

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


Aequitas 2023: Workshop on Fairness and Bias in AI | co-located with ECAI 2023, Kraków, Poland

✉ madeleine.waller@kcl.ac.uk (M. Waller); odinaldo.rodrigues@kcl.ac.uk (O. Rodrigues);

oana.cocarascu@kcl.ac.uk (O. Cocarascu)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

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 evaluations). 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, 33, 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

it. 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 methods¹ 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.

¹A 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 disparity³ 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}: Access to sensitive attributes The number of sensitive attributes available, and whether those attributes can be binary, multi-valued, or numerical impacts the applicability of methods.

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

⁴We denote recommendations related to applicability as R_a and to regulations as R_r .

R_{a2}: *Applicability to models* Whether a method can be applied in a scenario may depend on the model and access to that model.

R_{a3}: *Variability of evaluation* Experiments on different datasets and models, using different fairness metrics can give very different results. To understand a method’s effectiveness in various scenarios, it should be thoroughly evaluated.

R_{a4}: *Trade-offs after application of method* The impact on the system’s performance and differences in using various fairness notions/metrics.

R_{r1}: *Conflicts with legal definitions of fairness* Legal definitions of fairness may differ with technical definitions used in ML. Any new method should include which legal definitions of fairness it does (not) satisfy.

R_{r2}: *Data protection issues* Concerns about data protection and privacy may mean it is not realistic to assume sensitive attributes have been collected or stored.

R_{r3}: *Potential for positive discrimination* While bias mitigation methods aim to prevent negative discrimination, they may unintentionally cause illegal positive discrimination.

R_{r4}: *Transparency and explainability rights* The right to explainability and the transparency of a system are increasingly recognised as important aspects of legal and ethical AI. New bias mitigation methods should work towards this goal, not against it.

In order to consider the context of a system and not rely on pre-specified sensitive attributes which may not be available, we suggest exploring explainable AI (XAI) methods. Stakeholders can then consider the reasoning for the system’s classification and decide whether it is fair. There is a plethora of research into XAI, but its use cases in fairness are limited [20, 75, 76, 77]. There might still be issues with a stakeholder evaluating the system’s fairness but the concept of fairness is inherently human-oriented, context-specific and culturally dependent [78], meaning that it is difficult to automate, cannot be reduced to a metric, and requires some level of human input also for accountability purposes [79].

Overall, there is a need for cross-disciplinary considerations on issues such as fairness. Whilst not a new recommendation [8, 13, 18, 80, 81], it has not yet been universally adopted in current research around fairness and bias in AI.

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

⁵<https://safeandtrustedai.org>

References

- [1] Shared Intelligence and Local Government Association, Using predictive analytics in local public services, 2020. URL: <https://www.local.gov.uk/using-predictive-analytics-local-public-services>.
- [2] C. E. Church, A. J. Fairchild, In search of a silver bullet: child welfare's embrace of predictive analytics, *Juvenile and Family Court Journal* 68 (2017) 67–81.
- [3] Deloitte, AI and Automated Decision Making, 2021. URL: <https://www2.deloitte.com/uk/en/pages/deloitte-analytics/articles/ai-and-automated-decision-making.html>.
- [4] Z. Engin, P. C. Treleaven, Algorithmic government: Automating public services and supporting civil servants in using data science technologies, *Computer Journal* 62 (2019) 448–460. doi:10.1093/comjnl/bxy082.
- [5] Northpointe, Practitioner's Guide to COMPAS Core, 2019. URL: <https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf>.
- [6] J. Larson, S. Mattu, L. Kirchner, J. Angwin, How We Analyzed the COMPAS Recidivism Algorithm, 2016. URL: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- [7] J. Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, 2018. URL: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.
- [8] S. Wachter, B. D. Mittelstadt, C. Russell, Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI, *Comput. Law Secur. Rev.* 41 (2021) 105567. doi:10.1016/j.clsr.2021.105567.
- [9] M. Waller, P. Waller, Why predictive algorithms are so risky for public sector bodies, 2020. URL: <https://tinyurl.com/SoRisky>.
- [10] M. Waller, O. Rodrigues, O. Cocarascu, Bias mitigation methods for binary classification decision-making systems: Survey and recommendations, *arXiv preprint* (2023). arXiv:2305.20020.
- [11] D. Pessach, E. Shmueli, A review on fairness in machine learning, *ACM Comput. Surv.* 55 (2023) 51:1–51:44. doi:10.1145/3494672.
- [12] J. Dunkelau, Fairness-aware machine learning: An extensive overview, 2020. URL: https://www.sozwiss.hhu.de/fileadmin/redaktion/Fakultaeten/Philosophische_Fakultaet/Sozialwissenschaften/Kommunikations-_und_Medienwissenschaft_I/Dateien/Dunkelau__Leuschel__2019__Fairness-Aware_Machine_Learning.pdf.
- [13] M. Hort, Z. Chen, J. M. Zhang, F. Sarro, M. Harman, Bias mitigation for machine learning classifiers: A comprehensive survey, *arXiv preprint* (2022). arXiv:2207.07068.
- [14] The United States Department of Justice, The Equal Credit Opportunity Act, 2015. URL: <https://www.justice.gov/crt/equal-credit-opportunity-act-3>.
- [15] The United States Department of Justice, The Fair Housing Act, 2015. URL: <https://www.justice.gov/crt/fair-housing-act-1>.
- [16] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, *arXiv preprint* (2018).

