Algorithmic Fairness through the Lens of Causality and Privacy (AFCP) 2022

Awa Dieng AWADIENG@GOOGLE.COM
Google Research, Brain team

Miriam Rateike RATEIKE@CS.UNI-SAARLAND.DE
Saarland University

Golnoosh Farnadi FARNADIG@MILA.QUEBEC
HEC Montreal, Mila

Ferdinando Fioretto FFIORETT@SYR.EDU
Syracuse University

Matt Kusner MATT.KUSNER@GMAIL.COM
University College London

Jessica Schrouff SCHROUFF@DEEPMIND.COM
DeepMind

Editors: Awa Dieng, Miriam Rateike, Golnoosh Farnadi, Ferdinando Fioretto, Matt Kusner, Jessica Schrouff

1. Introduction

As machine learning models permeate every aspect of decision making systems in consequential areas such as healthcare and criminal justice, it has become critical for these models to satisfy trustworthiness desiderata such as fairness, interpretability, accountability, privacy and security. Initially studied in isolation, recent work has emerged at the intersection of these different fields of research, leading to interesting questions on how fairness can be achieved using a causal perspective and under privacy concerns.

Indeed, the field of causal fairness has seen a large expansion in recent years (Chiappa (2019); Khademi et al. (2019); Kusner et al. (2017); Wu et al. (2019)) notably as a way to counteract the limitations of initial statistical definitions of fairness (Friedler et al. (2016); Kleinberg et al. (2017); Lipton et al. (2018); Liu et al. (2018)). While a causal framing provides flexibility in modelling and mitigating sources of bias using a causal model, proposed approaches rely heavily on assumptions about the data generating process, i.e., the faithfullness and ignorability assumptions. This leads to open discussions on (1) how to fully characterize causal definitions of fairness, (2) how, if possible, to improve the applicability of such definitions, and (3) what constitutes a suitable causal framing of bias from a sociotechnical perspective? (Carey and Wu (2022); Fawkes et al. (2022); Kohler-Hausmann (2019); Kasirzadeh and Smart (2021); Kilbertus et al. (2019)).

Additionally, while most existing work on causal fairness assumes observed sensitive attribute data, such information is likely to be unavailable due to, for example, data privacy laws or ethical considerations. This observation has motivated initial work on training and evaluating fair algorithms without access to sensitive information (Andrus et al., 2021;

Kasirzadeh and Clifford, 2021; Kilbertus et al., 2018; Mougan et al., 2022; Yan et al., 2020) and studying the compatibility and trade-offs between fairness and privacy (Chang and Shokri, 2021; Cheng et al., 2021; Cummings et al., 2019; Ekstrand et al., 2018; Fioretto et al., 2022; Jagielski et al., 2019). However, such work has been limited, for the most part, to statistical definitions of fairness raising the question of whether these methods can be extended to causal definitions.

Given the interesting questions that emerge at the intersection of these different fields, we organized the Algorithmic Fairness through the Lens of Causality and Privacy workshop (AFCP[1]) as part of the Neural Information Processing Systems (NeurIPS[2]) conference in December 2022. Our aim was to deeply investigate how algorithmic fairness, causality and privacy relate, but also how they can *augment* each other to provide better or more suited definitions and mitigation strategies for algorithmic fairness. We were particularly interested in addressing open questions in the field, such as:

- Are causal definitions of fairness compatible with privacy constraints? If not, what are the trade-offs?
- How to build fair models without direct access to (or with encrypted) sensitive information?
- What causal assumptions hold in a fairness context?
- What are the ethical concerns and moral assumptions underlying causal-based notions of fairness?
- How can causality help in achieving intersectional fairness?

2. Workshop

The AFCP workshop was held in-person as a NeurIPS workshop on December 03, 2022. In order to make the workshop accessible to as many people as possible and accommodate different time-zones, we held a virtual morning session with livestreamed invited talks and roundtables. Additionally, all in-person talks were livestreamed and all accepted papers were able to pre-recorded a 3-minute video available on the website.

