A Causal Bayesian Networks Viewpoint on Fairness

Silvia Chiappa[1] and William S. Isaac[1]

DeepMind London, UK
{csilvia,williamis}@google.com

Abstract. We give a graphical interpretation of unfairness in a dataset as the presence of an unfair causal effect of a sensitive attribute in the underlying causal Bayesian network. This viewpoint allows us to highlight the danger of neglecting the data-generation mechanism common to popular machine learning fairness definitions focusing only on the properties of the model output. By being able to represent complex unfairness scenarios, causal Bayesian networks provide us with a powerful tool to measure unfairness in a dataset and to design techniques for alleviating unfairness in model outputs.

1 Introduction

Machine learning is increasingly used in a wide range of decision-making scenarios that have serious implications for individuals and society, including financial lending [10, 44], hiring [8, 27], online advertising [26, 38], pretrial and immigration detention [5, 40], child maltreatment screening [13, 45], health care [18, 31], and social services [1, 22]. Whilst this has the potential to overcome irrational aspects of human decision-making, there is raising concern that bias in the data and model inaccuracies can lead to decisions that treat unrepresented or historically discriminated groups unfavourably. The research community has therefore started to investigate how to ensure that learned models do not take decisions that are *unfair* with respect to *sensitive attributes* (*e.g.* race and gender).

This effort has led to the emergence of a fairness taxonomy [12, 14, 20, 23] providing researchers and practitioners with diagnostics tools to evaluate existing systems or with constraints when designing new ones. Fairness definitions in this taxonomy mainly focus on the relationship between the model output and the sensitive attribute. Many such definitions have been found to be mathematically incompatible [7, 12, 14, 15, 29], and this has been viewed as representing an unavoidable trade-off establishing fundamental limits on fair machine learning, or as an indication that certain definitions do not map on to traditional social or legal understandings of the concept [16].

In this manuscript, we show that fairness evaluation on a model requires careful considerations on the patterns of unfairness underlying the training data. We demonstrate this by viewing unfairness in a dataset as the presence of an unfair *causal effect* of the sensitive attribute in the causal Bayesian network

representing the data-generation mechanism. We show that, when applied to the COMPAS pretrial risk assessment tool, such considerations lead to doubting the appropriateness of currently considered fairness definitions. Finally, we show that causal Bayesian networks offer a powerful tool for representing and reasoning about complex fairness scenarios.

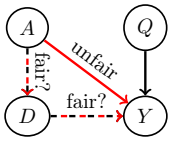
2 A Graphical View of (Un)fairness

Consider a dataset $\Delta = \{a^n, x^n, y^n\}_{n=1}^N$, corresponding to N individuals, where a^n indicates a sensitive attribute, and x^n a set of observations that can be used (together with a^n) to form a prediction $\hat{y}^n$ of y^n. We assume a binary setting $a^n, y^n, \hat{y}^n \in \{0, 1\}$ (unless otherwise specified), and indicate with $A, \mathcal{X}$, Y, and $\hat{Y}$ the (set of) random variables[1] corresponding to a^n, x^n, y^n, and $\hat{y}^n$.

Many fairness definitions are concerned with statistical properties of $\hat{Y}$ with respect to A. In this section we show at a high-level that such types of definitions cannot safely and correctly be used without an understanding of the patterns of unfairness underlying Δ, and therefore of the relationships among A, $\mathcal{X}$ and Y. More specifically we show that:

(i) Using the framework of causal Bayesian networks (CBNs), unfairness in Δ can be viewed as the presence of an unfair causal path from A to $\mathcal{X}$ or Y.

(ii) In order to determine which properties $\hat{Y}$ should possess to be fair, it is necessary to question and understand unfairness in Δ.



