






©2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Investigating the Legality of Bias Mitigation Methods in the United Kingdom

Mackenzie Jorgensen , Madeleine Waller , Oana Cocarascu , Natalia Criado , Odinaldo Rodrigues , Jose Such , Elizabeth Black 

I. INTRODUCTION

Algorithmic decision-making systems (ADMS)¹ fairness issues have been well highlighted over the past decade [1], including some facial recognition systems struggling to identify people of colour [2]. In 2021, Uber drivers filed a claim with the UK’s employment tribunal for unfair dismissal resulting from Automated Facial Recognition technology by Microsoft [3]. *Bias mitigation methods* have been developed to reduce discrimination from ADMS. These typically operationalise fairness notions as *fairness metrics* in order to minimise discrimination [4]. We refer to ADMS to which bias mitigation methods have been applied as “mitigated ADMS” or, in the singular, a “mitigated system.”

Bias mitigation methods aim to avoid ADMS from unlawfully discriminating. Progress has been made in assessing which fairness metrics are consistent with US [5] and EU non-discrimination law [6]. Similarly, bias mitigation methods have been investigated in the context of US non-discrimination law and affirmative action [7]. While the investigation of fairness metrics and their legality under EU law are likely applicable to the UK [6], the relationship between bias mitigation methods applied to ADMS and UK non-discrimination law has not been well analysed. There is a risk in applying these methods to ADMS since they could introduce discrimination and this is the focus of our article [8].

In contrast to similar literature, we specifically focus on the UK context. The UK Equality Act 2010 [9], referred to here as the *Equality Act*,² defines discrimination as treating someone less favourably because of a protected characteristic (e.g., age, sexual orientation, etc); discrimination can be *direct*, meaning it is linked directly to a protected (or perceived) characteristic of the individual (or someone connected); or *indirect*, when it results from a policy that unfairly affects an individual with a protected characteristic. In this article, when

we refer to legality, we mean compliance with provisions on discrimination in the Equality Act.

To better grasp the complexity of the Equality Act and its relationship to algorithmic fairness, we consider an example where an ADMS is being developed to predict whether an application for a benefit might be fraudulent. Assume the artificial intelligence (AI) practitioners decide not to include race in the data so the system cannot directly discriminate on race. If the practitioners are concerned about the system’s ability to indirectly discriminate due to other variables that could act as proxies for race, then they could conduct an analysis checking for disparities among different groups (e.g., by using fairness metrics). If they find unexpected differences between groups that cannot be explained and justified, then they would likely want to adjust the system to mitigate for indirect discrimination.

To mitigate for indirect discrimination, they could choose the “decision threshold optimiser” bias mitigation method which aims to adjust thresholds for groups to achieve parity between them [11]. However, the mitigated system now directly discriminates based on group membership since race is a factor in scoring people and making decisions; irrespective of whether the mitigated system decisions affect people negatively or positively, it is directly discriminating with no justification. The practitioners are likely legally worse off because the mitigated system directly discriminates in comparison to when they started and considered how the unmitigated system could indirectly discriminate (which is arguably trickier to demonstrate).

Here, we are mostly concerned with positive and negative direct discrimination that could come from mitigated ADMS, but also consider indirect discrimination where applicable. We provide the first exploration of how mitigated ADMS could contravene the Equality Act. We explicitly focus on discrimination from mitigated ADMS not present in the unmitigated ADMS as highlighted in the housing benefits example. We also present a bias mitigation method categorisation according to how dependent the resulting mitigated ADMS are on protected characteristics. We articulate these discussions and issues and use the categorisations around mitigated ADMS in a public sector case study on immigration. The issues we discuss are complex with multiple factors and we aim to draw out the challenges in the interpretation of the Equality Act for

Manuscript submitted 19th November 2023; revised 5th and 21st December 2023.

M. Jorgensen, M. Waller, E. Black, O. Cocarascu, O. Rodrigues and J. Such are with the Department of Informatics, King’s College London, London, UK (email: {firstname}.{surname}@kcl.ac.uk).

N. Criado and J. Such are with the Valencian Institute for Artificial Intelligence (VRAIN), Universidad Politècnica de València Innovation, València, Spain (email: {ncriado, jsuch}@upv.es).

