It's About Time: Counterfactual Fairness and Temporal Depth

Joshua R. Loftus[1]

[1]London School of Economics, Houghton Street, London, WC2A 2AE, United Kingdom

Abstract

Focusing on time opens up interesting lines of inquiry for algorithmic fairness. In the framework of counterfactual fairness, we can use temporal depth of counterfactuals to reason about common fairness ideals like opportunity, merit, and responsibility. In typical fairness applications greater temporal depth generally corresponds to stronger fairness requirements. We relate counterfactual depth to other causal criteria like direct and indirect effects, and comment on long-standing debates about causation without manipulation and the use of socially constructed traits like race and gender as variables with causal effects. There are diverse and potentially conflicting criteria for algorithmic fairness. Heuristics like temporal depth can help us reason about fairness in a unified way, compare differing criteria, and make good decisions.

Keywords

Algorithmic fairness, counterfactual fairness, causality, counterfactuals, temporal depth, time

1. Counterfactual fairness

Causal modeling methods can be used to study algorithmic fairness [1–8]. See [9] for a recent review on causal machine learning in general and [10, 11] on causal fairness in particular. In this work we focus on counterfactual fairness [2] or **CF**. An algorithm satisfies **CF** if its prediction/decision for a person is the same (probabilistically) as it would be in a counterfactual world where that person had a different value for a given sensitive attribute. For example, we may ask if the prediction for a man is the same as it would be if they were a woman. This requires answering:

1. What else about a person might be different in a counterfactual world?
2. How should any of these differences factor into making fair predictions?

Even in simple examples like Figure 1 there are a variety of approaches to answering the previous questions that lead to different conclusions about fairness, so this is a consequential choice.

With graphical models we can reason about different causal pathways, and path-specific counterfactual fairness or **PCF** weakens the requirement of **CF** by attempting to block the influence of the sensitive attribute on some pathways while allowing it on others if a variable in

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
J.R.Loftus@lse.ac.uk (J. R. Loftus)
https://joshualoftus.com/ (J. R. Loftus)
0000-0002-2905-1632 (J. R. Loftus)

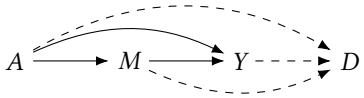

Figure 1: Graphical causal model. A sensitive attribute A has direct causal influences on an outcome Y and a mediator M. The mediator also has a direct influence on Y, resulting in two directed paths from A to Y. Finally, dashed arrows show these variables might be inputs into an algorithm or decision D.

that pathway is considered justifiable or resolving. Perhaps the narrowest version of fairness is a type of **PCF** that attempts to remove only the direct effect $A \dashrightarrow D$ of the sensitive attribute on the decision [12], but this has been extensively criticized [13–15]. Some of these critiques reject causal models like those in Figure 1 altogether, and may have common ground with Holland and Rubin's motto, "no causation without manipulation" [16]. But in some settings it is possible to manipulate the perception of a sensitive attribute or a proxy for it, which could be consistent with removing the direct effect [17, 1].

2. Temporal depth

2.1. Related work

Some previous works on philosophical justifications of fairness criteria [18] have applied heuristic spectra to do so, for example the narrow, middle, and broad views of equality of opportunity in [19]. Others have proposed important differences between causal effects earlier in life vs later [20, 21]. Some work focuses on time in the future [22], particularly those that consider interventions [23] or combinations of interventions and counterfactuals [7, 24]. We aim to build these ideas into a unified framework for understanding different notions of fairness through counterfactual reasoning at different temporal depths.

2.2. Temporality in fairness

In many fairness applications Figure 1 follows a natural temporal ordering. Sensitive attributes A are typically traits like gender or racial status which are (largely) determined early in a person's life. Mediators M are often predictors determined earlier than an outcome Y, and may represent something like preparation, merit, or talent [25, 26], while the outcome itself is some measure of success such as an exam or credit score. Different conceptions of fairness based on equality of opportunity then correspond to **PCF** allowing the pathways to D that pass through M or Y. Standard **CF** is temporally deeper, attempting to undo unfairness all the way back to the root node A.

Temporal depth is not only useful for reasoning about the normative dimension of fairness, but also the empirical or philosophical soundness of counterfactuals. One classic heuristic for counterfactual reasoning argues we should understand a counterfactual world to be as similar as possible to our own [27, 28]. Which notion of similarity we use depends on the counterfactual under consideration. Combining temporal depth and similarity, we may reason that if something had occurred differently at a more recent point in time the resulting counterfactual world would

be closer to our own present than if the difference had occurred earlier. From the example in Figure 1, if unfair discrimination had stopped affecting a person after their value of M was determined, then their counterfactual value of Y may be more similar to the observed Y in our (unfair) world. But if discrimination had ceased earlier, their counterfactual value of M may be improved and hence their counterfactual Y might be more different as well. When considering the same type of counterfactual, greater temporal depth corresponds to greater fairness, and hence also (necessarily) a world less similar to the actual present.

We do not mean to imply these normative and empirical dimensions are disjoint or even separable. It may be that different beliefs about how to achieve fairness correspond to different beliefs about the actual state of the world and/or the degree of similarity between a fair world and our own. And it may be natural for variability in such beliefs to be greater when considering greater temporal ranges. Continuing with a hypothetical hiring process as an example, it may be easier to measure how much unfairness is occurring at the time of the decision due to the presence of names or other social indicators in the applicants' data (CITE), but more difficult to guess how different the pool of applicants would be if the world had achieved robust equality of opportunity generations earlier. Similarly, if we ask how different we can or should try to make the world by tomorrow or in the near future the answers may be more similar than if we extend the time window to a decade, a century, or more. On this basis, so-called longtermism [29] may be a threat to our most trusted decision processes like democratic governance or scientific consensus since it provides more leeway and leverage to justify almost any extreme belief or action.

