JMLR: Workshop and Conference Proceedings 1:1–5, 2021 Algorithmic Fairness through the Lens of Causality and Robustness

Algorithmic Fairness through the Lens of Causality and Robustness (AFCR) 2021

Jessica Schrouff SCHROUFF@GOOGLE.COM
Google Research

Awa Dieng AWADIENG@GOOGLE.COM
Google Research, MILA

Miriam Rateike MIRIAM.RATEIKE@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems

Kweku Kwegyir-Aggrey MARK_KWEGYIR-AGGREY@BROWN.EDU
Brown University

Golnoosh Farnadi FARNADIG@MILA.QUEBEC
HEC Montréal, MILA

Editor: Jessica Schrouff, Awa Dieng, Miriam Rateike, Kweku Kwegyir-Aggrey, Golnoosh Farnadi

1. Introduction

Trustworthy machine learning (ML) encompasses multiple fields of research, including (but not limited to) robustness, algorithmic fairness, interpretability and privacy. Recently, relationships between techniques and metrics used across different fields of trustworthy ML have emerged, leading to interesting work at the intersection of algorithmic fairness, robustness, and causality.

On one hand, causality has been proposed as a powerful tool to address the limitations of initial statistical definitions of fairness (Kusner et al., 2017; Chiappa, 2019; Khademi et al., 2019; Wu et al., 2019). However, questions have emerged regarding 1) the applicability of such approaches due to strong assumptions inherent to causal questions (Kilbertus et al., 2019) and 2) the suitability of a causal framing for studies of bias and discrimination (Kohler-Hausmann, 2019; Hu and Kohler-Hausmann, 2020; Kasirzadeh and Smart, 2021).

On the other hand, the robustness literature has surfaced promising approaches to improve fairness in ML models. For instance, parallels can be shown between individual fairness and local robustness guarantees (Yurochkin et al., 2019; Nanda et al., 2021; Xu et al., 2021; Yeom and Fredrikson, 2020) or between group fairness metrics and robustness to distribution shift (Veitch et al., 2021). Beyond similarities, the interactions between fairness and robustness can help us understand how fairness guarantees hold under distribution shift (Singh et al., 2021; Subbaswamy and Saria, 2020) or adversarial/poisoning attacks (Solans et al., 2020; Liu et al., 2021), leading to *fair and robust* ML models.

To encourage further work at the intersection of these fields, we organized The *Algorithmic Fairness through the Lens of Causality and Robustness* workshop (AFCR*) as part of

*https://www.afciworkshop.org/

the Neural Information Processing Systems (NeurIPS[†]) conference in December 2021. Our aim was to investigate how these different topics relate, but also how they can *augment* each other to provide better or more suited definitions and mitigation strategies for algorithmic fairness. Examples of questions we were interested in addressing at the workshop include:

- *How can causally grounded fairness methods help develop more robust fairness algorithms in practice?*
- *What is an appropriate causal framing in studies of discrimination?*
- *How do approaches for adversarial/poisoning attacks target algorithmic fairness?*
- *How do fairness guarantees hold under distribution shift?*

2. Workshop

The AFCR workshop was held as a NeurIPS workshop on December 13th, 2021. In accordance with the virtual format of the conference, the program consisted in a mix of pre-recorded and live events.

2.1. Program

AFCR 2021 featured invited talks by Elias Bareinboim (Columbia University), Rumi Chunara (New York University), Silvia Chiappa (DeepMind), Isabel Valera (Saarland University), Aditi Raghunathan (UC Berkeley) and Hima Lakkaraju (Harvard University), six spotlight talks from authors of papers accepted at the venue, a panel discussion with Been Kim (Google Research), Ricardo Silva (University College London), Solon Barocas (Microsoft Research) and Rich Zemel (University of Toronto), two poster sessions and roundtable discussions. The latter consisted in live discussions between invited researchers of mixed seniority and workshop attendees, held virtually. They engaged more than 50 researchers and covered the following themes:

- *Causality for fairness.* Invited researchers: Issa Kohler-Haussman (Yale University), Matt Kusner (University College London), Maggie Makar (University of Michigan) and Ioana Bica (University of Oxford).
- *Robustness for fairness.* Invited researchers: Silvia Chiappa (DeepMind), Alexander D'Amour (Google Research) and Elliot Creager (University of Toronto).
- *General fairness.* Invited researchers: Isabel Valera (Saarland University), Ulrich Aïvodji (ETS Montréal), Keziah Naggita (Toyota Technological Institute at Chicago) and Stephen Pfohl (Stanford University).
- *Ethics.* Invited researchers: Luke Stark (University of Western Ontario), Irene Chen (Massachusetts Institute of Technology) and Lizzie Kumar (Brown University).

[†] https://neurips.cc/

2.2. Contributed papers

AFCR received 25 viable submissions, which were sent for peer reviewing. All papers received at least 3 reviews, which led to the acceptance of 16 works (acceptance rate: 64%). Among them, 5 papers were related to the use of causal methods for fairness, 4 works discussed the intersection of fairness and robustness, and 7 described applications, mitigation techniques or metrics for fairness. Among the selected works, 8 papers were considered for inclusion in the Proceedings, with the authors of 4 works choosing to do so. All contributed works were presented as posters during the conference, and were included in the live stream through pre-recorded 3 minutes video summaries. A 1-page abstract submission was also implemented on a rolling deadline, leading to 1 additional poster presented at the conference (out of 3 submissions).

