Survey on Causal-based Algorithmic Fairness Notions

Anonymous Author(s)

Abstract

In this paper, we discuss the importance of causal reasoning when it comes to creating fair algorithms for decision making such as understanding where sources of bias come from in order. We provide a literature review of existing approaches to fairness, describe concepts and work done in causality for understanding causal approaches, argue the importance of causality in fairness, and give a brief analysis of the some recent approaches to fairness based on causality.

1 Introduction

Fairness is becoming a very popular topics in the field of machine learning in the recent years. Machine learning algorithms are used for solving a variety of problems, but, it can introduce or create discriminatory decisions that are biased against certain individuals or groups. When designing automated decision-making systems, they can lead to discrimination against individuals or certain groups. For the affected individuals, the outcomes are unfair as they are caused by factors beyond their control. Hence, when designing algorithms, decisions should be be made independent from factors outside an an individual's control, such as gender, ethnicity, or place of birth [1]. Unfairness in machine learning systems is mainly due to human bias existing in the training data. Research into such problems is called algorithmic fairness. It is a challenging area of research for mainly two reasons: many features are linked to classes like race or gender, and secondly, what exactly algorithmic fairness means. As to whether they should always be fair when dealing with similar groups or all genders since not all these criteria can be satisfied at the same time [2].

Causal-based fairness notions differ from the previous statistical fairness approaches in that they are not totally based on data but consider additional knowledge about the structure of the world, in the form of a causal model [1]. This additional knowledge helps us understand how data is generated in the first place and how changes in variables propagate in a system. Most of these fairness notions are defined in terms of non-observable quantities such as interventions (to simulate random experiments) and counterfactuals (which consider other hypothetical worlds, in addition to the actual world).

Here, in this work, we discuss how we can rectify bias using causal reasoning, review existing notions of fairness in prediction problems, different tools used in causal reasoning, and how they can be combined using techniques like counterfactual fairness [3]. Section 2 explains the possible causes of unfairness. Section 3 describes the measures of bias in algorithmic fairness. Section 4 introduces about causal models. Section 5 is about the importance of causality in fairness. And finally, section 6 is about causal notions of fairness.

2 Causes of Unfairness

In [4, 5, 6], some causes of unfairness in machine learning are stated as:

A - has done survey on papers related to sections 1, 3 and 6.
B - has done survey on papers related to sections 2, 4, 5 and 7.

Submitted to *Causal Inference in Machine Learning* (CS886, 2020 Fall). Do not distribute.

- Datasets which are used for learning (based on biased device measurements) have biases included in them. Machine learning algorithms are designed to replicate these biases.
- Missing data cause biases as the the datasets are not representative of the target population.
- Algorithmic objectives (which aim at minimizing overall aggregated prediction errors and therefore discriminate against minorities) give rise to biases.
- Proxy attributes for sensitive attributes cause biases. Sensitive attributes are attributes such as race, gender and age, and are typically not used in decision making. Proxy attributes are non-sensitive attributes that can be used to derive sensitive attributes.

3 Measures of Bias in Algorithmic Fairness

Here, we use capital letters to refer to variables (A for protected attribute such as gender) and lower case letters to refer to a value an attribute can take (such as a for a woman or a' for a man). Y refers to the variable we are predicting and $\hat{Y}$ refers to the prediction of the variable Y. We also use $P(.|.)$ to represent conditional probability of events.

3.1 Equalized Odds

Equalized odds, also called Separation, Positive Rate Parity, was first proposed in [7]. It says that if a person has state y, the classifier will predict this at the same rate no matter what the value is of the protected attribute. It can be written as :

$$P(\hat{Y} = y|A = a, Y = y) = P(\hat{Y} = y|A = a', Y = y) \tag{1}$$

for all y, a, a'.

3.2 Calibration

This condition states that if the classifier predicts that a person has state y, their probability of actually having state y should be the same for all choices of attribute [8]. It reverses the condition of equalised odds.

$$P(Y = y|A = a, \hat{Y} = y) = P(Y = y|A = a', \hat{Y} = y) \tag{2}$$

for all y, a, a'.

3.3 Demographic Parity

Demographic Parity is defined as follows:

$$P(\hat{Y} = y|A = a) = P(\hat{Y} = y|A = a') \tag{3}$$

for all y, a, a'. Demographic Parity has been used for several purposes in the following works: [9, 10, 11, 12, 13, 14].