2.1. Program

AFCP 2022 featured invited talks by Deirdre Mulligan (UC Berkeley), Razieh Nabi (Emory University), Nicolas Papernot (University of Toronto), and Catuscia Palamidessi (INRIA), six spotlight talks from authors of accepted papers, an interdisciplinary panel discussion with Kristian Lum (University of Chicago), Joshua Loftus (London School of Economics), Rachel Cummings (Columbia University), Jake Goldenfein (Melbourne Law School), Sara Hooker (Cohere For AI), one poster session and roundtable discussions. The latter consisted in live discussions between invited researchers of mixed seniority and workshop attendees,

1. https://www.afciworkshop.org
2. https://neurips.cc

held virtually and in-person. They engaged more than 100 researchers and covered the following themes:

- Causality and fairness. Invited researchers: Joshua Loftus (London School of Economics), Dhanya Sridhar (University of Montreal, Mila), Aida Rahmattalabi (SonyAI), David Madras (University of Toronto), and Amanda Coston (CMU).
- Privacy and fairness. Invited researchers: Rachel Cummings (Columbia University), Ulrich Aïvodji (ETS Montreal), Fatemehsadat Mireshghallah (UCSD), and Sikha Pentyala (UW Tacoma).
- Ethics and fairness. Invited researchers: Negar Rostamzadeh (Google Research), Sina Fazelpour (Northeastern University), and Nyalleng Moroosi (Google Research).
- Interpretability and fairness. Invited researchers: Zachary Lipton (CMU), Julius Adebayo (MIT), and Amir-Hossein Karimi (MPI-IS, ETH Zurich).

2.2. Contributed papers and extended abstracts

AFCP had two tracks: a *Paper* track which called for 4-9 page manuscripts of novel work and an *Extended abstract* track which called for 1-page abstracts. We received 36 viable papers submissions and 15 extended abstracts, which were sent for peer reviewing. All submissions received at least 3 reviews, which led to the acceptance of 23 papers (acceptance rate: 64%) and 11 abstracts (acceptance rate: 68%). Among the accepted papers, 6 papers were related to the use of causal methods for fairness, 10 works discussed the intersection of fairness and privacy, and 7 described applications, mitigation techniques or metrics for fairness. Among the selected works, 14 papers were considered for inclusion in the Proceedings, with the authors of 4 works choosing to do so. All accepted works were presented as posters during the conference, and contributions in the *Paper* track were able to pre-recorded 3 minute video summaries which were available on the virtual NeurIPS website.

3. Themes and open questions

Throughtout the workshop, salient discussion topics were around how to conceptualize interventions on immutable sensitive attributes and the issues with modeling sensitive/immutable traits as (exogenous) causes. Participants also discussed the identifiability of counterfactual notions of fairness and how to address conflicting stakeholder views on causal graphs. The discussion around privacy centered on how differential privacy can negatively affect under-represented subgroups and what role synthetic datasets play in privacy-preserving fair analysis. Finally, several issues around data were brought up including dataset construction, data protection rules in face recognition applications, intellectual property and open datasets with harmful content.

Below, we highlight some takeaways from the discussion and open questions we hope to address in future editions:

- Causality and fairness: Domain Knowledge is important when modeling causal relationships. It is important to work with stakeholders to verify the plausibility of the causal

graph considered in context. *Causality inherently relies on manipulation. Given the impossibility of intervening on immutable characters, what can causality really bring to the work on fairness? Should we hold ourselves to the idea of causality through manipulation only?*

