Fairness in Machine Learning as ‘Algorithmic Positive Action’

Jan-Laurin Müller [1]

[1] *University of Bayreuth, Universitätsstraße 30, Bayreuth, 95447, Germany*

Abstract

In recent years the interdisciplinary discourse on ‘fairness in machine learning’ has produced a vast amount of technical approaches to guarantee non-discrimination in algorithmic decision-making “by design”. These fairness criteria and definitions have already been examined for their compatibility with moral and legal norms alike. However, non-discrimination law may not only require computer scientists to de-bias their algorithms. It may also limit the implementation of such fairness ensuring techniques: Both, EU-Member States and private actors using de-biasing and fairness-enhancing methods could be liable of violating the principle of equal treatment by undertaking ‘algorithmic positive action’. The article considers the legality of ‘fairness in machine learning’ from the perspective of EU positive action doctrine as well as U.S. affirmative action jurisprudence. It thereby makes three contributions to the debate about how to best enable algorithmic fairness: First, it demonstrates that private actors introducing technical fairness considerations to their models may be liable for executing “algorithmic positive action” under EU law. To do so, the paper puts the jurisprudence of the Court of Justice of the European Union in context of approaches from computer science literature. The paper’s second contribution is to help bridging the gap between computer science and law in the field of fair machine learning. It thereby offers normative guidance for computer science scholars and practitioners trying to implement fairness considerations into their models. Finally, the paper compares EU “positive action” doctrine with U.S. “affirmative action” jurisprudence regarding algorithmic decision making. It demonstrates both, similarities as well as differences across the Atlantic and shows that legal requirements for technological ‘fairness by design’ approaches are highly contextual and depended on socio-cultural settings. Building on these insights, the paper argues for a genuine European perspective on ‘fairness in machine learning’.

Keywords

Algorithmic Discrimination, EU Non-Discrimination Law, Affirmative Action, Positive Action

1. Introduction

The increasing use of algorithmic-decision-making-systems based on machine learning (ADM-systems) has triggered a broad and interdisciplinary debate on the ‘ethics of algorithms’ (Mittelstadt et al. 2016; Tsamados et al. 2021; Martens 2022). At the core of this debate lies the call for fairness, accountability and transparency in machine learning, with fairness comprising the values of privacy protection and non-discrimination. The potential of ADM-systems to discriminate against legally protected groups has been demonstrated across a wide range of social contexts, ranging from criminal risk assessment (Angwin et al. 2016), employment (Dastin 2018), voice and image recognition (Shankar et al. 2017; Buolamwini/Gebru 2018) to text processing and translation (Bolukbasi et al. 2016; Caliskan et al. 2017; Garg et al. 2018). Summarizing the debate, the American AI Now Institute stated famously: “The question is no longer whether there are harms and biases in AI systems. That debate has been settled: the evidence has mounted beyond doubt […]. The next task now is addressing these harms.“

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
EMAIL: jan-laurin.mueller@uni-bayreuth.de (A. 1)

(Whittaker et al. 2018, p. 42). Legal scholars on both sides of the Atlantic have demonstrated that legal non-discrimination regimes may fall short in doing so (see Barocas/Selbst 2016 for the U.S. and Hacker 2018 for the EU). Computer scientists are trying to fill this gap. By developing theoretical fairness definitions and bias metrics (cf. Hutchinson/Mitchell 2018; Verma/Rubin 2018; Dunkelau/Leuschel 2019; Chouldechova/Roth 2020; Pessach/Shmueli 2020) as well as practical bias detection and bias diminishing techniques (Bellamy et al. 2018; Saleiro et al. 2019) they try to ensure that ADM-systems match ethical and legal fairness considerations "by design". These fairness criteria and definitions have already been examined for their compatibility with moral norms (Hellman 2020) and applicable non-discrimination regimes (Wachter et al. 2021a; Wachter et al. 2021b; Hauer et al. 2021).

