CEUR-WS.org/Vol-3442/paper-11.pdf

Explainability Methods to Detect and Measure Discrimination in Machine Learning Models

Sofie Goethals[1,*], David Martens[1] and Toon Calders[2]

[1]Department of Engineering Management, University of Antwerp, Antwerp, 2000, Belgium

[2]Department of Computer Science, University of Antwerp, Antwerp, 2000, Belgium

Abstract

Today, it is common to use machine learning models for high-stakes decisions, but this can pose a threat to fairness as these models can amplify bias present in the dataset. At the moment, there is no consensus on a universal method to tackle this, and we argue that this is also not possible as the right method will depend on the context of each case. As a solution, our aim was to bring transparency in the fairness domain, and in earlier work, we proposed a counterfactual-based algorithm ($PreCoF$) to identify bias in machine learning models. This method attempts to counter the disagreement problem in Explainable AI, by reducing the flexibility of the model owner. We envision a future where transparency tools such as the latter are used to perform fairness audits by independent auditors who can judge for each case whether the audit revealed discriminatory patterns or not. This approach would be more in line with the current nature of EU legislation, as its requirements are often too contextual and open to judicial interpretation to be automated.

Artificial Intelligence (AI) is making decisions in more and more high-stakes domains of our life, such as justice, finance, education and healthcare. There is growing concern that the algorithms that generate these judgments may unintentionally encode and potentially exacerbate human bias as the influence and scope of these decisions expands [1]. This is why it is of huge importance to understand the decisions models are making and to ensure they are fair. We focus on fairness in classification, where the goal is to prevent discrimination against people based on their membership in a sensitive group, without compromising the utility of the classifier [2, 3]. Different statistical techniques are currently available to evaluate discrimination in machine learning models; however, they make implicit assumptions about the nature of bias in the data. There is a clear gap between these statistical measures of fairness and the context-sensitive and often intuitive metrics used by the European Court of Justice (ECJ) [4]. The right method to apply will be case-dependent and often policy-related, and the purpose of the data scientist should not be to make this call, but instead to make the nature of the algorithmic discrimination more transparent to support policymakers and legal scholars in decision making. As other authors have already argued [4, 5], it is misguided to focus on fairness without first obtaining transparency, as it is not fair that life-changing decisions would be made without entitlement to an explanation.

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

*Corresponding author.

sofie.goethals@uantwerpen.be (S. Goethals); david.martens@uantwerpen.be (D. Martens); toon.calders@uantwerpen.be (T. Calders)

0000-0003-3784-826X (S. Goethals)

© 2022 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)

In an earlier work [6], we proposed a counterfactual-based algorithm to identify unfairness in response to the request for more transparency in the fairness domain, as stated by Wachter et al. (2021) and Rudin et al. (2018) [4, 5]. Counterfactual explanations form the basis of an important class of explainable AI methods [7], and are defined as the smallest modification to a data instance that results in a different classification outcome [8, 9]. We named this metric $PreCoF$, which stands for *Predictive Counterfactual Fairness* [6]. We distinguish between *explicit bias*, which occurs when the model directly uses the sensitive attribute, and *implicit bias*, when there is a neutral attribute that substantially disadvantages the protected group. These are also known as *direct* and *indirect discrimination* respectively. Numerous legislations, such as the GDPR, focus on explicit bias by forbidding the collection and use of socially sensitive features in the decision-making model [10, 11]. However, given that any sufficiently rich data set is likely to contain proxy variables that have a strong correlation with the sensitive attributes, our findings and previous research indicate that simply eliminating these variables is ineffective [12].