3. Themes and open questions

Among the common themes during the workshop, the attendees discussed how sensitive attributes represent social constructs and the difficulties related to obtaining good proxies to assess the impact of machine learning on different subgroups of the population. In addition, the attendees questioned how the field can move beyond fairness metrics and audits, towards societal interventions that would lead to fair and/or equitable outcomes. The attendees also discussed at length how more effort needs to happen before we can see a practical impact of causal methods. Finally, they discussed different definitions of 'robustness' and how evaluation and mitigation techniques with regards to robustness depended on the data that was available. We leave with a set of open questions, that we hope to address in future editions:

- How should we model sensitive attributes?
- How can the ML community contribute to societal remedies of unfairness?
- How can we bring causal advances with its assumptions to practice?
- How can we ensure reliable models, decisions, explanations?

4. Acknowledgments

As AFCR organizers, we would like to thank all the invited speakers, panelists and roundtable researchers, as well as all the authors who contributed materials to the workshop. In addition, 46 reviewers contributed their time to ensure the quality of the workshop content. In alphabetical order, they are: Aboli Marathe (Pune Institute of Computer Technology, Symbiosis Centre for Applied Artificial Intelligence(SCAAI)), Akshayvarun Subramanya (UMBC), Alexander D'Amour (Google Brain), Amin Nikanjam (Université de Montréal), Anshuman Chhabra (UC Davis), Aria Khademi (Ford Motor Company), Arun Raja (Institute for Infocomm Research, A*STAR), Babak Salimi (Unievristy of California at San Diego), Candice Schumann (Google), Carolyn Ashurst (Alan Turing Institute), David Watson (University College London), Eli Sherman (Johns Hopkins University), Elizabeth Kumar (Brown University), Elliot Creager (University of Toronto), Flavien Prost (Google),

Grace Abuhamad (ServiceNow), Hadis Anahideh (University of Illinois at Chicago), Ibrahim Alabdulmohsin (Google Research), Jeremiah Liu (Harvard University), Jessica Dai (Arthur AI), Julius Adebayo (MIT), Julius von Kügelgen (MPI for Intelligent Systems, Tübingen, University of Cambridge), Kai-Wei Chang (UCLA), Krikamol Muandet (Max Planck Institute for Intelligent Systems), Krystal Maughan (ServiceNow), Laurent Charlin (HEC Montreal and Mila), Logan Stapleton (University of Minnesota), Maggie Makar (MIT), Malik Altakrori (McGill University), Megha Srivastava (Stanford), Minyechil tefera (symbiosis international university), Miruna Clinciu (Edinburgh Centre for Robotics), Negar Rostamzadeh (Google), Nenad Tomasev (DeepMind), Ricardo Silva (University College London), Roy Adams (Johns Hopkins University), Sahil Verma (University of Washington), sainyam galhotra (UMass Amherst), Samuel Dooley (University of Maryland), Sanghamitra Dutta (JP Morgan), Sarah Brown (Brown University), Sarah Tan (Facebook), Shubham Singh (University of Illinois at Chicago), Soumye Singhal (Mila, University of Montreal), Sunhee Hwang (LG Uplus) and Vedant Nanda (University of Maryland, College Park).

Importantly, we would like to thank the NeurIPS workshop Chairs Ndapa Nakashole, Anna Goldenberg, Sanmi Koyejo and Tristan Naumann, as well as Brian Nettleton, Terri Auricchio, Anastasia Kozitsyna and Viet Lai for their technical support. Finally, we thank Deborah Dormah Kanubala (Saarland University), Maryam Molamohammadi (MILA), Nisha George (Saarland University) and Priyanka Upadhyay (Saarland University) for their volunteer work during the workshop.

References

Silvia Chiappa. Path-specific counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7801–7808, Jul. 2019. doi: 10.1609/aaai.v33i01.33017801. URL https://ojs.aaai.org/index.php/AAAI/article/view/4777.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513, 2020.

Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. FAccT '21, page 228–236, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445886. URL https://doi.org/10.1145/3442188.3445886.

Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, 2019.

N. Kilbertus, P. J. Ball, M. J. Kusner, A. Weller, and R. Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, page 213. AUAI Press, July 2019. URL http://auai.org/uai2019/proceedings/papers/213.pdf.

Issa Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. 113 Nw. U. L. Rev. 1163,

2019. URL https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1374&context=nulr.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, Bo Li, and Yishi Lin. Stable adversarial learning under distributional shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8662–8670, 2021.

Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, 2021.

Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.

David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. *arXiv preprint arXiv:2004.07401*, 2020.

Adarsh Subbaswamy and Suchi Saria. I-spec: An end-to-end framework for learning transportable, shift-stable models. *arXiv preprint arXiv:2002.08948*, 2020.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/44a2e0804995faf8d2e3b084a1e2db1d-Paper.pdf.

Han Xu, Xiaorui Liu, Yaxin Li, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training, 2021. URL https://openreview.net/forum?id=vOchfRdvPy7.

Samuel Yeom and Matt Fredrikson. Individual fairness revisited: Transferring techniques from adversarial robustness. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 437–443. International Joint Conferences on Artificial Intelligence Organization, 7 2020. URL https://doi.org/10.24963/ijcai.2020/61. Main track.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*, 2019.