3.4 Individual Fairness

In [15], Individual Fairness is defined as follows:

$$P(\hat{Y}^{(p)} = y|X^{(p)}, A^{(p)}) \approx P(\hat{Y}^{(q)} = y|X^{(q)}, A^{(q)}),\ if\ d(p, q) \approx 0, \tag{4}$$

where p, q refer to two different individuals and (p), (q) are their associated data. The function $d(.,.)$ describes how a pair of individuals should be treated similarly in a fair world.

3.5 Causal Notions of Fairness

A number of recent works apply causal approaches to address fairness [16, 17, 18, 3, 18, 19]. We have tried to cover most of them in detail in Section 5. We have also described background on causal reasoning in Section 4. These works depart from the previous approaches in that they are not wholly

data-driven but require additional knowledge of the structure of the world, in the form of a causal model. This additional part is particularly valuable as it tells us how changes in variables propagate in a system, be it natural, engineered or social. Explicit causal assumptions remove ambiguity from methods that just depend upon statistical correlations.

4 Causal Models

Given two random variables X and Y, we say that X causes Y when there exist at least two different interventions on X that result in two different probability distributions of Y. This does not mean we will be able to define what an intervention is without using causal concepts. In this paper we will make use primarily of the structural causal model (SCM) framework advocated by [20], which shares a lot in common with the approaches by [21] and [22].

4.1 Structural Causal Model

We define a causal model as a triplet (U; V; F) of sets such that:

V is a set of observed random variables that forms the causal system.

U is a set of latent background variables that will represent all the possible causes of V and jointly follows distribution $P(U)$.

F is a set of functions $\{f_1, f_2, ..., f_n\}$, one of each $V_i \, \epsilon \, V$, such that $V_i = f_i(pa_i \, , U_{pa_i})$, $pa_i \subseteq V \backslash \{V_i\}$ and $U_{pai} \subseteq U$. Such equations are known as structural equations.

The notation pa_i is meant to capture the notion that a directed graph $\mathcal{G}$ can be used to represent the input-output relationship encoded in the structural equations: each vertex X in $\mathcal{G}$ corresponds to one random variable in $V \cup U$, with the same symbol used to represent both vertex and and the random variable; an edge $X \rightarrow V_i$ is added to $\mathcal{G}$ if X is one of the arguments of $V_i = f_i(\cdot)$. Hence, X is said to be a parent of V_i in $\mathcal{G}$. We also assume here that $\mathcal{G}$ is acyclic. A SCM is causal in the sense it allows us to predict effects of causes and to infer counterfactuals [23].

4.2 Interventions and Counterfactuals

A perfect $intervention$ on a variable V_i, at value v, corresponds to overriding $f_i(\cdot)$ with the equation $V_i = v$. Then the joint distribution of the remaining variables $V_{\backslash i} \equiv V \backslash \{V_i\}$ is given by the causal model. We will denote this operation as $P(V_{\backslash i} \mid do(Vi = v_i))$. This notion is extendable to a set of simultaneous interventions on multiple variables.

For example, consider the following structural equations:

$$Z = U_z \tag{5}$$

$$A = \lambda_{az} Z + U_A \tag{6}$$

$$Y = \lambda_{ya} A + \lambda_{yz} Z + U_Y \tag{7}$$

The graph for the above is shown in Figure 1(a). Assuming that the background variables follow a standard Gaussian with diagonal covariance matrix, standard algebraic manipulations allows us to calculate that $P(Y = y \mid A = a)$ has a Gaussian density with a mean that depends on λ_{az}, λ_{ya} and λ_{yz}. In contrast, $E[Y \mid do(A = a)] = \lambda_{ya} a$, which can be obtained by first erasing (6) and replacing A with a on the right-hand side of (7) followed by marginalizing the remaining variables. Figure 1(b), which represents a factual world and two parallel worlds where A is set to intervention levels a and a'. A joint distribution for $Y(a)$ and $Y(a')$ is implied by the model. Figure 1(c) shows the case for interventions on Y . It is not difficult to show, as Y is not an ancestor of A in the graph, that $A(y, u) = A(y', u) = A(u)$ for all u, y, y'. This shows that Y does not cause A.