Assume a dataset $\Delta = \{a^n, x^n = (q^n, d^n), y^n\}_{n=1}^N$ corresponding to a college admission scenario in which applicants are admitted based on three factors (qualifications Q, choice of department D, gender A) and where female applicants apply more often to certain departments. This scenario can be represented by the CBN on the left (see Appendix A for an overview of BNs, and §3 for a detailed treatment of CBNs). The causal path $A \to Y$ represents direct influence of gender A on admission Y, capturing the fact that two individuals with the same qualification and applying to the same department can be treated differently if of different gender. The indirect causal path $A \to D \to Y$ represents influence of A on Y through D, capturing the fact that female applicants more often apply to certain departments. Whilst the direct influence $A \to Y$ is certainly an unfair one, the paths $A \to D$ and $D \to Y$, and therefore $A \to D \to Y$, could either be considered as fair or as unfair. For example, rejecting women more often due to department choice could be considered fair when considering only the responsibility of the college. However, this could be considered unfair when considering societal responsibility if the departmental differences were a result of systemic historical or cultural factors (*e.g.* if female applicants apply to specific departments at lower rates because of overt or covert societal discouragement).

[1] Throughout the paper, we use capital and small letters for random variables and their values, and calligraphic capital letters for sets of variables.

Finally, if the college were to lower the admission rates for departments chosen more often by women, then the path $D \to Y$ would be unfair. Establishing whether a path is fair or unfair requires careful ethical and sociological considerations and/or might not be possible from a dataset alone. Nevertheless, this example illustrates that we can view unfairness in a dataset as the presence of an unfair causal path from the sensitive attribute A to $\mathcal{X}$ or Y.

Different (un)fair path labeling require $\hat{Y}$ to have different characteristics in order to be fair. In the case in which the causal paths from A to Y are all unfair (*e.g.* if $A \to Y$ is present and $A \to D \to Y$ is considered unfair), a prediction $\hat{Y}$ that is independent of A (indicated with $\hat{Y} \perp\!\!\!\perp A$) would not contain any of the unfair influence of A on Y. In such a case, $\hat{Y}$ is said to satisfy *Statistical Parity*.

Statistical Parity (SP). $\hat{Y}$ satisfies Statistical Parity if $\hat{Y} \perp\!\!\!\perp A$, *i.e.* $p(\hat{Y} = 1|A = 0) = p(\hat{Y} = 1|A = 1)$, where $p(\hat{Y} = 1|A = 0)$ can be estimated as

$$p(\hat{Y} = 1|A = 0) \approx \frac{1}{N}\sum_{n=1}^{N} p(\hat{y}^n = 1|a^n = 0, x^n) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\hat{y}^n=1,a^n=0}\,,$$

with $\mathbb{1}_{\hat{y}^n=1,a^n=0}$ equal to one if $\hat{y}^n = 1$ and $a^n = 0$, and to zero otherwise. Notice that many classifiers, such as for example logistic regression, rather than a prediction $\hat{y}^n \in \{0, 1\}$, would output the probability of class 1, $r^n = p(y^n = 1|a^n, x^n)$, also called *risk score* (in some cases, the risk score may not lie in the interval $[0, 1]$, *e.g.* $r^n \in \{1, \ldots, 10\}$ may represent a risk decile). To obtain the prediction $\hat{y}^n \in \{0, 1\}$ from the risk score r^n, it is common to use a threshold θ, *i.e.* $\hat{y}^n = \mathbb{1}_{r^n>\theta}$. In this case, we can rewrite the estimate for $p(\hat{Y} = 1|A = 0)$ as

$$p(\hat{Y} = 1|A = 0) \approx \frac{1}{N}\sum_{n=1}^{N} p(r^n > \theta|a^n = 0, x^n) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{r^n>\theta,a^n=0}\,.$$

Notice that requiring $R \perp\!\!\!\perp A$ would be much stronger than requiring SP, but this could be desirable to produce a model that satisfies SP for all values of θ.

In the case in which the causal paths from A to Y are all fair (*e.g.* if $A \to Y$ is absent and $A \to D \to Y$ is considered fair), and therefore A has a fair influence on Y, the following two criteria, requiring that the model does not contain dependence on A in the prediction error, could be reasonable.

Equal False Positive and Negative Rates (EFPRs/EFNRs). $\hat{Y}$ satisfies EFPRs and EFNRs if $\hat{Y} \perp\!\!\!\perp A|Y$, *i.e.* (EFPRs) $p(\hat{Y} = 0|Y = 1, A = 0) = p(\hat{Y} = 0|Y = 1, A = 1)$ and (EFNRs) $p(\hat{Y} = 1|Y = 0, A = 0) = p(\hat{Y} = 1|Y = 0, A = 1)$.