Let us first situate our methodology into existing literature about using explainability techniques to measure discrimination: Kusner et al. introduced Counterfactual Fairness, which studies fairness-aware machine learning from a causal perspective [13]. A major drawback with this method is that you have to assume that the causal relations between all the variables in a dataset are known, while in reality this is often not the case. Other researchers use counterfactual explanations to assess fairness by focusing on the distance to the counterfactual instance, which thus assesses whether the effort to reach the required outcome is equal across groups [14, 15]. We will move away from the *algorithmic recourse* literature and not focus on *plausible* and *actionable* counterfactual explanations, because they can actually conceal bias in our case. For example, the counterfactual explanation to *'Change your native language to English'* is both not actionable and not plausible (for certain population groups), but this is exactly the kind of explanation we are interested in to identify bias. Sokol et al. (2019) suggest using counterfactual explanations to identify explicit bias at the individual level, by looking for explanations that include one protected attribute change [16]. Lastly, the use of Shapley values to identify algorithmic fairness has also been studied [17, 18]. The crucial difference between counterfactual explanations and Shapley values is that the former explain a decision and the latter a prediction score; we focus on fair decision making and hence use counterfactual explanations. Our results show that both techniques can indeed result in fairly different results.

Our algorithm can be used both to detect explicit bias, by searching for explanations that only contain the sensitive attributes, as well as implicit bias, by comparing the counterfactual explanations of different sensitive groups and determining which attributes are more frequently responsible for a negative decision for each group.

In a previous study [6], we applied this algorithm to assess explicit and implicit bias in models trained on tabular datasets that are well known in fairness-aware machine learning research [19]. We can give an example of something our metric detected when it was applied on the Catalonia juvenile dataset, a dataset of juvenile offenders that is used to predict recidivism (where foreign status is the sensitive attribute that is used to measure explicit bias). When investigating the explicit bias, our method found that the explanation *'If you would have been a local instead of a foreigner, you would have been predicted to not reoffend'* is present for 25% of foreigners, while the reverse explanation: (*'If you would have been a foreigner instead of a local, you would have been predicted to not reoffend'*) is never present. This implies that, if all other features were equal, 25% of foreigners who have been predicted as likely to reoffend, would have been predicted as not likely to reoffend, just by changing their foreign status. This shows an example of explicit bias, but our metric also allows us to look at implicit bias. When we remove foreign from the dataset, and retrain the machine learning model again to measure implicit bias, we find that foreigners are advised to change their *national group* more frequently than locals. This is a clear proxy for foreign status and should have also been removed when race attributes are not allowed, and $PreCoF$ can be useful for flagging these proxy attributes.

In this case, it might have been straightforward to detect the proxies right away, but in other situations, intuition may fail us. After all, we cannot assume that automated systems will discriminate in the same ways as people do: new and counterintuitive proxies for traditionally protected attributes can emerge, but will not necessarily be detected [4]. If such an attribute is found that substantially disadvantages the protected group, this is not necessarily a problem: Some attributes can be justified, depending on the context of the case and relevant legislation. *Justified indirect discrimination* occurs when the 'proportionality test' is passed, meaning that this attribute is both legally necessary and proportionate [20]. Our algorithm was created with this idea in mind: can we find the attributes that explain why sensitive groups are more often predicted with a negative outcome? A discussion on whether or not these attributes are justifiable can follow from this. This methodology is more in line with the current nature of EU legislation than the statistical fairness metrics that are currently in use. The current requirements of the EU are too contextual, reliant on intuition and open to judicial interpretation to be automated and legal scholars emphasize that an one size-fits-all solution is not applicable to algorithmic fairness, but that an approach that provides transparency into the context of an algorithm, can guarantee a fairer outcome [21, 4].

However, replacing statistical fairness metrics with transparency methods does open up the risk of misinterpretation or manipulation by the owner of the machine learning model. As we see in earlier research, and as supported by experimental results, different explainability methods can yield significantly different results, often in disagreement with each other [22]. Moreover, even a single explanation method can produce a multitude of possible explanations, depending on the choice of parameters [23]. In an adversarial situation, where the model owner acts as the adversary, this flexibility allows them to selectively choose and present explanations that conceal biases [24, 23]. Furthermore, financial incentives could potentially lead to the creation of fabricated explanations [25].

To address these issues, we proposed a technique that eliminates the reliance on modifiable input parameters. Our approach involves conducting a greedy search over all possible explanations, independent of their order of return. By exploring the entire space of explanations, we aim to minimize the influence of model parameters and the model owners' manipulation.