arXiv:1810.01943.

- [17] T. L. Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsi, A survey on datasets for fairness-aware machine learning, *WIREs Data Mining and Knowledge Discovery* 12 (2022). doi:10.1002/widm.1452.
- [18] T. Van Nuenen, X. Ferrer, J. M. Such, M. Coté, Transparency for whom? Assessing discriminatory Artificial Intelligence, *Computer* 53 (2020) 36–44.
- [19] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel, Fairness through awareness, in: *Innovations in Theoretical Computer Science*, 2012, pp. 214–226. doi:10.1145/2090236.2090255.
- [20] J. Chakraborty, K. Peng, T. Menzies, Making fair ML software using trustworthy explanation, in: *35th IEEE/ACM International Conference on Automated Software Engineering, ASE, IEEE*, 2020, pp. 1229–1233. doi:10.1145/3324884.3418932.
- [21] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 329–338. doi:10.1145/3287560.3287589.
- [22] G. Ristanoski, W. Liu, J. Bailey, Discrimination aware classification for imbalanced datasets, in: *22nd ACM International Conference on Information and Knowledge Management, CIKM*, 2013, pp. 1529–1532. doi:10.1145/2505515.2507836.
- [23] W. Zhang, E. Ntoutsi, FAHT: an adaptive fairness-aware decision tree classifier, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, 2019, pp. 1480–1486. doi:10.24963/ijcai.2019/205.
- [24] B. Fish, J. Kun, Á. D. Lelkes, A confidence-based approach for balancing fairness and accuracy, in: *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016, pp. 144–152. doi:10.1137/1.9781611974348.17.
- [25] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, R. Puri, Bias mitigation post-processing for individual and group fairness, in: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2019, pp. 2847–2851. doi:10.1109/ICASSP.2019.8682620.
- [26] T. Calders, A. Karim, F. Kamiran, W. Ali, X. Zhang, Controlling attribute effect in linear regression, in: *2013 IEEE 13th International Conference on Data Mining*, 2013, pp. 71–80. doi:10.1109/ICDM.2013.114.
- [27] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. J. Kearns, J. Morgenstern, S. Neel, A. Roth, A convex framework for fair regression, arXiv preprint (2017). arXiv:1706.02409.
- [28] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, M. Pontil, Fair regression via plug-in estimator and recalibration with statistical guarantees, in: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems NeurIPS*, 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/ddd808772c035aed516d42ad3559be5f-Abstract.html>.
- [29] P. Putzel, S. Lee, Blackbox post-processing for multiclass fairness, 2022. URL: http://ceur-ws.org/Vol-3087/paper_36.pdf.
- [30] W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. W. Michalak, S. Asoodeh, F. P. Calmon, Beyond adult and COMPAS: fairness in multi-class prediction, arXiv preprint (2022). arXiv:2206.07801.