However, non-discrimination law may not only require computer scientists to de-bias their algorithms. It may also *limit* the implementation of such fairness ensuring techniques: Public and private actors using de-biasing and fairness-enhancing methods could be liable of violating the principle of equal treatment, both under EU positive action doctrine as well as under U.S. affirmative action jurisprudence. Affirmative and positive action alike describe a variety of measures going beyond simply refraining from discrimination, but rather serving to promote substantial equality for groups that have suffered particular disadvantages in the past. Such measures range from mere outreach schemes (like selective advertising for protected groups) all the way up to reverse discrimination, such as hiring quotas (for a taxonomy see McCrudden 2011). According to EU anti-discrimination law, Member States or private actors may implement such approaches to achieve 'full equality of opportunity' even though they conflict with a purely formal conception of equal treatment. The article argues that this conflict may arise in algorithmic decision making as well: two individuals who differ regarding a protected attribute, may receive different scores and therefore be treated unequally even though they would have been treated equally before the intervention of the adjusted model. This could be considered "algorithmic reverse discrimination" and violate the principle of equal treatment. Differing from the U.S. context, where the issue is discussed as "algorithmic affirmative action" (Bent 2020; Ho/Xiang 2020, Kim 2022), in Europe the possibility of fairness in machine learning (FML) leading to "algorithmic positive action" has not yet been considered. The article therefore aims to answer two questions: (1.) Does EU non-discrimination law limit the use of technical de-biasing methods? (2.) If so, to what extent?

2. The Legality of 'Algorithmic Positive Action'

By answering these questions, the article makes three contributions to the debate about how to best enable algorithmic fairness:

2.1. 'Algorithmic Positive Action' Under EU Law

First, it demonstrates that private actors introducing technical fairness considerations to their models may be liable for executing "algorithmic positive action" under EU law. To do so, the paper analyzes the jurisprudence of the Court of Justice of the European Union (CJEU) and puts it in context of approaches from computer science literature that systematize technical fairness metrics. The paper particularly refers to the distinction between group- and individual-fairness (cf. Verma/Rubin 2018; Barocas/Hardt/Narayanan 2019, Pessach/Shmueli 2020; critical Binns 2020). It finds that 'group-fairness metrics' in particular may violate the non-discrimination principle. Unlike 'individual-fairness metrics' they aim to equal the distribution of certain statistical variables between groups and do not solely focus on individual persons. Thus, they closely resemble quota systems, which have been highly controversial in the CJEU's jurisprudence. In *Badeck* (C-158/97 – Badeck), the Court famously held that "a measure which is intended to give priority in promotion to women in sectors of the public service where they are under-represented must be regarded as compatible with Community law if […] [1.] women and men are equally qualified, and [2.] the candidatures are the subject of an objective assessment which takes account of the specific personal situations of all candidates." Two aspects are decisive: First, the corrective measures must consider some notion of qualification (in employment), creditworthiness (in banking) or risk (in insurance). This is challenging for certain fairness-metrics, foremost for (pure) 'statistical parity'. Second, an objective case-by-case examination needs to ensure that the disadvantaged person's interests are sufficiently taken into account. The CJEU therefore considered strict quotas to violate the non-discrimination principle because they pursue a notion of equality of results (C-450/93 – Kalanke) and flexible quotas pursuing equality of chances to be lawful (C-409/95 – Marschall). Users of ADM-systems could therefore be required to couple algorithmic fairness measures with a final evaluation leaving the ultimate decision of whether or not to adopt the corrective measure to a human (Hacker 2018, p. 1181). This would be a great challenge for the project of ensuring 'fairness *by design*'. However, in *Badeck* the CJEU made an exception to its strict approach and allowed rigid quotas for traineeship positions and job interviews (C-158/97 – Badeck): Employers were permitted to allocate a fixed number of traineeship positions to women and invite a fixed number of them to job interviews. The paper shows that these exceptions seamlessly fit into the Court's context-based notion of substantive equal opportunity because they were considered mere preconditions for the access to the labor market. By doing so, the article offers normative guidance for computer science scholars and practitioners trying to implement fairness considerations into their models. Its second contribution therefore is to help bridging the gap between computer sciences and law in the field of fair machine learning.