The name $counterfactual$ means that, if the corresponding event already took place, then any such alternative outcomes would be contrary to the realised facts.

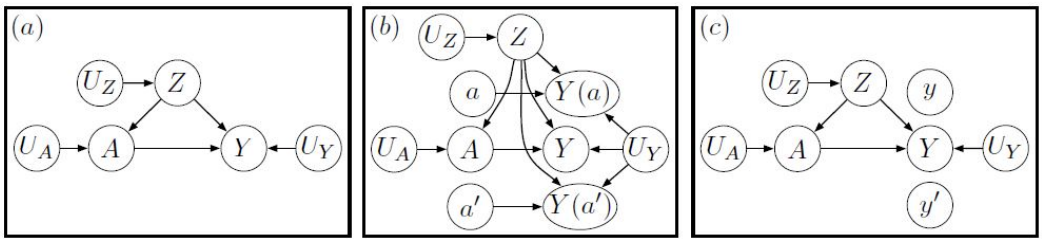

Figure 1: (a) A causal graph for three observed variables A, Y, Z. (b) A joint representation with explicit background variables, and two counterfactual alternatives where A is intervened at two different levels. (c) Similar to (b), where the interventions take place on Y [23].

Another commonly used term is potential outcomes, a terminology meaning that such outcomes are not truly counterfactual until an exposure effectively takes place.

4.3 Counterfactuals Require Untestable Assumptions

Structural equations cannot be tested for correctness unless they depend on observed variables only. Different structural equations with the same interventional distributions mean different joint distributions over counterfactuals [23]. The traditional approach for causal inference in statistics tries to avoid any estimand that cannot be expressed by the marginal distributions of the counterfactuals. Models that follow this approach and only specify the univariate marginals of a counterfactual joint distribution are sometimes called single-world models [24]. However, cross-world models seem better for algorithmic fairness. They are mostly required for non-trivial statements that concern fairness at an individual level as opposed to fairness measures averaged over groups of individuals.

5 Importance of Causality in Fairness

It is important to minimise or remove causal dependence on factors that are not under individual control while designing systems that impact people's life [23]. For example, birth place or their perceived race. These kind of factors have influence on people's life and is considered important for determining fairness.

Many notions of algorithmic fairness have attempted to adjust the covariates. It is possible to produce similar predictions or decisions with a model that is mathematically equivalent. The design decisions underlying a covariate adjustments are sometimes based on implicit causal reasoning, if causal assumptions or interpretations are not provided. The Berkeley Admissions example is explained below to understand the fundamental benefit from the explicit statement of these assumptions.

5.1 Berkeley Admissions

Berkeley Admissions is a classic example of bias in graduate admissions [25]. This is mostly used to explain Simpson's paradox [26] and to understand the importance of adjusting of covariates. In 1973, 34.6% of women and 44.3% of men applied to graduate studies were admitted at Berkeley. There was no evidence that the admission decisions were biased against women. The decision was made by the department and they admitted proportions of men and women at approximately the same rate. But a larger proportion of women applied to the most selective department in a lower overall acceptance rate for women.

After controlling for choice of department, the overall outcome seems relatively fair. The outcome can still be considered to be unfair, not due to biased admissions decisions, but rather to the causes of differences in choice of department, such as socialisation.

5.2 Selection Bias and Causality

Unfairness can also occur from bias in how data is sampled or collected. For example, if a cop stops an individual on the street to check possession of drugs. If the stopping criteria is based on

the discrimination that targets particular individuals, then this creates feedback loop that justifies discriminatory practice. The data gathered by the police suggests that $P(Drugs = yes|Race = a) > P(Drugs = yes|Race = a')$, this can be exploited to justify an unbalanced stopping process when police resources are limited. We can assess its fairness by postulating structures analogous to the Berkeley example, where a mechanism such as Race $\rightarrow$ Economic status $\rightarrow$ Drugs explains the pathway.

5.3 Fairness Requires Intervention

Algorithmic fairness's approaches usually involve imposing constraints on the algorithm. We can view this as an intervention on the predicted outcome $\hat{Y}$. And then, we can try to understand the causal implications for the system we are intervening on. We can use an SCM to model the causal relationships between variables in the data, between those and the predictor $\hat{Y}$ that we are intervening on, and between $\hat{Y}$ and other aspects of the system that will be impacted by decisions made based on the output of the algorithm. In particular, not imposing fairness can also be a deliberate intervention, albeit one of inaction.