¹By ADMS, we refer to machine learning systems that have classification or risk scoring abilities which can make or support decisions in real-world contexts. We use “system” as the singular form of ADMS.

²There is another UK Equality Act 2006 [10], but we focus on the 2010 version.

II. BACKGROUND

We provide background to the Equality Act, present a case study, outline existing fairness metrics, and give insight into related work.

A. Interpreting the UK Equality Act 2010

We define discrimination according to the Equality Act [9] which outlines the UK’s legislation enforcing non-discrimination law with the aim of equality across different sectors. *Direct discrimination* with negative treatment happens “when someone is put at a disadvantage or treated less favourably because of a protected characteristic” [12]. Meanwhile, *indirect discrimination* happens “when a working practice, policy or rule is the same for everyone but has a worse effect on someone because of a protected characteristic.” We highlight that the notion of favourability is subjective and extremely context dependent.

Positive action involves taking action to treat a “group that shares a protected characteristic more favourably than others, where this is a proportionate way to enable or encourage members of that group to: overcome or minimise a disadvantage, have their different needs met, participate in a particular activity” [13]. Positive action is voluntary and not required. However, *positive discrimination*, unlawful direct discrimination, is when a group is treated more favourably based on a protected characteristic and it does not meet the criteria outlined for positive action. It is unclear how one should interpret the boundary between lawful positive action and unlawful positive discrimination in the context of a mitigated system. Our article attempts to explore these legal concepts with a technical lens and to identify key considerations.

B. Case Study: Immigration

The public sector case study we use throughout the article is from immigration services. We consider a visa application system that predicts whether an applicant should be streamlined (have their application processed much faster than normal) for a UK visa.³ Throughout this article, we consider race but our analysis similarly applies to other protected characteristics. A positive classification means an application is streamlined for a visa while a negative classification means it is not. The latter will likely mean flagging the application resulting in delays and a higher likelihood of rejection later. A *true positive* (TP) decision is when an applicant is correctly streamlined and a *true negative* (TN) decision is when an applicant is correctly *not* streamlined. A *false positive* (FP) decision is when an applicant is unduly streamlined, bypassing necessary checks and potentially leading to incorrectly granted visas. A *false negative* (FN) decision is when an applicant is incorrectly *not* streamlined, resulting in further checks, and potentially seriously impacting the applicant’s ability to travel.

This example is based on a similar service previously used in the UK’s Home Office until 2020 [15], giving green, amber

or red labels to applications to inform caseworking processes. The algorithm was dropped after legal action was brought against the Home Office from the Joint Council for the Welfare of Immigrants and Foxglove [16]. They claimed that the model’s use of nationality as a factor could cause indirect racial discrimination, breaching the Equality Act. Under the UK Public Sector Equality Duty (a part of the Equality Act), advancing equality of opportunity in immigration services in relation to race is not required [17]. However, the UK’s Home Office example highlights that there is still a potential to breach the Equality Act if discrimination can be shown. In this article, we consider what bias mitigation could look like to alleviate racial discrimination from a similar but hypothetical system.

C. Detecting Bias in ADMS

The algorithmic fairness community has developed fairness metrics for detecting bias often with respect to model decisions (e.g., TP, TN, FP, and FN) and protected characteristics (also referred to as *protected or sensitive attributes*)⁴ or proxies for them [4]. For example, an individual’s name is often a proxy for that individual’s race or gender [18].

Fairness metrics quantify either individual or group fairness. Individual fairness ensures fairness towards an individual, e.g., by evaluating to what extent similar individuals receive the same decision (the notion of similarity depends on the context) [19]. Group fairness, which we focus on, ensures that two groups defined by values of a single protected characteristic are treated similarly. We define the fairness metrics that are considered in our article next.

Demographic Parity (DP)⁵ aims for the same positive classification rates across protected groups [19]. Optimising for DP disregards potentially legitimate disparity between classifications for protected groups and removing this disparity could decrease the accuracy of ADMS. Other group fairness metrics ensure that decisions from ADMS reflect the training data. *Equalised odds* (EO) aims for the same proportion of correct and incorrect positive classifications, respectively TPs and FPs, across protected groups [11]. *Equality of opportunity* (EOO) is similar but measures only TP rates across protected groups [11]. *Error rate parity* (ERP) quantifies the difference in FP and FN rates across protected groups [20].