Calibration. $\hat{Y}$ satisfies Calibration if $Y \perp\!\!\!\perp A|\hat{Y}$. In the case of risk score output r^n, this condition is often instead called *Predictive Parity* at threshold θ, $p(Y = 1|R > \theta, A = 0) = p(Y = 1|R > \theta, A = 1)$, and Calibration defined as requiring $Y \perp\!\!\!\perp A|R$.

In the case in which at least one causal path from A to Y is unfair (*e.g.* if $A \to Y$ is present), EFPRs/EFNRs and Calibration are inappropriate criteria. This is because these metrics only test if $\hat{Y}$ contains additional dependence on A besides the one in Y, rather than if the unfair influence through $A \to Y$ is absent in $\hat{Y}$ (*e.g.* a perfect model, $\hat{Y} = Y$, would automatically satisfy EFPRs/EFNRs and Calibration, but would contain the unfair influence through $A \to Y$).

The second case is particularly relevant to the recent debate surrounding the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) pretrial risk assessment tool. We revisit this debate in the next section.

2.1 The COMPAS Debate

Over the past few years, numerous state and local governments around the United States have sought to reform their pretrial court systems with the aim of reducing unprecedented levels of incarceration, and specifically the population of low-income defendants and racial minorities in America's prisons and jails [2, 24, 30]. As part of this effort, quantitative tools for determining a person's likelihood for reoffending or failure to appear, called *risk assessment instruments* (RAIs), were introduced to replace previous systems driven largely by opaque discretionary decisions and money bail [6, 25]. However, the expansion of pretrial RAIs has unearthed new concerns of racial discrimination which would nullify the purported benefits of these systems and adversely impact defendants' civil liberties.

An intense ongoing debate, in which the research community has also been heavily involved, was triggered by an exposé from investigative journalists at ProPublica [5] on the COMPAS pretrial RPI developed by Equivant (formerly Northpointe) and deployed in Broward County in Florida. The COMPAS general recidivism scale (GRRS) and violent recidivism risk scale (VRRS), the focus of ProPublica's investigation, sought to leverage machine learning techniques to improve the predictive accuracy of recidivism compared to older RAIs such as the Level of Service Inventory-Revised (LSI-R) [3] which were primarily based on theories and techniques from a sub-field of psychology known as the psychology of criminal conduct (PCC) [4, 9][2].

ProPublica's criticism of COMPAS centered on two concerns. First, the authors argued that the distribution of risk scores $R \in \{1, \ldots, 10\}$ exhibited discriminatory patterns ($R \not\perp A$), as black defendants displayed a fairly uniform

[2] While the exact methodology underlying GRRS and VRRS is proprietary, publicly available reports suggest that the process begins with a defendant being administered a 137 point assessment during intake. This is used to create a series of dynamic risk factor scales such as the Criminal Involvement Scale and History of Violence Scale. In addition, COMPAS also includes static attributes such as the defendant's age and prior police contact (number of prior arrests). The raw COMPAS scores are transformed into decile values by ranking and calibration with a normative group to ensure an equal proportion of scores within each scale value. Lastly, to aid practitioner interpretation, the scores are grouped into three risk categories. The scale values are displayed to court officials as either Low (1-4), Medium (5-7), and High (8-10) risk.