- [31] F. Chierichetti, R. Kumar, S. Lattanzi, S. Vassilvitskii, Fair clustering through fairlets, in: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, 2017, pp. 5029–5037. URL: <https://proceedings.neurips.cc/paper/2017/hash/978fce5bcc4eccc88ad48ce3914124a2-Abstract.html>.
- [32] I. M. Ziko, J. Yuan, E. Granger, I. B. Ayed, Variational fair clustering, in: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, 2021, pp. 11202–11209. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/17336>.
- [33] M. Abbasi, A. Bhaskara, S. Venkatasubramanian, Fair clustering via equitable group representations, in: *FACCT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 504–514. doi:10.1145/3442188.3445913.
- [34] A. Backurs, P. Indyk, K. Onak, B. Schieber, A. Vakilian, T. Wagner, Scalable fair clustering, in: *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97 of *Proceedings of Machine Learning Research*, 2019, pp. 405–413. URL: <http://proceedings.mlr.press/v97/backurs19a.html>.
- [35] C. Rösner, M. Schmidt, Privacy preserving clustering with constraints, in: *45th International Colloquium on Automata, Languages, and Programming, ICALP*, volume 107 of *LIPICs*, 2018, pp. 96:1–96:14. doi:10.4230/LIPICs.ICALP.2018.96.
- [36] S. K. Bera, D. Chakrabarty, N. Flores, M. Negahbani, Fair algorithms for clustering, in: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS*, 2019, pp. 4955–4966. URL: <https://proceedings.neurips.cc/paper/2019/hash/fc192b0c0d270dbf41870a63a8c76c2f-Abstract.html>.
- [37] V. Iosifidis, E. Ntoutsis, FABBOO - online fairness-aware learning under class imbalance, in: *Discovery Science - 23rd International Conference, DS*, volume 12323 of *Lecture Notes in Computer Science*, 2020, pp. 159–174. doi:10.1007/978-3-030-61527-7_11.
- [38] A. Roy, J. Horstmann, E. Ntoutsis, Multi-dimensional discrimination in law and machine learning - A comparative overview, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FACCT 2023, Chicago, IL, USA, June 12-15, 2023, ACM*, 2023, pp. 89–100. doi:10.1145/3593013.3593979.
- [39] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, K. P. Gummadi, Fairness constraints: Mechanisms for fair classification, in: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 54, 2017, pp. 962–970. URL: <http://proceedings.mlr.press/v54/zafar17a.html>.
- [40] Y. Choi, G. Farnadi, B. Babaki, G. V. den Broeck, Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns, in: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, AAAI Press, 2020, pp. 10077–10084. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/6565>.
- [41] N. Quadrianto, V. Sharmanska, Recycling privileged learning and distribution matching for fairness, in: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, 2017, pp. 677–688. URL: <https://proceedings.neurips.cc/paper/2017/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html>.
- [42] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: *Proceedings of the 21th ACM SIGKDD*, 2015, pp. 259–268. doi:10.1145/2783258.2783311.
- [43] T. Calders, S. Verwer, Three naive bayes approaches for discrimination-free classification,

Data Mining and Knowledge Discovery 21 (2010) 277–292. doi:10.1007/s10618-010-0190-x.