D. Related Work

The work closest to our article is an investigation of fairness metrics and their relationship to EU non-discrimination law [6]. The *conditional demographic parity* (CDP) metric represents the notion of fairness in EU non-discrimination law, by measuring the difference in the proportion of positive classifications for different protected groups, conditional on a legitimate characteristic. For example, a bank offering the same proportion of loans to men and women in the same

³We assume that the government department would have access to the data needed to train the model, including protected characteristics. However, the data available might not be fully representative; there will always be “invisible data” such as people not included [14].

⁴We say protected characteristic throughout the article to use the Equality Act rhetoric and refer to a specific identity group within this category as a protected group.

⁵DP is sometimes called statistical parity in the literature.

income bracket would satisfy CDP. This requires domain knowledge of what characteristics are legitimate.

An analysis of algorithmic fairness literature discovered most works focus on the US legal landscape and would not transfer to the EU [21]. The legality of fairness metrics under EU non-discrimination law is discussed and the metrics are categorised into two categories, bias preserving and bias transforming [6]. They provide a checklist for guidance in choosing the right fairness metric to detect bias in a system. The motivation aligns with our work and some discussions are transferable to a UK context. However, instead of metrics, we focus on the impact of bias mitigation methods and how their application can not only preserve existing bias in data but also introduce discrimination that was not present before in unmitigated ADMS. Scholars have developed and tested the effectiveness of many bias mitigation methods [4]. We focus on three methods which we define below. We do not discuss related work that considers the relationship between US non-discrimination law and algorithmic fairness since they are less applicable in the UK context [5], [7].

III. CATEGORISING AND APPLYING BIAS MITIGATION METHODS

We present a bias mitigation method categorisation based on the reliance of mitigated ADMS on protected characteristics after the methods have been applied. This reliance has implications for whether the mitigated ADMS could introduce direct discrimination not present in the unmitigated ADMS. We only talk about three methods, one from each category, but from analysing state-of-the-art methods (e.g., ones found in the *Fairlearn* toolkit [22]) we find our categorisation covers them. No category of methods is better than another.

We define each method, explain its categorisation, discuss how its use of protected characteristics could introduce discrimination, and consider the effects of the mitigated system within the case study described in the previous section. We emphasise that any protected group⁶ could be discriminated against and categorise bias mitigation methods according to the following properties:

- No reliance: Methods prevent the protected characteristics from directly influencing ADMS decisions.
- Medium reliance: Methods attempt to minimise the influence of the protected characteristic on ADMS decisions.
- High reliance: Methods make no attempt to minimise the influence of the protected characteristic on ADMS decisions.

A. “No Reliance” Method Example and Application

Fairness through unawareness [23] removes protected characteristics from the training data. This method aims to address direct discrimination since no protected characteristics are included in the decision-making. We categorise Fairness Through Unawareness as “no reliance” because by removing

the protected characteristics from the training dataset, the method prevents them from having any direct influence on the decision, and by definition the mitigated system cannot directly discriminate. However, there is still the risk of indirect discrimination if proxies are not removed (e.g., an individual’s postcode can be an indicator of their race [24]). If proxies remain in the dataset, the mitigated system could more heavily weigh them, exacerbating indirect discrimination already present in the unmitigated system. Applying a method from this category to a system could result in poorer accuracy and even poorer fairness metric results.

If an AI practitioner for our case study applies *fairness through unawareness* to an ADMS, a visa applicant’s protected characteristics have no direct impact on whether the application is streamlined. There is no potential for the mitigated system to directly discriminate (e.g., treating individuals favourably or unfavourably based on protected characteristics). However, by ignoring these, no attempt is made to balance positive classifications across groups (e.g., no attempt at having the same proportion of white and non-white applicants streamlined). Not considering the difference in classifications for different groups could lead to mitigated ADMS indirectly discriminating if applicants from a protected group are flagged for more checks over applicants in another group. As aforementioned, indirect discrimination is still likely and could be exacerbated by the method’s application if there are proxies in the data. Assessing whether ADMS indirectly discriminate (with or without the application of a method) is more difficult than assessing direct discrimination.

