Fairness After Intervention: Towards a Theory of Substantial Fairness for Machine Learning

Sebastian Zezulka[1]

[1]Universität Tübingen, Cluster of Excellence "Machine Learning", Maria-von-Linden-Str. 6, 72076 Tübingen

Abstract

Implementing an algorithmically-informed policy represents a significant intervention into existing social structures. How such an intervention will affect society is a "naive", but arguably central, question for fair machine learning. I argue that this question is not adequately addressed by current "backward-looking" approaches, which focus on constraints in a historical, pre-interventional distribution. This paper makes two contributions. First, I specify two methodological challenges for answering the "naive" question, *intervention* and *feedback* effects, and suggest methods to address these challenges. Second, I introduce a detailed case study from public policy: statistical profiling of registered unemployed by public employment services, focusing especially on Germany. Thereby, I also answer the call for greater engagement in the algorithmic fairness literature with concrete and context-rich use cases.

Keywords

Fairness in Machine Learning, Statistical Profiling of Unemployed, Substantial Fairness

1. Retrospective and Prospective Fairness

The algorithmic fairness literature usually focuses on formal properties of an algorithm and its predictions [1], [2]. In one typical presentation, we are concerned with learning a function that takes as input some features X and a sensitive attribute A, and outputs a risk score R. Formal fairness requires that some constraint is met on either the joint distribution $p_0(A, X, Y, R)$ or on the causal structure D_0 giving rise to it (see the left-hand graph in Figure 1). Both group-based and causal fairness constraints can be represented in this way. Introducing a suitable similarity metric on (X, A) allows us to understand individual fairness approaches as imposing constraints on p_0 as well. In this sense, algorithmic fairness has a *retrospective* perspective that evaluates the fairness of an algorithm in the historical, pre-interventional distribution D_0 from which the training data were drawn.

This *retrospective* perspective on fair ML, focusing only on D_0, does not adequately address our "naive" question of how society will change once we implement a policy informed by our predictor. This is for three reasons. First, results from test-set drawn from p_0 are not valid fairness estimates because implementing an algorithmically informed policy constitutes an *intervention* and, therefore, changes the joint distribution of (A, X, Y, R) and the causal structure that gives rise to it [2]. Take college admission as example. Algorithmically informed admissions will differ from non-algorithmic ones—that is, after all, part of the motivation for introducing them—,

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
sebastian.zezulka@uni-tuebingen.de (S. Zezulka)

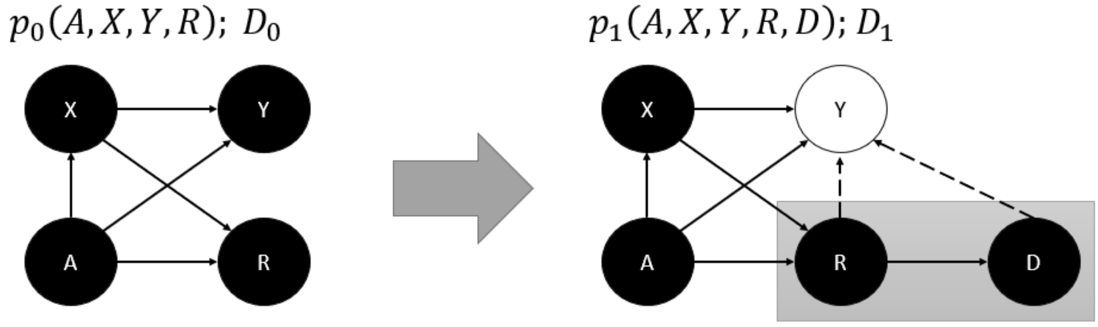

Figure 1: On the left is the joint distribution p_0 over sensitive attributes A, further features X, and the outcome variable Y generated from them. The predictions, say a risk score R, are the output of a learned function with arguments A and X. *Retrospective* fairness formulates constraints on this predictor. The right-hand side shows the joint distribution p_1 after implementing an algorithmically (informed) policy, with predictions R and decisions D both affecting the outcome variable. *Prospective* fairness requires to evaluate the consequences of intervening on the structure of D_0 and moving to D_1.

implying that some applicants that previously would not be admitted will be, and vice versa. The induced distribution shift implies that a predictor that satisfies some fairness constraint on D_0 will not necessarily satisfy it after the intervention [3]. Second, applicants will adapt to the new decision processes, a phenomenon commonly attributed to *Goodhart's law* and studied under the various guises of long-term fairness [4], [5], performative prediction [6] and strategic manipulation [7]. Third, formal fairness accounts tend to conflate predictions and decisions. Many standard case studies in fair ML encourage this equivocation. A function which predicts whether a student is likely to graduate in four years naturally suggests, but is not identical to, the admission policy which admits precisely those students that are classified as likely to graduate. Beigang [8] rightfully stresses the different normative requirements for prediction and decision making, respectively.

