Introduction to The Special Section on Bias and Fairness in AI

Toon Calders, Eirini Ntoutsi, Mykola Pechenizkiy, Bodo Rosenhahn, Salvatore Ruggieri
U. Antwerpen, Freie Universität Berlin, TU Eindhoven, Leibniz University Hannover, Università di Pisa
toon.calders@uantwerpen.be, eirini.ntoutsi@fu-berlin.de, m.pechenizkiy@tue.nl, rosenhahn@tnt.uni-hannover.de, salvatore.ruggieri@unipi.it

ABSTRACT

Fairness in Artificial Intelligence rightfully receives a lot of attention these days. Many life-impacting decisions are being partially automated, including health-care resource planning decisions, insurance and credit risk predictions, recidivism predictions, etc. Much of work appearing on this topic within the Data Mining, Machine Learning and Artificial Intelligence community is focused on technological aspects. Nevertheless, fairness is much wider than this as it lies at the intersection of philosophy, ethics, legislation, and practical perspectives. Therefore, to fill this gap and bring together scholars of these disciplines working on fairness, the first workshop on Bias and Fairness in AI was held online on September 18, 2020 at the ECML-PKDD 2020 conference. This special section includes six articles presenting different perspectives on bias and fairness from different angles.

Keywords

Bias, fairness, discrimination, fairness-aware machine learning, responsible artificial intelligence

1. INTRODUCTION

Artificial Intelligence (AI) techniques based on big data and algorithmic processing are increasingly used to guide decisions in important societal spheres, including hiring decisions, university admissions, loan granting, and crime prediction. However, there are growing concerns with regard to the epistemic and normative quality of AI evaluations and predictions. In particular, there is strong evidence that algorithms may sometimes amplify rather than eliminate existing bias and discrimination, and thereby have negative effects on social cohesion and on democratic institutions.

Despite the increased amount of work in this area in the last few years, we still lack a comprehensive understanding of how pertinent concepts of bias or discrimination should be interpreted in the context of AI and which socio-technical options to combat bias and discrimination are both realistically possible and normatively justified. The main objective of the workshop on Bias and Fairness in AI held online[1] on September 18, 2020 at the ECML-PKDD 2020 conference is a contribution to the understanding of "How can standards of unbiased attitudes and non-discriminatory practices be met in (big) data analysis, AI and algorithm-based decision-making?".

We introduce topics in Bias and Fairness in AI and describe how they were covered in the program of the workshop in Section 2 and provide a brief overview of the contributed articles to this special section in Section 3.

2. TOPICS IN AI BIAS AND FAIRNESS

Research on fairness in machine learning and data mining took off in 2008-2010 with some of the first works on discrimination discovery in databases [1] and learning classification models with (non-discrimination) independency constraints [2; 3]. These papers were followed by an exponential explosion of papers in major AI conferences, and an emergence of new cross-disciplinary workshops and conferences such as most notably FAccT[2] and AIES[3]. A recent snapshot of the frontiers of fairness in machine learning research can be found in [4].

Much of the research on fairness in machine learning can be framed in an optimization context [5], where the goal is to maintain good predictive performance while satisfying a number of group-level or individual fairness constraints. This combination can be achieved via modeling and removing representation bias and/or labeling bias in the training data, via fairness-aware representation learning [6; 7], model induction, model selection, regularization, or post-processing of specific [8] or any [9] trained models or model outputs.

In parallel, temporal dynamics of fairness in algorithmic decision making [10] and its long-term impact [11] has been studied to address feedback loops that may amplify discrimination.

Next to algorithmic approaches, also progress has been made with respect to theoretical analysis to better understand the possibility or impossibility of fairness with its different often conflicting notions [12].

Another recent avenue of fairness-aware machine learning research includes causality. The notion of counterfactual fairness and approaches of counterfactual inference have been proposed to make predictions fair across different subpopulations. Considering classification as an optimization problem with fairness constraints entailed by competing causal explanations, Russell et al. [13] demonstrated that it is possible to be approximately fair with respect to multiple pos-

[1] https://sites.google.com/view/bias-2020/programme

[2] https://facctconference.org/

[3] https://www.aies-conference.com/

sible causal models at once, thus mitigating the bottleneck of exact causal specification.