B. “Medium Reliance” Method Example and Application

Adversarial debiasing [25] trains two machine learning models. The first model tries to optimise for a fairness metric (DP, EO, or EOO) and accuracy and tries to fool a second model which attempts to guess individuals’ protected characteristics. If the second model can guess the protected characteristic correctly, then the first model treats the groups differently and is likely discriminating. We classify adversarial debiasing as “medium reliance” because of the objective to reduce the influence of the protected characteristics on the decisions. This method requires a fairness metric which in many use cases could actively promote equality for a justifiable reason (e.g., the practitioner has reason to believe that their historical data is biased). This could be considered positive action on an individual level. After applying this method, some decisions may change for individuals because of their protected characteristics, so the mitigated ADMS could still directly discriminate. The legality of this method should be decided in the courts. Future legislation might specify the relationship between promoting equality through positive action and direct discrimination more clearly in mitigated ADMS.

If an AI practitioner applies *adversarial debiasing* to a visa streamlining system, they would need to choose a fairness metric. They might choose to optimise for EO due to the priority of streamlining the correct visa applicants (TPs) and not streamlining applicants that should be flagged for further application checks (FPs) across groups. Adversarial debiasing

⁶There can be multiple protected characteristics at play (e.g., race and religion) and within those protected characteristics, there will be multiple protected groups (e.g., for religion, the groups could be Christianity, Islam, and Judaism).

with EO aims to satisfy EO and accuracy by only considering non-protected attributes for the classifications, if possible. After applying the method, the mitigated system should be less dependent on protected characteristics. This is still a challenging case because even though the method attempts to reduce the mitigated system's reliance on protected characteristics, the mitigated ADMS could still introduce direct discrimination not present before. For example, a previously flagged non-white applicant might be streamlined and a previously streamlined white applicant might be flagged for checks given the mitigated system. This could be an example of positive discrimination. That example could also highlight negative discrimination towards the white applicant.

C. "High Reliance" Method Example and Application

Reductions via constrained optimization [26] addresses direct discrimination because it optimises for a fairness metric (DP, EO, EOO, or ERP) and accuracy. The fairness metric is tied to a protected characteristic with the idea of uplifting a protected group that is historically underprivileged or underrepresented. We categorise this method as "high reliance" because the protected characteristic is crucial for the method's application. The protected characteristic directly influences the mitigated ADMS decisions because the method optimises for a fairness metric defined by that protected characteristic. The risk of the mitigated ADMS directly discriminating in a way not present in the unmitigated ADMS is high.

If an AI practitioner applies *reductions via constrained optimization* to a system, they need to choose a fairness metric to optimise for alongside accuracy. They might choose to optimise for ERP due to the priority of not streamlining applicants that should be flagged for further checks (FPs) and, more importantly, not flagging applicants for further checks that should be streamlined (FNs). The method might be successful in reducing direct discrimination through the mitigated ADMS in comparison to the unmitigated ADMS; however, since the method is "high reliance," the protected characteristic is crucial for the application of the method and individual decisions are likely to change solely based on protected characteristics. As a result, mitigated ADMS would likely introduce direct discrimination that the unmitigated ADMS did not have. For example, the method could have been applied to mitigate for negative direct discrimination but then the mitigated system ends up positively discriminating.

The mitigated ADMS could also introduce indirect discrimination that was not present in unmitigated ADMS. If the mitigated system improves ERP with respect to one protected characteristic and disadvantages a protected group with respect to another protected characteristic, this could be indirect discrimination. For example, applying the method could ensure ERP is satisfied across racial protected groups but this could reduce the ERP for groups under another protected characteristic (e.g., disability). The practitioner could argue that by optimising for ERP with respect to race they are attempting to be as fair as possible towards racial groups; however, the disability indirect discrimination exacerbated or introduced by the mitigated system is problematic.