- [44] Y. Bechavod, K. Ligett, Learning fair classifiers: A regularization-inspired approach, arXiv preprint (2017). arXiv:1707.00044.
- [45] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES, ACM, 2018, pp. 335–340. doi:10.1145/3278721.3278779.
- [46] F. Kamiran, T. Calders, M. Pechenizkiy, Discrimination aware decision tree learning, in: ICDM, The 10th IEEE International Conference on Data Mining, 2010, pp. 869–874. doi:10.1109/ICDM.2010.50.
- [47] T. Hu, V. Iosifidis, W. Liao, H. Zhang, M. Y. Yang, E. Ntoutsi, B. Rosenhahn, Fairnn - conjoint learning of fair representations for fair decisions, in: Discovery Science - 23rd International Conference, DS, volume 12323, 2020, pp. 581–595. doi:10.1007/978-3-030-61527-7_38.
- [48] S. Caton, C. Haas, Fairness in machine learning: A survey, arXiv preprint (2020). arXiv:2010.04053.
- [49] A. Aggarwal, P. Lohia, S. Nagar, K. Dey, D. Saha, Black box fairness testing of machine learning models, in: M. Dumas, D. Pfahl, S. Apel, A. Russo (Eds.), Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2019, Tallinn, Estonia, August 26-30, 2019, ACM, 2019, pp. 625–635. doi:10.1145/3338906.3338937.
- [50] Y. Savani, C. White, N. S. Govindarajulu, Intra-processing methods for debiasing neural networks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/1d8d70dddf147d2d92a634817f01b239-Abstract.html>.
- [51] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (2011) 1–33. doi:10.1007/s10115-011-0463-8.
- [52] Centre for Data Ethics and Innovation, Review into bias in algorithmic decision-making, 2020. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into_bias_in_algorithmic_decision-making.pdf.
- [53] V. Grari, B. Ruf, S. Lamprier, M. Detyniecki, Fair adversarial gradient tree boosting, in: 2019 IEEE International Conference on Data Mining, ICDM, 2019, pp. 1060–1065. doi:10.1109/ICDM.2019.00124.
- [54] V. Iosifidis, E. Ntoutsi, Adafair: Cumulative fairness adaptive boosting, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM, 2019, pp. 781–790. doi:10.1145/3357384.3357974.
- [55] E. Krasanakis, E. S. Xioufis, S. Papadopoulos, Y. Kompatsiaris, Adaptive sensitive reweighting to mitigate bias in fairness-aware classification, in: Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW, 2018, pp. 853–862. doi:10.1145/3178876.3186133.
- [56] K. Mancuhan, C. Clifton, Combating discrimination using bayesian networks, Artificial Intelligence and Law 22 (2014) 211–238. doi:10.1007/s10506-014-9156-4.

- [57] F. Kamiran, I. Zliobaite, T. Calders, Quantifying explainable discrimination and removing illegal discrimination in automated decision making, *Knowledge and Information Systems* 35 (2013) 613–644. doi:10.1007/s10115-012-0584-8.
- [58] I. Zliobaite, F. Kamiran, T. Calders, Handling conditional discrimination, in: 11th IEEE International Conference on Data Mining, ICDM, 2011, pp. 992–1001. doi:10.1109/ICDM.2011.72.
- [59] R. L. Cardoso, W. M. Jr., V. A. F. Almeida, M. J. Zaki, A framework for benchmarking discrimination-aware models in machine learning, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES, 2019, pp. 437–444. doi:10.1145/3306618.3314262.
- [60] V. Iosifidis, E. Ntoutsis, Dealing with bias via data augmentation in supervised learning scenarios, *Jo Bates Paul D. Clough Robert Jäschke* 24 (2018). URL: https://ceur-ws.org/Vol-2103/paper_5.pdf.
- [61] L. Oneto, M. Donini, A. Elders, M. Pontil, Taking advantage of multitask learning for fair classification, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES, 2019, pp. 227–237. doi:10.1145/3306618.3314255.
- [62] European Commission, Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>.
- [63] A. Ruoss, M. Balunovic, M. Fischer, M. T. Vechev, Learning certified individually fair representations, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems NeurIPS, 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/55d491cf951b1b920900684d71419282-Abstract.html>.
- [64] S. Schelter, Y. He, J. Khilnani, J. Stoyanovich, Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions, in: Proceedings of the 23rd International Conference on Extending Database Technology, EDBT, 2020, pp. 395–398. doi:10.5441/002/edbt.2020.41.
- [65] F. Kamiran, A. Karim, X. Zhang, Decision theory for discrimination-aware classification, in: 12th IEEE International Conference on Data Mining, ICDM, 2012, pp. 924–929. doi:10.1109/ICDM.2012.45.
- [66] A. Xiang, I. D. Raji, On the legal compatibility of fairness definitions, arXiv preprint (2019). arXiv:1912.00761.
- [67] R. Xenidis, Tuning EU equality law to algorithmic discrimination: Three pathways to resilience, *Maastricht Journal of European and Comparative Law* 27 (2020) 736–758. doi:10.1177/1023263X20982173.
- [68] GDPR, General Data Protection Regulation (GDPR) – Official Legal Text, 2016. URL: <https://gdpr-info.eu/>.
- [69] Centre for Data Ethics and Innovation, Enabling responsible access to demographic data to make AI systems fairer, 2023. URL: <https://www.gov.uk/government/publications/enabling-responsible-access-to-demographic-data-to-make-ai-systems-fairer/report-enabling-responsible-access-to-demographic-data-to-make-ai-systems-fairer#executive-summary>.
- [70] M. A. Haeri, K. A. Zweig, The crucial role of sensitive attributes in fair classification, in: IEEE Symposium Series on Computational Intelligence, SSCI, 2020, pp. 2993–3002.