It is for these reasons that the *retrospective* perspective, focusing only on D_0, does not answer our original question: how will implementing an algorithmic policy impact society? and will the algorithmic policy ameliorate or entrench the injustices present in the historical structure D_0? To answer these difficult question and to move towards a substantial theory of fairness for machine learning, we must understand the problem as an intervention on structure [9]. In other words, an *prospective* perspective on algorithmic fairness must consider the post-interventional distribution $p_1(A, X, Y, R, D)$, shown on the right in Figure 1, with a new causal structure D_1 arising after the intervention. Here, predictions, R, and decisions, D, are conceptually separated. Further, the predictions and decisions affect the outcome variable, Y. Additionally, we must ask whether we have sufficiently accounted for (potential) feedback loops induced by the policy, or whether we should expect further changes to the joint distribution—is D_1 in a stable state, or should we expect it to continue to evolve?

Answering the above posed "naive" question of fair ML thus requires one to specify whether moving from a society represented by D_0 to one represented by D_1, with a algorithmic (informed) policy in place, is an improvement, or at least no deterioration in standards of justice. Doing so requires the introduction of some contextual holistic measure φ of the fairness of D_0 and D_1 and a comparison of $\varphi(D_0)$ with $\varphi(D_1)$. For example, φ could be the degree to which membership in the disadvantaged group predicts negative outcomes. Alternatively, it could be some suitable measure of the causal effect of the sensitive attribute on the outcome, or the degree to which relevantly similar individuals experience similar outcomes. Crucially, this proposal depends on comparing structural properties of D_0 and D_1, it cannot be adjudicated with knowledge of D_0 alone.

A number of challenging methodological questions arise from this conceptual clarification. It is natural to worry that the structure of D_1 is simply underdetermined by D_0 and even a detailed algorithmic policy proposal. This difficulty is real, but not insurmountable. Simulation studies using Markov Decision Processes [4] or structural causal models with dynamics [10] can be used to explore various scenarios and quantify the extent to which formal fairness constraints answer to our substantive fairness goals. If we are sincere in our concern about algorithmically informed policy, we should be willing to explore their consequences in the same way we study the potential consequences of high-stakes proposals in climate policy or public health. In this spirit, the following case study focuses on statistical profiling of registered unemployed.

2. Statistical profiling of long-term unemployed

The welfare systems in OECD countries have changed drastically in the last three decades. Public employment services (PES) have been transformed under a twofold *activation* regime. One is directed towards the unemployed by making participation in active labour market programs (ALMP) a pre-condition for receiving benefits. The other is directed towards public administration itself and aims at cost-effective provision of public goods by introducing organisational principles from private firms. By now, statistical profiling has been used to inform public administration decisions in a variety of fields. These tools are often framed as introducing objectivity and effectiveness in the provision of public goods. In their focus on statistical methods, they align with demands for evidence-based policy and digitisation in public administration.

Statistical profiling of registered unemployed is current practice in various OECD countries such as Australia, the Netherlands, and Belgium. Supervised learning techniques are employed to identify people at risk of becoming long-term unemployed (LTU). Building on studies like Kern et al. [11] and Körtner and Bonoli [12], I evaluate the prediction of the individual risk of becoming long-term unemployed using survey data from Germany as a case study on fairness in machine learning. I utilise the *IZA Evaluation Dataset Survey*[1] covering $8,915$ newly registered unemployed from Germany eligible for (type one) unemployment benefits. Using only survey data, I achieve accuracy rates between .66 and .808. These results are similar to those reported for statistical profiling tools used in practice [13].

This project has two parts, the first of which is already realised. First, a *retrospective* and

[1]This study uses the IZA ED Survey as provided by the International Data Service Center (IDSC) of the Institute for the Study of Labor (IZA). The IZA ED Survey consists of survey information on individuals who entered unemployment between June 2007 and May 2008 in Germany (see Arni, Caliendo, Künn, and Zimmermann, 2014).

group-based fairness analysis was conducted for the sensitive attributes *gender* and *migration background* and two exemplary allocation policies. In the first, the PES *prioritises* those predicted to be long-term unemployed. In the second, access to certain ALMPs is *restricted* to increase cost-effectiveness. The first policy is modelled after the example of Flanders, Belgium [14], the second after the proposed but so-far unrealised Austrian "AMS-Algorithm" [15]. Focusing on unconstrained logistic regressions as an example, the observational fairness measures *Independence* and *Separation* are violated for gender, whereas *Sufficiency* is almost satisfied. In other words, women are predicted to become LTU more often than the respective base rate suggests, and their false-positive rate is higher compared to men. For migration background, the three group-based formal fairness constraints are approximately satisfied. Based on this, I have further conducted a qualitative fairness evaluation discussing potential harms and benefits for registered unemployed under the two policies. Statistical profiling of registered unemployed is an intricate case study because the effects of ALMPs are heterogeneous across programs and social groups [16]. Strong welfare gains are possible if allocation to ALMPs can be made more targeted to individual needs, as shown by Goller et al. [17] and others. Utilising the empirical evidence from the social sciences, this case study provides the relevant context for a normative evaluation and demonstrates the relevance of the specific policies that are to be informed by statistical profiling.