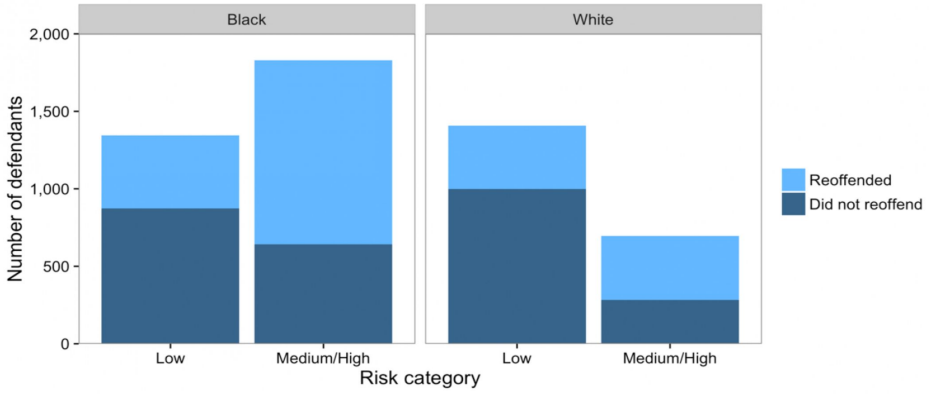


Fig. 1. Number of black and white defendants in each of two aggregate risk categories [15]. The overall recidivism rate for black defendants is higher than for white defendants (52% vs. 39%), *i.e.* $Y \not\perp\!\!\!\perp A$. Within each risk category, the proportion of defendants who reoffend is approximately the same regardless of race, *i.e.* $Y \perp\!\!\!\perp A|\hat{Y}$. Black defendants are more likely to be classified as medium or high risk (58% vs. 33%) *i.e.* $\hat{Y} \not\perp\!\!\!\perp A$. Among individuals who did not reoffend, black defendants are more likely to be classified as medium or high risk than white defendants (44.9% to 23.5%). Among individuals who did reoffend, white defendant are more likely to be classified as low risk than black defendants (47.7% vs 28%), *i.e.* $\hat{Y} \not\perp\!\!\!\perp A|Y$.

distribution across each decile value, while white defendants exhibited a right skewed distribution across decile values, suggesting that the COMPAS recidivism risk scores disproportionately rated white defendants as lower risk than black defendants. Second, the authors claimed that the GRRS and VRRS do not satisfy EFPRs and EFNRs, as FPRs $= 44.9\%$ and FNRs $= 28.0\%$ for black defendants, whilst FPRs $= 23.5\%$ and FNRs $= 47.7\%$ for white defendants (see Fig. 1). This evidence led ProPublica to conclude that COMPAS had a disparate impact on black defendants, leading to public outcry over potential biases in RAIs and machine learning writ large.

In response, Equivant published a technical report [19] refuting the claims of bias made by ProPublica and concluded that COMPAS is sufficiently calibrated, in the sense that it satisfies Predictive Parity at key thresholds. Subsequent analyses [12, 14, 29] confirm Equivant's claims of calibration, but also demonstrate the incompatibility of EFPRs/EFNRs and Calibration due to differences in baseline rates across groups ($Y \not\perp\!\!\!\perp A$) (see Appendix B). Moreover, the finding conclude that attempting to satisfy these competing forms of fairness force unavoidable trade-offs between criminal justice reformers' purported goals of racial equity and public safety.

As explained in §2, EFPRs/EFNRs and Calibration require that the prediction error does not depend on A, and are therefore not concerned with the existing dependence of Y on A. The fact that these criteria cannot be satisfied at the

same time if $Y \not\perp\!\!\!\perp A$ simply reflects the fact that dependence of Y on A, if present in $\hat{Y}$, would also be in the difference between Y and $\hat{Y}$ and therefore appear as $\hat{Y} \not\perp\!\!\!\perp A|Y$ and/or $Y \not\perp\!\!\!\perp A|\hat{Y}$. If such a dependence is fair, this is not problematic. If such a dependence includes influence of A in Y through an unfair causal path, both criteria are inadequate.

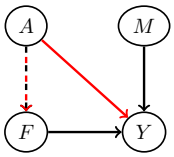


Fig. 2. Possible CBN underlying COMPAS.

As previous research has shown [28, 34, 41], modern policing tactics center around targeting a small number of neighborhoods — often disproportionately populated by non-white and low income residents — with recurring patrols and stops. This uneven distribution of police attention, as well as other factors such as funding for pretrial services [30, 43], means that differences in base rates between racial groups are not reflective of ground truth rates. We can rephrase these findings as indicating the presence of a direct path $A \to Y$ (through unobserved neighborhood) in the CBN underlying the data (Fig. 2). Furthermore, such tactics also imply an influence of A on Y through F containing number of prior arrests. In addition, the influence of A on Y through $A \to Y$ and $A \to F \to Y$ could be more prominent or contain more unfairness due to racism. If any of these issues is present, EFPRs/EFNRs and Calibration would be inappropriate criteria.