2.3. Limitations and conclusion

As a heuristic, temporal depth certainly has exceptions. In some settings removing disparities in the data could result in greater unfairness. For example, a social category may correlate with greater exposure to risk factors for a poor health outcome, so a risk prediction algorithm which is **PCF** may be the more fair, more appropriate choice. In this case, using **CF** based on a greater temporal depth could hide the fact of a person's greater exposure from the algorithm. But there may also be settings where someone considers a certain variable to justify disparities and believes **PCF** is appropriate (e.g. equality of opportunity, especially in a narrow sense) because they have simply settled for a less fair outcome. Future work can elaborate the use of temporal depth using a variety of simple archetypal causal models for different domains or for different types of unfairness as in [30].

Imagining how things could be better may not be necessary or sufficient for achieving improvement, but it can be a useful start. We can imagine greater differences by considering deeper counterfactuals, worlds that diverged from our own further in the past. Such counterfactuals may be less straightforwardly applicable to reasoning about practical interventions in the present, but could motivate us to decide what we want to change and understand why [21].

References

[1] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, Avoiding discrimination through causal reasoning, Advances in neural information processing

systems 30 (2017).
[2] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual Fairness, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.
[3] R. Nabi, I. Shpitser, Fair inference on outcomes, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
[4] J. R. Loftus, C. Russell, M. J. Kusner, R. Silva, Causal Reasoning for Algorithmic Fairness, arXiv:1805.05859 [cs] (2018). URL: http://arxiv.org/abs/1805.05859, arXiv: 1805.05859.
[5] J. Zhang, E. Bareinboim, Fairness in decision-making—the causal explanation formula, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
[6] S. Chiappa, Path-specific counterfactual fairness, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 7801–7808.
[7] M. Kusner, C. Russell, J. Loftus, R. Silva, Making Decisions that Reduce Discriminatory Impacts, in: Proceedings of the 36th International Conference on Machine Learning, PMLR, 2019, pp. 3591–3600. URL: https://proceedings.mlr.press/v97/kusner19a.html, iSSN: 2640-3498.
[8] M. J. Kusner, J. R. Loftus, The long road to fairer algorithms, Nature 578 (2020) 34–36. URL: https://www.nature.com/articles/d41586-020-00274-3. doi:10.1038/d41586-020-00274-3, bandiera_abtest: a Cg_type: Comment Number: 7793 Publisher: Nature Publishing Group Subject_term: Computer science, Ethics, Society.
[9] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, R. Silva, Causal machine learning: A survey and open problems, arXiv preprint arXiv:2206.15475 (2022).
[10] K. Makhlouf, S. Zhioua, C. Palamidessi, Survey on causal-based machine learning fairness notions, arXiv preprint arXiv:2010.09553 (2020).
[11] D. Plecko, E. Bareinboim, Causal fairness analysis, arXiv preprint arXiv:2207.11385 (2022).
[12] J. Pearl, Causal inference in statistics: An overview, Statistics Surveys 3 (2009) 96 – 146. URL: https://doi.org/10.1214/09-SS057. doi:10.1214/09-SS057.
[13] I. Kohler-Hausmann, Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination, Nw. UL Rev. 113 (2018) 1163.
[14] L. Hu, What is "race" in algorithmic discrimination on the basis of race, Journal of Moral Philosophy (2021).
[15] L. Hu, I. Kohler-Hausmann, What's sex got to do with machine learning?, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 513–513.
[16] P. W. Holland, Statistics and causal inference, Journal of the American statistical Association 81 (1986) 945–960.
[17] M. Bertrand, S. Mullainathan, Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination, American economic review 94 (2004) 991–1013.
[18] C. Hertweck, C. Heitz, M. Loi, On the moral justification of statistical parity, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 747–757.
[19] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, fairmlbook.org, 2019. http://www.fairmlbook.org.

[20] T. J. VanderWeele, W. R. Robinson, On causal interpretation of race in regressions adjusting for confounding and mediating variables, Epidemiology (Cambridge, Mass.) 25 (2014) 473.
[21] C. Glymour, M. R. Glymour, Commentary: race and sex are causes, Epidemiology 25 (2014) 488–490.
[22] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, M. Hardt, Delayed impact of fair machine learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 3150–3158.
[23] K. Imai, Z. Jiang, Principal fairness for human and algorithmic decision-making, Statistical Science 1 (2023) 1–12.
[24] L. Bynum, J. Loftus, J. Stoyanovich, Counterfactuals for the future, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023.
[25] M. Kasy, R. Abebe, Fairness, equality, and power in algorithmic decision-making, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 576–586.
[26] M. S. A. Lee, L. Floridi, J. Singh, Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics, AI and Ethics 1 (2021) 529–544.
[27] W. Todd, Counterfactual conditionals and the presuppositions of induction, Philosophy of Science 31 (1964) 101–110.
[28] D. Lewis, Counterfactuals, Blackwell, 1973.
[29] W. MacAskill, What we owe the future, Basic books, 2022.
[30] A. Castelnovo, R. Crupi, N. Inverardi, D. Regoli, A. Cosentini, Investigating bias with a synthetic data generator: Empirical evidence and philosophical interpretation, in: 1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, 2022.