- Privacy and fairness: It is important to providing contextualized privacy explanations (e.g., value of epsilon in different applications domains) and to study application-specific problems. *Open challenges in differential privacy include streaming, allocating privacy budget, data heterogeneity.*
- Interpretability and fairness: Counterfactual explanations can provide recommendations with regards to causal explanation and can be restricted based on real world assumptions. However they can increase security risks. It is important to consider real cases, contexts, and the end-users for effective explanations. *While LIME/SHAP and other techniques are valuable, how can we move to a post LIME/SHAP world with more interactive model correction and feedback from domain experts?*
- Ethics and fairness: Participatory machine learning approaches are important in order to co-develop and allow different contexts to be taken into account (domain, politics, regulation, etc). When teaching ethics to machine learning and statistic students, social and historical perspectives play an important role in order to better understand data and ethics. *Given the field's focus on different fairness tradeoffs and impossibility results, what is an ethical understanding of these findings?*

4. Acknowledgments

As AFCP 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, 65 reviewers and 9 meta-reviewers contributed their time to ensure the quality of the workshop content. In alphabetical order, the meta-reviewers are: Alexander D'Amour, Aurélien Bellet, Babak Salimi, Graham Cormode, Michael Hay, Milan Shen, Sarah Brown, Sarah Tan, Xi He. The reviewers are: Adrián Arnaiz-Rodríguez, Agoritsa Polyzou, Aishwarya Sivaraman, Alan Mishler Amin Nikanjam, Amrita Roy Chowdhury, Anderson Nascimento, Andrii Krutsylo, Anoop Mishra, Anshuman Chhabra, Benedikt Höltgen, Bhagyashree Puranik, Canyu Chen, Carsten Baum, Cuong Tran, Daniela Cialfi, David Watson, Debashis Ghosh, Di Wang, Elette Boyle, Elliot Creager, Faisal Hamman, Fatemeh Mireshghallah, Flavien Prost, Ganesh Kini, Gintare Dziugaite, Hanieh Hashemi, I. Elizabeth Kumar, Ibrahim Alabdulmohsin, Ivan Habernal, Ivan Habernal, Jake Fawkes, Jan Ramon, Jeremiah Liu, Jessica Dai, Jianfeng Chi, Jonas Ngnawe, Jonathan Passerat-Palmbach, Jose Alvarez, Julien Ferry, Julius Adebayo, Krystal Maughan, Kweku Kwegyir-Aggrey, Malik Altakrori, Manish Nagireddy, Marc Juarez, Marianne Abemgnigni Njifon, Martin Jullum, Masoud Mansoury, Megha Srivastava, Mercy Asiedu, Olawale Salaudeen, Paritosh Ramanan, Rakshit Naidu, Ranya Aloufi, Raouf Kerkouche, Rob Romijnders, Sahil Verma, Sanghamitra Dutta, Satyapriya Krishna, Shubham Singh, Sihong He, Srinath Namburi, Stacey Truex, Vishal Bhalla.

Importantly, we would like to thank the NeurIPS workshop Chairs Hanie Sedghi, Hsuan-Tien Lin, Sungjin Ahn, Tristan Naumann, as well as Brad Brockmeyer, Brian Nettleton,

Lee Campbell, Mary Ellen Perry, Max A Wiesner , Stephanie Willes, and Terri Auricchio for their technical support.

References

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 249–260, 2021.

Alycia N Carey and Xintao Wu. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in Big Data*, 5, 2022.

Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE, 2021.

Victoria Cheng, Vinith M Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 149–160, 2021.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315, 2019.

Michael D Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, pages 35–47. PMLR, 2018.

Jake Fawkes, Robin Evans, and Dino Sejdinovic. Selection, ignorability and challenges with causal fairness. In *Conference on Causal Learning and Reasoning*, pages 275–289. PMLR, 2022.

Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. In *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, page to appear, 2022.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR, 2019.

Atoosa Kasirzadeh and Damian Clifford. Fairness and data protection impact assessments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 146–153, 2021.

Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236, 2021.

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.

Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, pages 2630–2639. PMLR, 2018.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Issa Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. 113 Nw. U. L. Rev. 1163, 2019.

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

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135, 2018.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

Carlos Mougan, Jose M Alvarez, Gourab K Patro, Salvatore Ruggieri, and Steffen Staab. Fairness implications of encoding protected categorical attributes. *arXiv preprint arXiv:2201.11358*, 2022.

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.

Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2020.