, 271–281. https://doi.org/10.1145/3506860.3506902

A SUMMARY TABLE

Table 1: A summary of the cases regarding which metrics detect a bias, whether it is considered problematic, how to treat it, what metrics detect a bias after the treatment, the effect of the treatment on performance and individual predictions, and whether the model is still problematic.

Case	Detects Bias	Problem	Treatment	Detects Bias II	Performance	Ind. Prediction	Problem II
C1a	DP	no	-	-	-	-	-
C1b	DP	yes	remove sensitive feature	PE, EOP, EO	decreases	some changes	no
C2	PE/EOP/EO	potentially	-	-	-	-	-
C3	DP, maybe PE/EOP/EO	(yes)	remove sensitive feature	DP, maybe PE/EOP/EO	no change	maybe	no
C4	DP, maybe PE/EOP/EO	yes	remove sensitive feature, bias mitigation technique	PE, EOP, EO	decreases	some changes	no
C5a	CDP	no	-	-	-	-	-
C5b	DP	yes	remove sensitive feature, bias mitigation technique	PE, EOP, EO, maybe DP	strongly decreases	many changes	yes
C6a	all	(yes)	remove sensitive feature	all	no change/ slight decrease	no/few changes	no
C6b	all	yes	remove sensitive feature	PE, EOP, EO, maybe DP	decreases	some changes	no