IV. DISCUSSION

Our article opens a discussion on the key issues surrounding mitigated ADMS that introduce discrimination not previously present in the unmitigated ADMS. An AI practitioner could intend to remove harmful bias from a system by applying a bias mitigation method to it. However, even with the intention of mitigating negative discrimination, for example, the mitigated ADMS could end up positively discriminating. The complexity of positive action versus positive discrimination from mitigated ADMS is a question technical scholars cannot answer alone. A comprehensive investigation of relevant case law would help towards this; although, case law in this rapidly developing area is rather sparse. A limitation of our work is that we mainly focused on a single protected characteristic although the Equality Act Section 14 brings up intersectional or combined discrimination [9]. Next, we discuss the importance of context, our categorisation of bias mitigation methods, and reflections on the case study analysed.

A. Context is Key for Fairness Considerations

Two of the bias mitigation methods we considered optimise for a fairness metric. Similar to the choice of a bias mitigation method, the choice of metric is nontrivial [27]. Understanding the context is essential because some fairness metrics are relevant or meaningful in certain scenarios but not appropriate in others [28]. Our case study in the previous section considered some options, but an AI practitioner may identify further options.

We emphasise the importance of AI practitioners understanding how the choice of a fairness metric is related to the decisions of ADMS. System decisions can have positive or negative impacts on the protected groups and by applying bias mitigation methods to ADMS, these now mitigated system decisions can be altered [29]. Those potential impacts on protected groups could be the result of direct discrimination (including negative or positive discrimination) or indirect discrimination.

B. Cautiously Use Bias Mitigation Methods

We recommend using fairness metrics to monitor bias in ADMS, and, with caution, to use bias mitigation methods to attempt to make them less discriminatory. However, as highlighted, the use of existing bias mitigation methods on ADMS does not currently guarantee that they will follow the Equality Act and their application could result in mitigated ADMS discriminating in ways not present before. We mostly focused on direct discrimination that could arise from applying methods to ADMS. Indirect discrimination could come from ADMS with or without the application of bias mitigation methods and use of protected characteristics in the training dataset. Of the categorisations we provided, *no one category of methods is better than another*; rather, the categorisations are meant to help practitioners think about how applying methods to ADMS will affect the treatment of protected groups. We argue that discrimination introduced by mitigated ADMS is greatly affected by the context.

C. Reflections on the Case Study

An AI practitioner in our case study could check for discrepancies in the decisions across the protected groups as a result of the application of a bias mitigation method by measuring bias with fairness metrics. Assuming a “medium” or “high reliance” method was used, if there is an increase in bias or a change in accuracy that could negatively or positively affect a protected group, the mitigated system is likely introducing direct discrimination. Otherwise, these changes could be attributed to indirect discrimination. We note that if the mitigated ADMS is less biased than the unmitigated ADMS, the practitioner could argue that applying the method improves equality.

V. CONCLUSION

The assumption that using bias mitigation methods is always beneficial for non-discrimination efforts is very naive. Through this article, from a technical perspective, we discussed the issues from the Equality Act that arise as a result of applying bias mitigation methods to ADMS. We focused on discrimination coming from mitigated ADMS that was not present before applying the methods. Our categorisation of methods assists our discussion since the methods often directly influence how important protected characteristics are for mitigated ADMS. By using a public sector case study on immigration, we analysed the discrimination implications when applying three bias mitigation methods to ADMS. Although the relationship between the Equality Act and algorithmic fairness research is not easily interpretable, we start a much needed discussion for non-discrimination purposes.

ACKNOWLEDGMENTS

The work of Mackenzie Jorgensen and Madeleine Waller was supported by the U.K. Research and Innovation under Grant EP/S023356/1 in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence. The work of Madeleine Waller was supported by The Alan Turing Institute’s Enrichment Scheme. This work was supported in part by Project TED2021-131295B-C32, in part by Project MCIN/AEI/10.13039/501100011033, and in part by the European Union NextGenerationEU/PRTR. Mackenzie Jorgensen began this research under the mentorship of Mark Durkee while interning at the Centre for Data Ethics & Innovation (CDEI) which is a part of the Department for Science, Innovation & Technology; the authors are incredibly grateful for Mark Durkee’s feedback on this work even after Mackenzie Jorgensen finished her internship. This article does not represent an official view of the CDEI. Mackenzie Jorgensen, Madeleine Waller, and Elizabeth Black are affiliates of the King’s Institute for Artificial Intelligence. The authors thank Victoria Hendrickx, Yasaman Yousefi, and Brent Mittelstadt for their valuable feedback on previous article versions.