The BIAS2020 workshop solicited contributions on bias and fairness in all areas of AI (supervised and unsupervised learning, reinforcement learning, information retrieval and recommender systems, human-computer interaction, constraint solving, complex systems and networks, etc.) and encouraging interdisciplinary studies including law, philosophy and social sciences. 21 full paper submissions were received of which 7 were selected to the workshop program after peer-review. The program also featured four invited talks and a concluding panel discussion. Revised and extended contributions were invited for this special section.

3. CONTRIBUTED ARTICLES

The special section includes six contributed articles spanning a variety of topics: philosophical viewpoints on discrimination [14], applicability of different ML fairness notions [15], a new measure for viewpoint fairness in ranking applications [16], gender perception in online platforms [17], fair classification via ethical adversaries [18], and why not only serendipity but also equity should be considered to mitigate historical discrimination effects [19].

Two Kinds of Discrimination in AI-Based Penal Decision-Making. Hubner in [14] presents a viewpoint on discrimination in algorithmic decision making from the standpoint of practical philosophy and ethics of science. In his work, he distinguishes two kinds of discrimination that need to be addressed in AI-based penal decision-making: the problem of inevitable trade-offs between incompatibility of statistical fairness measures as became widely known due to the COMPAS study and analyzed theoretically in [20], and the problem referred to as the so-called *discursive fairness* that applies when *humans make decisions based on empirical evidence*. Hubner discusses the fundamental differences in approaching requirements of non-discriminatory action within the penal sector for each of these two kinds of discrimination. Whereas in the case of statistical fairness, the focus is on measuring dependency between race and (correctly and/or wrongly) predicted recidivism, in case of discursive fairness, it is necessary to analyze what types of information must be provided when justifying a court's decisions based on a machine learning model's predictions. This leads to seeking answers to the core question: *What reasons must a judge as a human decision maker provide for* ***her*** *each and every decision to grant or deny parole.*

On the Applicability of Machine Learning Fairness Notions. While many notions of fairness were introduced and many machine learning approaches and techniques have been development that can help to optimize for those notions, we also know that it is impossible to optimizing several of the competing notions of fairness at the same time [12]. Hence, a natural practitioner's question is which notion of fairness should be used. Makhlouf et al. [15] introduce a survey of fairness notions that should help find an answer to the question "which notion of fairness is most suited to a given real-world scenario and why?". The authors identify a set of fairness-related characteristics of real-world scenarios and analyze the relevance of corresponding fairness notions to these characteristics. Their findings are summarized in a decision diagram that may help different research communities, practitioners and policy makers to understand and navigate the space of fairness notions studied in fairness-aware machine learning.

Blind Spots in AI: the Role of Serendipity and Equity in Algorithm-Based Decision-Making. Van Leeuwen et al. [19] argue that designing an algorithm-based decision-making system focusing solely on serendipity might not be enough to avoid historical discrimination and therefore they suggest to also include equity in the development process. To this end, they propose a design rationale that incorporates the principles of serendipity (diversifiability) and equity (intersectionality, reflexivity and power balance) for the development of such systems.

Gendering algorithms in social media. Fosch-Villaronga et al. [17] investigate the impact of algorithmic bias on inadvertent privacy violations and the reinforcement of social prejudices of gender and sexuality. In particular, they conducted an online survey to understand whether and how Twitter inferred the gender of users. They found that gender-related stereotypes persist both online and offline, and platforms often appear to fail to understand that gender is not binary (male/female). Beyond Twitter's binary understanding of gender and the inevitability of the gender inference as part of Twitter's personalization trade-off, they also found that the misgendering rate is much higher for gay men (32%) and straight women (16%) as compared to straight males (8%). Their results call for attention to gender in gender classifiers to avoid amplification of existing biases that affect especially marginalized communities.