Our concern that insufficient consideration has been given to the systemic nature of unfairness underlying RAIs mirror concerns raised by social scientists and legal scholars on mismeasurement and unrepresentative data in the US criminal justice system. Multiple studies [21, 33, 35, 43] have argued that the core premise of RAIs, to assess the likelihood a defendant reoffends, is impossible to measure and that the empirical proxy used (*e.g.* arrest or conviction) introduces embedded biases and norms which renders existing fairness tests unreliable.

This section used the CBNs framework to describe at a high-level different patters of unfairness that can underlie a dataset and to point out issues with current deployment of and debate around fairness definitions. In the reminder of the manuscript, we use this framework more extensively to further advance our analysis on fairness. Before doing that, we give some background on CBNs [17, 36, 37, 39, 42], assuming that variables are continuous.

3 Causal Bayesian Networks

A *Bayesian network* is a *directed acyclic graph* where nodes and edges represent random variables and statistical dependencies. Each node X_i in the graph is associated with the conditional distribution $p(X_i|\text{pa}(X_i))$, where $\text{pa}(X_i)$ is the set of *parents* of X_i. The joint distribution of all nodes, $p(X_1, \ldots, X_I)$, is given by the product of all conditional distributions, *i.e.* $p(X_{1:I} \equiv X_1, \ldots, X_I) = \prod_{i=1}^{I} p(X_i|\text{pa}(X_i))$ (see Appendix A for more details on Bayesian networks).

When equipped with causal semantic, namely when describing the process underlying the data generation, Bayesian networks can be used to visually express causal relationships. More specifically, CBNs enable us to give a graphical

definition of causes and causal effects: if there exists a *directed path* from A to Y, then A is a *potential cause* of Y. Directed paths are also called *causal paths*.

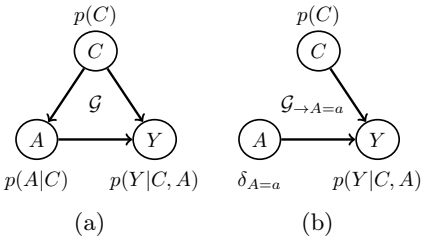


Fig. 3. (a): CBN with a confounder C for the effect of A on Y. (b): Modified CBN resulting from intervening on A.

The causal effect of A on Y can be seen as the information traveling from A to Y through causal paths, or as the conditional distribution of Y given A restricted to causal paths. This implies that, to compute the causal effect, we need to disregard the information that travels along non-causal paths, which happens if such paths are *open*. Since all paths with an arrow emerging from A are either causal or closed (*blocked*) by a *collider*, the problematic paths are only those with an arrow pointing into A, called *back-door paths*. Unless they contain a collider, back-door paths are open and thus transfer non-causal information.

Therefore, if there exist at least one open back-door path between A and Y then the causal effect of A on Y differs from $p(Y|A)$. An example of such a path is $A \leftarrow C \rightarrow Y$ in the CBN $\mathcal{G}$ of Fig. 3(a): the variable C is said to be a *confounder* for the effect of A on Y, as it confounds the causal effect with non-causal information. As an example, assume that A represents hours of exercise in a week, Y cardiac health, and C age. Observing cardiac health conditioning on exercise level from $p(Y|A)$ is insufficient to understand the effect of exercise on cardiac health, since there is always the possibility that dependence between the two is because of the confounder of age.

Each parent-child relationship in a CBN represents an autonomous mechanism, and therefore it is conceivable to change one such relationship without changing the others. This enables us to express the causal effect of $A = a$ on Y as the conditional distribution $p_{\rightarrow A=a}(Y|A = a)$ on the modified CBN $\mathcal{G}_{\rightarrow A=a}$ of Fig. 3(b), resulting from replacing $p(A|C)$ with a Dirac delta distribution $\delta_{A=a}$ (thereby removing the link from C to A) and leaving the remaining conditional distributions $p(Y|A, C)$ and $p(C)$ unaltered – this process is called *intervention* on A. The distribution $p_{\rightarrow A=a}(Y|A = a)$ can be estimated as $p_{\rightarrow A=a}(Y|A = a) = \int_C p_{\rightarrow A=a}(Y|A = a, C)p_{\rightarrow A=a}(C|A = a) = \int_C p(Y|A = a, C)p(C)$. This is a special case of the following back-door adjustment formula.