REFERENCES

[1] X. Ferrer, T. van Nuenen, J. Such, M. Cote, and N. Criado, “Bias and discrimination in ai: a cross-disciplinary perspective,” *IEEE Technology and Society*, vol. 20, no. 2, pp. 72–80, 2021.

[2] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Proc. 1st Conference on Fairness, Accountability and Transparency*, vol. 81, 23–24 Feb 2018, pp. 77–91.

[3] C. Vallance, “Legal action over alleged uber facial verification bias,” *BBC*, 2021.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Comput. Surv.*, vol. 54, no. 6, 2021.

[5] A. Xiang and I. D. Raji, “On the legal compatibility of fairness definitions,” *CoRR*, vol. abs/1912.00761, 2019.

[6] S. Wachter, B. Mittelstadt, and C. Russell, “Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law,” *West Virginia Law Review*, vol. 123, no. 3, 2021.

[7] P. T. Kim, “Race-aware algorithms: Fairness, nondiscrimination and affirmative action,” *California Law Review*, vol. 110, pp. 1539–1596, 2022.

[8] Centre for Data Ethics and Innovation, “Review into bias in algorithmic decision-making,” Tech. Rep., 2020.

[9] UK Public General Acts, “Equality Act c.15,” 2010.

[10] —, “Equality Act c.3,” 2006.

[11] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 1–9.

[12] Advisory, Conciliation and Arbitration Service, “Discrimination and the equality act 2010,” 2023.

[13] UK Government, “Positive action in the workplace guidance,” 2023.

[14] M. Onuoha, “Broadway won’t document its dramatic race problem, so a group of actors spent five years quietly gathering this data themselves,” *Quartz*, 2016.

[15] BBC Technology Inc, “Home office drops ‘racist’ algorithm from visa decisions,” *BBC*, 2020.

[16] The Joint Council for the Welfare of Immigrants, “We won! home office to stop using racist visa algorithm,” *JCWI Latest News*, 2020.

[17] Equality and H. R. Commission, “Public Sector Equality Duty Advice and Guidance,” 2023.

[18] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, “Avoiding discrimination through causal reasoning,” in *Proc. 31st International Conference on Neural Information Processing Systems*, 2017, p. 656–666.

[19] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, “Fairness through awareness,” in *Innovations in Theoretical Computer Science 2012*. ACM, 2012, pp. 214–226.

[20] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[21] S. Wachter, B. D. Mittelstadt, and C. Russell, “Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI,” *Comput. Law Secur. Rev.*, vol. 41, p. 105567, 2021.

[22] S. Bird, M. Dudík, H. Wallach, and K. Walker, “Fairlearn: A toolkit for assessing and improving fairness in ai,” Microsoft, Tech. Rep., 2020.

[23] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, “The case for process fairness in learning: Feature selection for fair decision making,” in *Proc. NIPS Symposium on Machine Learning and Law*, vol. 1, 2016.

[24] B. Wiggins, in *Calculating Race: Racial Discrimination in Risk Assessment*. Oxford University Press, 11 2020.

[25] B. H. Zhang, B. Lemoine, and M. Mitchell, “Mitigating unwanted biases with adversarial learning,” in *Proc. 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, p. 335–340.

[26] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, “A reductions approach to fair classification,” in *Proc. 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 60–69.

[27] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, “The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making,” *Commun. ACM*, vol. 64, no. 4, p. 136–143, 2021.

[28] R. Binns, “Fairness in machine learning: Lessons from political philosophy,” in *Proc. 1st Conference on Fairness, Accountability and Transparency*, vol. 81, 2018, pp. 149–159.

[29] M. Jorgensen, H. Richert, E. Black, N. Criado, and J. Such, “Not so fair: The impact of presumably fair machine learning models,” in *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, 2023.