2.2. Comparison with U.S. Affirmative Action Doctrine

Finally, the paper compares EU "positive action" doctrine with U.S. "affirmative action" jurisprudence regarding algorithmic decision making. In *Ricci v. DeStefano* (557 U.S. 557 – Ricci v. DeStefano) the Supreme Court (SC) held that Title VII does not prohibit an employer from considering racial disparities "before administering a test or practice" (cf. 539 U.S. 244 – Gratz v. Bollinger; 539 U.S. 306 – Grutter v. Bollinger). But once the test or practice has been established, the equal treatment principle sharply disapproves of any altering of the result on grounds of race (see Primus 2010, p. 1369-1374; Siegel 2015, p. 682-683; Bagenstos 2016, p. 1151). These requirements resemble the technical distinction between pre-processing, in-processing and post- processing methods (cf. Dunkelau/Leuschel 2019, Martens 2022). This reading of the case law enables the article to demonstrate that under U.S. non-discrimination law, post-processing methods will likely be subject to stricter scrutiny than pre-processing and in-processing approaches. However, even pre- and in-processing methods will largely be considered illegal. This is because in the U.S., decision-making processes (like college admissions) are increasingly required not to consider protected attributes at all. This development will probably continue with the Supreme Court's decision in *Students for Fair Admissions v. President and Fellows of Harvard College* and *Students for Fair Admissions v. University of North Carolina* which is expected to overrule existing precedents (for the arguments in the oral hearing see Liptak 2022).

Despite similar historical starting points, EU positive action doctrine and U.S. affirmative action jurisprudence have taken opposite routes. While EU non-discrimination law allows for a contextual and substantive reading of equality (of opportunity), recent American jurisprudence (cf. Areheart 2012; Bagenstos 2016) opted for a rather formal notion of equality. Considering protected attributes may be a legitimate strategy in breaking down historical structures of oppression against protected groups under EU but not under U.S. non-discrimination law. The article therefore demonstrates both, similarities as well as differences across the Atlantic and thereby shows that legal requirements for technological "fairness by design" approaches are highly contextual and depended on socio-cultural settings.

3. The Necessity of a European Perspective

Building on these insights, the paper argues for a genuine European perspective on 'fairness in machine learning'. It provides a normative point of view on the workshop's research agenda: Instead of asking "what, if anything, *is* specifically European in the debate about fairness in machine learning?" it tackles the question "why *should* there *be* a specifically European debate about fairness in machine learning?".

4. Acknowledgements

5. References

[1] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] B. Areheart, The Anticlassification Turn in Employment Discrimination Law, Alabama Law Review 63 (2012) 955-1006. URL: https://ir.law.utk.edu/cgi/viewcontent.cgi?article=1223&context=utklaw_facpubs.

[3] S. Bagenstos, Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After Inclusive Communities, Cornell Law Review 101 (2016) 1115-1169. URL: https://repository.law.umich.edu/cgi/viewcontent.cgi?article=2822&context=articles.

[4] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning - Limitations and Opportunities, 2019. URL: https://fairmlbook.org.

[5] S. Barocas, A. Selbst, Big Data's Disparate Impact, California Law Review 104 (2016) 671-732. doi:10.15779/Z38BG31.

[6] R. Bellamy et al., Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, arXiv:1810.01943 (2018). doi: 10.48550/arXiv.1810.01943.

[7] J. Bent, Is Algorithmic Affirmative Action Legal?, The Georgetown Law Journal 108 (2020) 803-853. URL: https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2020/04/Is-Algorithmic-Affirmative-Action-Legal.pdf.

[8] R. Binns, On the apparent conflict between individual and group fairness, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, ACM Press, New York, NY, 2020. doi:10.1145/3351095.3372864.

[9] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, A. Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, ACM Press, New York, NY, 2016. doi:10.5555/3157382.3157584.

[10] J. Buolamwini, T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency 2018, FAT*'18, ACM Press, New York, NY, 2018. URL: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[11] A. Caliskan, J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183-186. doi:10.1126/science.aal4230.

[12] A. Chouldechova, A. Roth, A Snapshot of the Frontiers of Fairness in Machine Learning, Communications of the ACM 63 (2020) 82-89. doi:10.1145/3376898.

[13] 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.

[14] J. Dunkelau, M. Leuschel, Fairness-Aware Machine Learning. An Extensive Overview, 2019. URL: https://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Sozialwissenschaften/Kommunikations-_und_Medienwissenschaft/KMW_I/Working_Paper/Dunkelau___Leuschel__2019__Fairness-Aware_Machine_Learning.pdf.