Back-door Adjustment. If a set of variables $\mathcal{C}$ satisfies the back-door criterion relative to $\{A, Y\}$, the causal effect of A on Y is given by $p_{\rightarrow A}(Y|A) = \int_{\mathcal{C}} p(Y|A, \mathcal{C})p(\mathcal{C})$. $\mathcal{C}$ satisfies the back-door criterion if (a) no node in $\mathcal{C}$ is a *descendant* of A and (b) $\mathcal{C}$ blocks every back-door path from A to Y.

The equality $p_{\rightarrow A=a}(Y|A = a, \mathcal{C}) = p(Y|A = a, \mathcal{C})$ follows from the fact that $\mathcal{G}_{A\rightarrow}$, obtained by removing from $\mathcal{G}$ all the links emerging from A, retains all (and only) the back-door paths from A to Y. As $\mathcal{C}$ blocks all those paths, $Y \perp\!\!\!\perp A|\mathcal{C}$ in $\mathcal{G}_{A\rightarrow}$. This means that no, non-causal, information travels from A to Y when conditioning on $\mathcal{C}$ and therefore conditioning on A coincides with intervening.

Conditioning on C to block an open back-door path may open a closed path on which C is a collider. For example, in the CBN of Fig. 4(a), conditioning on C closes the paths $A \leftarrow C \leftarrow X \rightarrow Y$ and $A \leftarrow C \rightarrow Y$, but opens the path $A \leftarrow E \rightarrow C \leftarrow X \rightarrow Y$ (additional conditioning on X would close this path).

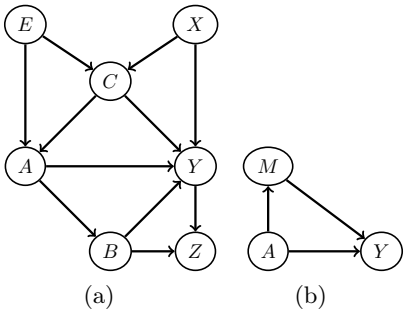


Fig. 4. (a): CBN in which conditioning on C closes the paths $A \leftarrow C \leftarrow X \rightarrow Y$ and $A \leftarrow C \rightarrow Y$, but opens the path $A \leftarrow E \rightarrow C \leftarrow X \rightarrow Y$. (b): CBN with one direct and one indirect causal path from A to Y.

The back-door criterion can also be derived from the rules of do-calculus [36, 37], which indicate whether and how $p_{\rightarrow A}(Y|A)$ can be estimated using observations from $\mathcal{G}$: for many graph structures with unobserved confounders the only way to compute causal effects is by collecting observations directly from $\mathcal{G}_{\rightarrow A}$ – in this case the effect is said to be *non-identifiable*.

Potential Outcome Viewpoint. Let $Y_{A=a}$ be the random variable with distribution $p(Y_{A=a}) = p_{\rightarrow A=a}(Y|A=a)$. $Y_{A=a}$ is called *potential outcome* and, when not ambiguous, we will refer to it with the shorthand Y_a. The relation between Y_a and all the variables in $\mathcal{G}$ other than Y can be expressed by the graph obtained by removing from $\mathcal{G}$ all the links emerging from A and by replacing Y with Y_a. If Y_a is independent on A in this graph, then[3] $p(Y_a) = p(Y_a|A=a) = p(Y|A=a)$. If Y_a is independent of A in this graph when conditioning on $\mathcal{C}$, then

$$p(Y_a) = \int_{\mathcal{C}} p(Y_a|\mathcal{C})p(\mathcal{C}) = \int_{\mathcal{C}} p(Y_a|A=a,\mathcal{C})p(\mathcal{C}) = \int_{\mathcal{C}} p(Y|A=a,\mathcal{C})p(\mathcal{C})\,,$$

i.e. we retrieve the back-door adjustment formula.

In the reminder of the section we show that, by performing different interventions on A along different causal paths, it is possible to isolate the contribution of the causal effect of A on Y along a group of paths.