6 Causal Notions of Fairness

The following section explains how counterfactual fairness relates to some well-known notions of statistical fairness and ways in which a causal perspective contributes to their interpretation [23].

6.1 Counterfactual Fairness

Informally, $counterfactuals$ are outcomes resulting from alternative interventions on the same $unit$. A $unit$ here can be understood as the snapshot of a system at a specific context. A predictor $\hat{Y}$ is said to satisfy $counterfactual\ fairness$ if

$$P(\hat{Y}(a, U) = y|X = x, A = a) = P(\hat{Y}(a', U) = y|X = x, A = a), \tag{8}$$

for all y, a, a'. Here, U are background variables that describe a particular individual at some point in time. This concept means that background variables being equal, the prediction would not change in the parallel world where only A would have changed. The application of counterfactual fairness require a causal model $\mathcal{M}$ to be formulated, and training data for the observed variables X, A and target variable Y.

Counterfactual fairness relaxes the rigid restriction on any descendant to less strict limitations. For example, the graph does not suffer from proxy discrimination [27] if the predicted label is not dependent on any proxy of the sensitive attribute. The graph also does not suffer from unresolved discrimination if the predicted label is not dependent on any resolving variable (a resolving variable is influenced by the sensitive feature but is accepted by practitioners as non-discriminatory) [4].

6.2 Counterfactual Fairness and its relation to Common Statistical Notions

In this section, we see how counterfactual fairness can be related to non-causal, statistical notions of fairness [23].

6.2.1 Equalised Odds and Calibration

A sufficient condition for $Y \perp\!\!\!\perp A$ (Y is independent of A) is that there are no causal paths between Y and A [22]. In this situation, it can be shown via a graph that a counterfactually fair $\hat{Y}$ made from non-descendants of A in the graph will respect both equalised odds ($\hat{Y} \perp\!\!\!\perp A \mid Y$) and calibration ($Y \perp\!\!\!\perp A \mid \hat{Y}$). It can also be argued that if $Y \not\perp\!\!\!\perp A$, then neither equalised odds nor calibration are desirable. For example, if A is an ancestor of Y, then Y should not be tried to be reproduced since it carries bias according to the counterfactual definition ($Y = \hat{Y}$).

6.2.2 Demographic Parity

If background variables U are uniquely determined by observed data $\{X = a, A = a\}$, and $\hat{Y}$ is a function only of U and observed variables independent of A, then a counterfactually fair predictor

will satisfy eq. (3). In this case, counterfactual fairness can be seen as a counterfactual analogue of demographic parity.

6.2.3 Individual Fairness

As defined in eq. (4), if the two individuals are thought to be matched by a statistical matching procedure [28], then they can be thought of as counterfactual versions of each other. In this case, the counterfactual version of individual p that is used to estimate a causal effect is in reality an observed case q in a sample of controls, such that p and q are close [28]. The term 'closeness' can be specified by the distance metric in eq. (4). Here, the interpretation of the individual fairness condition is similar to a particular instantiation of counterfactual fairness (via matching).

6.3 Framework to Categorise Causal Reasoning in Fairness

In this section, [23] suggest categorizing causal fairness methods according to the following dimensions:

- Explicit vs Implicit structural equations: Even though counterfactual models need untestable assumptions, not all assumptions are equally created. In some cases, we can achieve some degree of fairness by postulating independence constraints among counterfactuals but this might lead to losing significant information.
- Prediction vs Explanation: The task might be to explain in which way a process is discriminatory instead of creating fair predictors.
- Individual vs Group level causal effects: Counterfactual reasoning takes place at the individual level. Even though there advantages when it comes to group effects (as they do not require postulating a joint distribution of two or more outcomes which cannot be observed together, where the existence of counterfactuals can be treated as a metaphysical concept [29]), fairness is more commonly understood at an individual level, where in many cases, unit-level assumptions are needed.

The paper [23] classifies counterfactual fairness as a concept that (i) operates at the individual level, (ii) has explicit structural equations and, (iii) is used for prediction tasks. Using this framework, we can see how existing notions are categorised and how they relate to algorithmic fairness.