Ethical Adversaries: Towards Mitigating Unfairness with Adversarial Machine Learning. It is now common in machine learning research to address non-discrimination by introducing independency constraints into the predictive modeling process. One generic approach to do this was presented in [5]. Delobelle et al. [18] continue on this track and introduce the idea of using adversarial training for improving fairness of classification. The authors introduce a framework that makes use of two models. One model is optimized for preventing the correct guessing of the values of protected attributes, while staying as accurate as possible. The other adversary model leverages evasion attacks to generate new examples that will be misclassified and provides them to the training of the first model. The experimental evaluation of this framework on common benchmarks like the COMPAS datasets demonstrates promising results for achieving group level fairness including demographic parity and equality of opportunity.

Assessing Viewpoint Diversity in Search Results Using Ranking Fairness Metrics. Fairness-awareness is being considered in a variety of applications of autonomous decision making by machine learning based scoring mechanisms. Considering biases and fairness in recommender systems and web search, a graph-based algorithm that post-processes generated recommendations for improving aggregate diversity was proposed in [21]. The paper of Draws et al. [16] included in the special section, highlights the importance of researching how to measure and assess *viewpoint diversity* in real search result rankings. Depending on how the items are ranked in search results, more homogeneous or more diverse items or viewpoints will be exposed to the user. The authors show that assessing the viewpoint diversity might not be as straightforward as it may seem, considering and experimenting with a few ranking fairness metrics in a controlled simulation study.

We hope you will enjoy reading the papers on bias and fairness in AI in this special section and find them an inspiration for formulating and addressing many of the open challenges in this socio-technical problem space, advancing the current state of the art further and further.

Acknowledgements

Our special thanks go to the invited speakers, all authors who submitted to and presented their work at the BIAS2020 workshop, to the program committee members and ad hoc reviewers, and to all the participants. The workshop was part of the FAccT network `https://facctconference.org/network/`. The work of E. Ntoutsi and S. Ruggieri was partially supported by the European Community H2020 project *NoBIAS* (`nobias.eu`, G.A. 860630).

4. REFERENCES

[1] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08*, page 560–568, New York, NY, USA, 2008. ACM.

[2] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops, ICDMW '09*, page 13–18, USA, 2009. IEEE Computer Society.

[3] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21(2):277–292, 2010.

[4] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, April 2020.

[5] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

[6] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 325–333. PMLR, 2013.

[7] Tongxin Hu, Vasileios Iosifidis, Wentong Liao, Hang Zhang, Michael Ying Yang, Eirini Ntoutsi, and Bodo Rosenhahn. FairNN - conjoint learning of fair representations for fair decisions. In *Proceedings of the 23rd International Conference on Discovery Science, DS 2020*, volume 12323 of *LNCS*, pages 581–595. Springer, 2020.

[8] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *Proceedings of the 10th IEEE International Conference on Data Mining, ICDM 2010*, pages 869–874. IEEE Computer Society, 2010.

[9] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems, NIPS 2016*, pages 3315–3323, 2016.

[10] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni 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*, page 525–534, New York, NY, USA, 2020. ACM.

[11] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 3156–3164, 2018.

[12] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143, March 2021.

[13] Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems, NIPS 2017*, pages 6414–6423, 2017.

[14] Dietmar Hübner. Two kinds of discrimination in AI-based penal decision-making. *SIGKDD Explorations*, 23(1), 2021.

[15] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. *SIGKDD Explorations*, 23(1), 2021.

[16] Tim Draws, Nava Tintarev, and Ujwal Gadiraju. Assessing viewpoint diversity in search results using ranking fairness metrics. *SIGKDD Explorations*, 23(1), 2021.

[17] Eduard Fosch-Villaronga, Adam Poulsen, Roger A. Søraa, and Bart Custers. Gendering algorithms in social media. *SIGKDD Explorations*, 23(1), 2021.

[18] Pieter Delobelle, Paul Temple, and Bettina Berendt. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *SIGKDD Explorations*, 23(1), 2021.

[19] Cora van Leeuwen, Annelien Smets, and An Jacobs. Blind spots in ai: the role of serendipity and equity in algorithm-based decision-making. *SIGKDD Explorations*, 23(1), 2021.

[20] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science*, page 43:1–43:23, 2017.

[21] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Fairmatch: A graph-based approach for improving aggregate diversity in recommender systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '20*, page 154–162, New York, NY, USA, 2020. ACM.