In addition, we believe that the responsibility for verifying model fairness should not rest solely with the model owner. Given their vested interest in the outcome, they may have incentives to overlook discriminatory biases. Instead, we envision a future where transparency tools like $PreCoF$ are used for fairness audits conducted by independent third-party auditors, in line with Raji et al. [26]. These auditors would possess the necessary expertise to assess the context of each case and determine whether the audit reveals discriminatory patterns. The question of whether the identified patterns are justified or not should be resolved through collaboration with Member States courts and the ECJ [27]. Implementing procedures like these would assist companies in adhering to the General Data Protect Regulation (GDPR) which mandates fair, transparent, and accountable automated decision-making processes. Moreover, such measures can foster trust in the decision-making processes of these companies.

By employing independent auditors and involving legal and regulatory authorities, we can establish a more robust and unbiased system for evaluating fairness in machine learning models. our approach reduces the potential for manipulation, ensures comprehensive exploration of explanations, and facilitates the enforcement of fairness standards in line with legal requirements and regulations.

Acknowledgments

Funding was provided by Research Foundation – Flanders (Grant No.11N7723N) and by Flanders AI Research Program.

References

[1] S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning, arXiv preprint arXiv:1808.00023 (2018).

[2] S. Caton, C. Haas, Fairness in machine learning: A survey, arXiv preprint arXiv:2010.04053 (2020).

[3] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012, pp. 214–226.

[4] S. Wachter, B. Mittelstadt, C. Russell, Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and AI, Computer Law & Security Review 41 (2021) 105567.

[5] C. Rudin, C. Wang, B. Coker, The age of secrecy and unfairness in recidivism prediction, arXiv preprint arXiv:1811.00731 (2018).

[6] S. Goethals, D. Martens, T. Calders, PreCoF: counterfactual explanations for fairness, Machine Learning (2023) 1–32.

[7] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access 6 (2018) 52138–52160.

[8] D. Martens, F. Provost, Explaining data-driven document classifications, MIS quarterly 38 (2014) 73–100.

[9] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[10] G. M. Johnson, Algorithmic bias: on the implicit biases of social technology, Synthese 198 (2021) 9941–9961.

[11] M. van Bekkum, F. Z. Borgesius, Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception?, Computer Law & Security Review 48 (2023) 105770.

[12] P. T. Kim, Auditing algorithms for discrimination, U. Pa. L. Rev. Online 166 (2017) 189.

[13] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, Advances in Neural Information Processing Systems 30 (2017).

[14] S. Sharma, J. Henderson, J. Ghosh, CERTIFAI: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models, arXiv preprint arXiv:1905.07857 (2019).

[15] J. von Kügelgen, A.-H. Karimi, U. Bhatt, I. Valera, A. Weller, B. Schölkopf, On the fairness of causal algorithmic recourse, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 9584–9594.

[16] K. Sokol, R. Santos-Rodriguez, P. Flach, FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency, arXiv preprint arXiv:1909.05167 (2019).

[17] J. M. Hickey, P. G. Di Stefano, V. Vasileiou, Fairness by explicability and adversarial shap learning, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III, Springer, 2021, pp. 174–190.

[18] M. Mase, A. B. Owen, B. B. Seiler, Cohort shapley value for algorithmic fairness, arXiv preprint arXiv:2105.07168 (2021).

[19] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsi, A survey on datasets for fairness-aware machine learning, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2022) e1452.

[20] 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.

[21] A. Elyounes, et al., Contextual fairness: A legal and policy analysis of algorithmic fairness, Journal of Law, Technology and Policy, Forthcoming (2019).

[22] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, H. Lakkaraju, The disagreement problem in explainable machine learning: A practitioner's perspective, arXiv preprint arXiv:2202.01602 (2022).

[23] S. Bordt, M. Finck, E. Raidl, U. von Luxburg, Post-hoc explanations fail to achieve their purpose in adversarial contexts, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 891–905.

[24] S. Barocas, A. D. Selbst, M. Raghavan, The hidden assumptions behind counterfactual explanations and principal reasons, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 80–89.

[25] T. Greene, S. Goethals, D. Martens, G. Shmueli, Monetizing explainable ai: A double-edged sword, arXiv preprint arXiv:2304.06483 (2023).

[26] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 33–44.

[27] P. Hacker, Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law, Common Market Law Review 55 (2018).