[15] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proceedings of the National Academy of Sciences of the United States of America 115 (2018) E3635-E3644. doi:10.1073/pnas.1720347115.

[16] P. Hacker, Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law, Common Market Law Review 55 (2018) 1143-1185.

[17] M. Hauer, J. Kevekordes, M. A. Haeri, Legal perspective on possible fairness measures – A legal discussion using the example of hiring decisions, Computer Law & Security Review 42 (2021). doi: 10.1016/j.clsr.2021.105583.

[18] D. Hellman, Measuring Algorithmic Fairness, Virginia Law Review 106 (2020) 811-866. URL: https://www.virginialawreview.org/wp-content/uploads/2020/06/Hellman_Book.pdf.

[19] D. Ho, A. Xiang, Affirmative Algorithms: The Legal Grounds for Fairness as Awareness, The University of Chicago Law Review Online, 2020. URL: https://lawreviewblog.uchicago.edu/2020/10/30/aa-ho-xiang/.

[20] B. Hutchinson, M. Mitchell, 50 Years of Test (Un)fairness: Lessons for Machine Learning, in: FAT* '19: Conference on Fairness, Accountability, and Transparency, FAT* '19, ACM Press, New York, NY, 2019, pp. 49-58. doi:10.1145/3287560.3287600.

[21] A. Liptak, Supreme Court Seems Ready to Throw Out Race-Based College Admissions, 2022, URL: https://www.nytimes.com/2022/10/31/us/supreme-court-harvard-unc-affirmative-action.html.

[22] D. Martens, Data Science Ethics: Concepts, Techniques, and Cautionary Tales, 1st. ed., Oxford University Press, New York, NY, 2022.

[23] C. McCrudden, A Comparative Taxonomy of 'Positive Action' and 'Affirmative Action' Policies, in: R. Schulze, Non-Discrimination in European Private Law, 1st. ed., Tübingen, Germany, 2011, pp. 157-180.

[24] B. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: Mapping the debate, Big Data & Society 3 (2016). doi:10.1177/2053951716679679.

[25] P. Kim, Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action, California Law Review 110 (2022) 1539-1596. doi: 10.15779/Z387P8TF1W.

[26] D. Pessach, E.Shmueli, A Review on Fairness in Machine Learning, ACM Computing Surveys 55 (2022). doi:10.1145/3494672.

[27] R. Primus, The Future of Disparate Impact, Michigan Law Review 108 (2010) 1341-1387. URL: https://repository.law.umich.edu/cgi/viewcontent.cgi?article=1517&context=articles.

[28] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. Rodolfa, R. Ghani, Aequitas: A Bias and Fairness Audit Toolkit, arXiv.1811.05577 (2019). doi:10.48550/arXiv.1811.05577.

[29] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, D. Sculley, No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, arXiv:1711.08536 (2017). doi:10.48550/arXiv.1711.08536.

[30] R. Siegel, Race-Conscious but Race-Neutral: The Constitutionality of Disparate Impact in the Roberts Court, Alabama Law Review 66 (2015) 653-689. URL: https://openyls.law.yale.edu/bitstream/handle/20.500.13051/4540/66AlaLRev653.pdf?sequence=2&isAllowed=y.

[31] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo, L. Floridi, The ethics of algorithms: key problems and solutions, AI & SOCIETY (2022) 215-230. doi:10.1007/s00146-021-01154-8.

[32] S. Wachter, B. Mittelstadt, C. Russel, Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI, Computer Law & Security Review 41 (2021). doi:10.1016/j.clsr.2021.105567.

[33] S. Verma, J. Rubin, Fairness Definitions Explained, in: IEEE/ACM International Workshop on Software Fairness, FairWare '18, ACM Press, New York, NY, 2018. doi:10.1145/3194770.3194776.

[34] S. Wachter, B. Mittelstadt, C. Russel, Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, West Virginia Law Review 123 (2021) 735-790. URL: https://researchrepository.wvu.edu/cgi/viewcontent.cgi?article=6331&context=wvlr.

[35] M. Whittaker, K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. Myers West, R. Richardson, J. Schultz, O. Schwartz, AI Now Report 2018, New York, NY, 2018. URL: https://ainowinstitute.org/AI_Now_2018_Report.pdf.