6.3.1 Interventional Notions

Since the nature of counterfactuals are ultimately untestable, it is better to avoid several of its assumptions whenever possible. To do this, constraints can be defined on the interventional distributions $P(\hat{Y}|\, do(A = a), X = x)$. The work done in [21], shows this aspect. The interventional notion by [27] is the constraint:

$$P(\hat{Y}|\, do(A = a)) = P(\hat{Y}|\, do(A = a'))$$

A predictor $\hat{Y}$ is made from starting from the causal model $\mathcal{M}$. Modifications are made to $\mathcal{M}$ to create a family of models so that the total effect of A on $\hat{Y}$ is cancelled. The family of models itself can be parameterised so that minimising error with respect to Y is possible within this constrained space.

6.3.2 Excluding Explicit Structural Equations

Since the goal is to minimize the number of untestable assumptions, one way to achieve it is to avoid making any explicit assumptions about the structural equations of the causal model. In this field, we can make assumptions about the directed edges in a causal graph, parametric formulation for observed distributions and independence constraints among counterfactual variables. But explicit structural equations are not allowed. The work done in [30] has similar ideas. It makes a direct connection to the graphical causal models by providing algorithms for identifying causal estimands. Approaches which do not assume any particular parametric contribution from latent variables, provide a function $P(V)$ (V is the set of all observed variables) that is equal to a casual effect of interest or report such a transformation is not possible [20]. In [30], the main idea is to identify which causal

effects of protected attribute A on outcome Y should be zero. The model can be then fit subject to such a constraint and then predict Y as $\hat{Y}$.

Even though structural equations have advantages, there are disadvantages. Information from any descendant of A that is judged to be on an "unfair path" from A to Y is lost when such constraints are enforced. In many cases, the causal effect of interest cannot be identified. Even if it is, the constrained fitting procedure can be both computationally challenging and not have a clear relationship to the actual loss function of interest in predicting Y. This is because it first requires assuming a model for Y and fitting a projection of the model to the space of constrained distributions. Even though explicit says that structural equations are avoided, the approach still relies on assumptions of counterfactuals being independent [23].

6.3.3 Path-Specific Effects

A common alternative to counterfactual fairness [27, 30, 19] is the focus on path-specific effects. For example, for understanding path-specificity and its relation to fairness, we need to refer to the previously discussed case study of gender bias in the admissions to the University of California at Berkeley in the 1970s: gender (A) and admission (Y). This was found to be associated in the data, which lead to questions about fairness of the admission process. One possible explanation is that this was due to the choice of department each individual was applying to (X). By postulating the causal structure $A \rightarrow X \rightarrow Y$, the paper [23] claims that, even though A is a cause of Y , the mechanism by which it changes Y is "fair" in the sense that free-will is assumed in the choice of department made by each applicant. The problem gets more complicated if edge $A \rightarrow Y$ is also present.

7 Conclusion

In this paper, we presented an overview of causal-based algorithmic fairness approaches and notions. We started by describing the main causes of unfairness, followed by common definitions and measures of fairness, and the importance of causality in fairness. We then presented how causal reasoning can relate to algorithmic fairness. Within a causal-based framework, any mathematical form of fairness can be studied. This approach provides tools for making assumptions that are used in the notions related to fairness. The notion of fairness is important because if it does not agree with the actual causal relationships in the data, then it can result in the outcome of misleading and incorrect outputs. Thus, we can make fair decisions by understanding and using accurate modelling techniques [23]. Causal fairness models can also help to overcome many of the challenges faced with respect to fair prediction tasks.

To conclude, since the use of algorithms is expanding to all aspects of our lives, demanding that automated decisions be more ethical and fair is inevitable. We should aspire to not only develop fairer algorithms, but also to design procedures to reduce biases in the data. Such procedures may rely for example on integrating both humans and algorithms in the decision pipeline.

Acknowledgement

We would like to thank prof. Yaoliang Yu for inspiring us to work on this project.

References

[1] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. "Survey on Causal-based Machine Learning Fairness Notions". *arXiv:2010.09553* (2020).

[2] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores". *arXiv preprint:1609.05807* (2016).

[3] M. Kusner, J. Loftus, C. Russell, and R. Silva. "Counterfactual fairness". *Advances in Neural Information Processing Systems*, vol. 30 (2017), pp. 4066–4076.