Direct and Indirect Effect

Consider the CBN of Fig. 4(b), containing the direct path $A \rightarrow Y$ and one indirect causal path through a variable M. Let $Y_a(M_{\bar{a}})$ be the random variable that results from the interventions $A = a$ along $A \rightarrow Y$ and $A = \bar{a}$ along $A \rightarrow M \rightarrow Y$. The *average direct effect* (ADE) of $A = a$ with respect to $A = \bar{a}$, defined as

$$\text{ADE}_{\bar{a}a} = \langle Y_a(M_{\bar{a}})\rangle_{p(Y_a(M_{\bar{a}}))} - \langle Y_{\bar{a}}\rangle_{p(Y_{\bar{a}})}\,,$$

[3] The equality $p(Y_a|A=a) = p(Y|A=a)$ is called *consistency*.

where *e.g.* $\langle Y_a \rangle_{p(Y_a)} = \int_{Y_a} Y_a p(Y_a)$, measures difference in flow of causal information from A to Y when $A = a$ along $A \to Y$ and $A = \bar{a}$ along $A \to M \to Y$ with respect to when $A = \bar{a}$ along both paths.

Analogously, the *average indirect effect* (AIE) of $A = a$ with respect to $A = \bar{a}$, is defined as $\text{AIE}_{\bar{a}a} = \langle Y_{\bar{a}}(M_a) \rangle_{p(Y_{\bar{a}}(M_a))} - \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}})}$.

The difference $\text{ADE}_{\bar{a}a} - \text{AIE}_{a\bar{a}}$ gives the *average total effect* (ATE) $\text{ATE}_{\bar{a}a} = \langle Y_a \rangle_{p(Y_a)} - \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}})}$[4].

Path-Specific Effect

To estimate the effect along a specific group of causal paths, we can generalize the formulas for the ADE and AIE by replacing the variable in the first term with the one resulting from performing the intervention $A = a$ along the group of interest and $A = \bar{a}$ along the remaining causal paths. For example, consider the CBN of Fig. 5 (top) and assume that we are interested in isolating the effect of A on Y along the direct path $A \to Y$ and the paths passing through M, $A \to M \to, \ldots, \to Y$, namely along the red links. The *path-specific effect* (PSE) of $A = a$ with respect to $A = \bar{a}$ for this group of paths is defined as

$$\text{PSE}_{\bar{a}a} = \langle Y_a(M_a, L_{\bar{a}}(M_a)) \rangle - \langle Y_{\bar{a}} \rangle \,,$$

where $p(Y_a(M_a, L_{\bar{a}}(M_a)))$ can be computed as

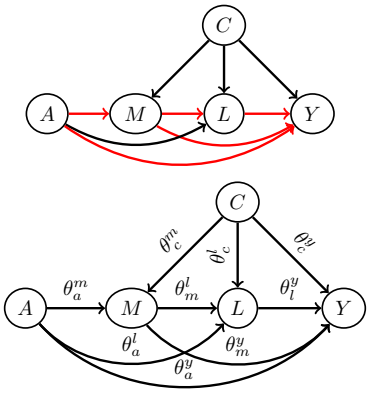


Fig. 5. Top: CBN with direct path and paths passing through M highlighted in red. Bottom: CBN corresponding to Eq. (1).

$$\int_{C,M,L} p(Y|A=a,C,M,L)p(L|A=\bar{a},C,M)p(M|A=a,C)p(C)\,.$$

In the simple case in which the CBN corresponds to a linear model, *e.g.*

$$\begin{aligned} &A \sim \text{Bern}(\pi),\ C = \epsilon_c\,, \\ &M = \theta^m + \theta^m_a A + \theta^m_c C + \epsilon_m,\ L = \theta^l + \theta^l_a A + \theta^l_c C + \theta^l_m M + \epsilon_l\,, \\ &Y = \theta^y + \theta^y_a A + \theta^y_c C + \theta^y_m M + \theta^y_l L + \epsilon_y\,, \end{aligned} \tag{1}$$

where ϵ_c, ϵ_m, ϵ_l and ϵ_y are unobserved independent zero-mean Gaussian variables, we can compute $\langle Y_{\bar{a}} \rangle$ by expressing Y as a function of $A = \bar{a}$ and the Gaussian variables, by recursive substitutions in C, M and L, *i.e.*

$$\begin{aligned} Y = {} & \theta^y + \theta^y_a \bar{a} + \theta^y_c \epsilon_c + \theta^y_m(\theta^m + \theta^m_a \bar{a} + \theta^m_c \epsilon_c + \epsilon_m) \\ & + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_c \epsilon_c + \theta^l_m(\theta^m + \theta^m_a \bar{a} + \theta^m_c \epsilon_c + \epsilon_m) + \epsilon_l) + \epsilon_y\,, \end{aligned}$$