The second part of this project will take a *prospective* view of the problem. Building on an initial simulation study by Scher et al. [18] and utilising individualised treatment effect estimates from Knaus et al. [19], this project aims at answering how to rigorously study the "naive" question of fair machine learning. Thus, it must answer how different ALMP allocation policies informed by fairness-constraint predictors impact labour market outcomes across demographics. Simulation studies are promising as they allow us to study equilibrium effects and quantify the effect of different *retrospective* fairness constraints after deployment.

Various *intervention* and *feedback* effects are to be considered here. Different combinations of (1) (fairness-constrained) risk predictors and (2) algorithmically-informed allocation policies will induce different distributions of labor market outcomes. The heterogeneity in program effects implies further variation in labour-market outcomes under different choices for (1) and (2). Understanding the effects of these various combinations is essential for crafting policies that make gains in substantive fairness.

To summarise, the paper contributes a novel conceptual understanding of the methodological requirements for substantial fairness in machine learning, *fairness after intervention*, and illustrates the proposal by a case study of statistical profiling of registered unemployed.

Acknowledgments

The author is supported by the Deutsche Forschungsgemeinschaft (BE5601/4-1; Cluster of Excellence "Machine Learning—New Perspectives for Science", EXC 2064, project number 390727645) and the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

[1] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019.

[2] R. Berk, A. K. Kuchibhotla, E. T. Tchetgen, Fair Risk Algorithms, Annual Review of Statistics and Its Application 10 (2022).

[3] A. Mishler, N. Dalmasso, Fair When Trained, Unfair When Deployed: Observable Fairness Measures are Unstable in Performative Prediction Settings, 2022.

[4] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, Y. Halpern, Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, ACM, 2020, p. 525–534.

[5] X. Zhang, R. Tu, Y. Liu, M. Liu, H. Kjellstrom, K. Zhang, C. Zhang, How do fair decisions fare in long-term qualification?, Advances in Neural Information Processing Systems (2020) 1–13.

[6] J. Perdomo, T. Zrnic, C. Mendler-Dünner, M. Hardt, Performative prediction, in: International Conference on Machine Learning, PMLR, 2020, pp. 7599–7609.

[7] L. Hu, N. Immorlica, J. W. Vaughan, The Disparate Effects of Strategic Manipulation, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 259–268.

[8] F. Beigang, On the Advantages of Distinguishing Between Predictive and Allocative Fairness in Algorithmic Decision-Making, Minds and Machines 32 (2022) 655–682.

[9] D. Malinsky, Intervening on structure, Synthese 195 (2018) 2295–2312.

[10] Y. Hu, L. Zhang, Achieving Long-Term Fairness in Sequential Decision Making, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 9549–9557.

[11] C. Kern, R. L. Bach, H. Mautner, F. Kreuter, Fairness in algorithmic profiling: A German case study, arXiv preprint arXiv:2108.04134 (2021).

[12] J. Körtner, G. Bonoli, Predictive Algorithms in the Delivery of Public Employment Services, SocArXiv (2021).

[13] S. Desiere, K. Langenbucher, L. Struyven, Statistical profiling in public employment services: An international comparison, Employment and Migration Working Papers 224 (2019).

[14] S. Desiere, L. Struyven, Using artificial intelligence to classify jobseekers: the accuracy-equity trade-off, Journal of Social Policy 50 (2021) 367–385.

[15] D. Allhutter, F. Cech, F. Fischer, G. Grill, A. Mager, Algorithmic profiling of job seekers in Austria: how austerity politics are made effective, Frontiers in Big Data 3 (2020) 1–17.

[16] D. Card, J. Kluve, A. Weber, What works? A meta analysis of recent active labor market program evaluations, Journal of the European Economic Association 16 (2018) 894–931.

[17] D. Goller, T. Harrer, M. Lechner, J. Wolff, Active labour market policies for the long-term unemployed: New evidence from causal machine learning, arXiv preprint arXiv:2106.10141 (2021).

[18] S. Scher, S. Kopeinik, A. Trügler, D. Kowald, Modelling the long-term fairness dynamics of data-driven targeted help on job seekers, Scientific Reports 13 (2023).

[19] M. C. Knaus, M. Lechner, A. Strittmatter, Heterogeneous employment effects of job search programs a machine learning approach, Journal of Human Resources 57 (2022) 597–636.