[4] Dana Pessach and Erez Shmueli. "Algorithmic Fairness". *arXiv:2001.09784v1* (2020).

[5] Alexandra Chouldechova and Aaron Roth. "The Frontiers of Fairness in Machine Learning" (2018).

[6] Fernando Martínez-Plumed, Cèsar Ferri, David Nieves, and José Hernández-Orallo. "Fairness and Missing Values" (2019).

[7] Moritz Hardt, Eric Price, and et al Nati Srebro. "Equality of Opportunity in Supervised Learning". *Advances in neural information processing systems* (2016), pp. 3315–3323.

[8] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". *Big data*, vol. 5(2) (2017), pp. 153–163.

[9] Harrison Edwards and Amos Storkey. "Censoring Representations with an Adversary". *arXiv preprint arXiv:1511.05897* (2015).

[10] Faisal Kamiran and Toon Calders. "Classifying without discriminating". *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on* (2009), pp. 1–6.

[11] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. "Fairness-Aware Classifier with Prejudice Remover Regularizer". *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2012), pp. 35–60.

[12] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. "The Variational Fair Autoencoder". *arXiv preprint arXiv:1511.00830* (2015).

[13] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. "Fairness Constraints: Mechanisms for Fair Classification". *arXiv preprint arXiv:1507.05259* (2017).

[14] Rich Zemel, Kevin Swersky Yu Wu, Toni Pitassi, and Cynthia Dwork. "Learning Fair Representations". *International Conference on Machine Learning* (2013), pp. 325–333.

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness". *Proceedings of the 3rd innovations in theoretical computer science conference* (2012), pp. 214–226.

[16] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. "Interventions over predictions: Reframing the ethical debate for actuarial risk assessment". *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of Proceedings of Machine Learning Research* (2018), pp. 62–76.

[17] S. Chiappa and T. Gillam. "Path-specific counterfactual fairness". *arXiv:1802.08139* (2018).

[18] C. Russell, M. Kusner, J. Loftus, and R. Silva. "When worlds collide: integrating different counterfactual assumptons in fairness". *Advances in Neural Information Processing Systems*, vol. 30 (2017), pp. 6417–6426.

[19] J. Zhang and E. Bareinboim. "Fairness in Decision-Making: the Causal Explanation Formula". *32nd AAAI Conference on Artificial Intelligence* (2018).

[20] J. Pearl. "Causality: Models, Reasoning and Inference". *Cambridge University Press* (2000).

[21] J. Robins. "A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect". *Mathematical Modelling*, vol. 7 (1986), pp. 1395–1512.

[22] P. Spirtes, C. Glymour, and R. Scheines. "Causation, Prediction and Search". *Lecture Notes in Statistics 81. Springer* (1993).

[23] Joshua R. Loftus, Chris Russell2, Matt J. Kusner, and Ricardo Silva. "Causal Reasoning for Algorithmic Fairness". *arXiv:1805.05859v1* (2018).

[24] T.S. Richardson and J. Robins. "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality". *Working Paper Number 128, Center for Statistics and the Social Sciences, University of Washington* (2013).

[25] Peter J Bickel, Eugene A Hammel, and J William O'Connell. "Sex Bias in Graduate Admissions: Data from Berkeley". *Science*, vol. 187(4175) (1975), pp. 398–404.

[26] Edward H Simpson. "The Interpretation of Interaction in Contingency Tables". *Journal of the Royal Statistical Society. Series B (Methodological)* (1951), pp. 238–241.

[27] N. Kilbertus, M. R. Carulla, G. Parascandolo M. Hardt, D. Janzing, and B. Sch¨olkopf. "Avoiding Discrimination through Causal Reasoning". *Advances in Neural Information Processing Systems*, vol. 30 (2017), pp. 656–666.

[28] S. Morgan and C. Winship. "Counterfactuals and Causal Inference: Methods and Principles for Social Research". *Cambridge University Press* (2015).

[29] A. P. Dawid. "Causal Inference without Counterfactuals". *Journal of the American Statistical Association*, vol. 95 (2000), pp. 407–424.

[30] R. Nabi and I. Shpitser. "Fair Inference on Outcomes". *32nd AAAI Conference on Artificial Intelligence* (2018).