doi:10.1109/SSCI47803.2020.9308585.

- [71] S. Wachter, B. Mittelstadt, C. Russell, Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law, *W. Va. L. Rev.* 123 (2020) 735. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772.
- [72] UK Public General Acts, Equality Act 2010, 2010. URL: <https://www.legislation.gov.uk/ukpga/2010/15/contents>.
- [73] S. Wachter, B. Mittelstadt, L. Floridi, Why a right to explanation of automated decision-making does not exist in the general data protection regulation, *International Data Privacy Law* 7 (2017) 76–99. doi:10.1093/idpl/ix005.
- [74] J. Schöffner, Y. Machowski, N. Kuehl, A study on fairness and trust perceptions in automated decision making, in: D. Glowacka, V. R. Krishnamurthy (Eds.), *Joint Proceedings of the ACM IUI 2021 Workshops co-located with 26th ACM Conference on Intelligent User Interfaces (ACM IUI 2021)*, College Station, United States, April 13-17, 2021, volume 2903 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: <https://ceur-ws.org/Vol-2903/IUI21WS-TESS-12.pdf>.
- [75] T. Begley, T. Schwedes, C. Frye, I. Feige, Explainability for fair machine learning, *arXiv preprint* (2020). arXiv:2010.07389.
- [76] P. A. Grabowicz, N. Perello, A. Mishra, Marrying fairness and explainability in supervised learning, in: *FACt '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul, Republic of Korea, June 21 - 24, 2022, ACM, 2022, pp. 1905–1916. doi:10.1145/3531146.3533236.
- [77] R. Calegari, F. Sabbatini, The psyche technology for trustworthy artificial intelligence, in: A. Dovier, A. Montanari, A. Orlandini (Eds.), *AIxIA 2022 – Advances in Artificial Intelligence*, Springer International Publishing, Cham, 2023, pp. 3–16.
- [78] C. Barabas, C. Doyle, J. B. Rubinovitz, K. Dinakar, Studying up: reorienting the study of algorithmic fairness around issues of power, in: M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, G. Zanfir-Fortuna (Eds.), *FAT* '20: Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, January 27-30, 2020, ACM, 2020, pp. 167–176. doi:10.1145/3351095.3372859.
- [79] M. Veale, M. V. Kleek, R. Binns, Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making, in: R. L. Mandryk, M. Hancock, M. Perry, A. L. Cox (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018*, Montreal, QC, Canada, April 21-26, 2018, ACM, 2018, p. 440. doi:10.1145/3173574.3174014.
- [80] X. F. Aran, T. van Nuenen, J. M. Such, M. Côté, N. Criado, Bias and discrimination in AI: A cross-disciplinary perspective, *IEEE Technol. Soc. Mag.* 40 (2021) 72–80. doi:10.1109/MTS.2021.3056293.
- [81] L. Cheng, K. R. Varshney, H. Liu, Socially responsible AI algorithms: Issues, purposes, and challenges, *J. Artif. Intell. Res.* 71 (2021) 1137–1181. doi:10.1613/jair.1.12814.