[4] Often the AIE of $A = a$ with respect to $A = \bar{a}$ is defined as $\text{AIE}^a_{\bar{a}a} = \langle Y_a \rangle_{p(Y_a)} - \langle Y_a(M_{\bar{a}}) \rangle_{p(Y_a(M_{\bar{a}}))} = -\text{AIE}_{a\bar{a}}$, which differs in setting A to a rather than to $\bar{a}$ along $A \to Y$. In the linear case, the two definitions coincide (see Eqs. (2) and (3)). Similarly the ADE can be defined as $\text{ADE}^a_{\bar{a}a} = \langle Y_a \rangle_{p(Y_a)} - \langle Y_{\bar{a}}(M_a) \rangle_{p(Y_{\bar{a}}(M_a))} = -\text{ADE}_{a\bar{a}}$.

and then take mean, obtaining $\langle Y_{\bar{a}} \rangle = \theta^y + \theta^y_a \bar{a} + \theta^y_m(\theta^m + \theta^m_a \bar{a}) + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_m(\theta^m + \theta^m_a \bar{a}))$. Analogously

$$\langle Y_a(M_a, L_{\bar{a}}(M_a)) \rangle = \theta^y + \theta^y_a a + \theta^y_m(\theta^m + \theta^m_a a) + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_m(\theta^m + \theta^m_a a)) \,.$$

For $a = 1$ and $\bar{a} = 0$, this gives

$$\text{PSE}_{\bar{a}a} = \theta^y_a(a - \bar{a}) + \theta^y_m \theta^m_a (a - \bar{a}) + \theta^y_l \theta^l_m \theta^m_a (a - \bar{a}) = \theta^y_a + \theta^y_m \theta^m_a + \theta^y_l \theta^l_m \theta^m_a \,.$$

The same conclusion could have been obtained by looking at the graph annotated with path coefficients (Fig. 5 (bottom)). The PSE is obtained by summing over all causal paths from A to Y the products of the coefficients in each path, $\theta^y_a + \theta^m_a \theta^l_m \theta^y_l + \theta^m_a \theta^y_m + \theta^l_a \theta^y_l$.

Notice that $\text{AIE}_{\bar{a}a}$, given by

$$\begin{aligned}
\text{AIE}_{\bar{a}a} &= \langle Y_{\bar{a}}(M_a, L_a(M_a)) \rangle - \langle Y_{\bar{a}} \rangle \\
&= \theta^y + \theta^y_a \bar{a} + \theta^y_m(\theta^m + \theta^m_a a) + \theta^y_l(\theta^l + \theta^l_a a + \theta^l_m(\theta^m + \theta^m_a a)) \\
&\quad - \theta^y + \theta^y_a \bar{a} + \theta^y_m(\theta^m + \theta^m_a \bar{a}) + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_m(\theta^m + \theta^m_a \bar{a})) \\
&= \theta^y_m \theta^m_a (a - \bar{a}) + \theta^y_l(\theta^l_a(a - \bar{a}) + \theta^l_m \theta^m_a (a - \bar{a})) \,, \qquad (2)
\end{aligned}$$

coincides with $\text{AIE}^a_{\bar{a}a}$, given by

$$\begin{aligned}
\text{AIE}^a_{\bar{a}a} &= \langle Y_a \rangle - \langle Y_a(M_{\bar{a}}, L_{\bar{a}}(M_{\bar{a}})) \rangle \\
&= \theta^y + \theta^y_a a + \theta^y_m(\theta^m + \theta^m_a a) + \theta^y_l(\theta^l + \theta^l_a a + \theta^l_m(\theta^m + \theta^m_a a)) \\
&\quad - \theta^y + \theta^y_a a + \theta^y_m(\theta^m + \theta^m_a \bar{a}) + \theta^y_l(\theta^l + \theta^l_a \bar{a} + \theta^l_m(\theta^m + \theta^m_a \bar{a})) \,. \qquad (3)
\end{aligned}$$

Effect of Treatment on Treated (ETT). Consider the conditional distribution $p(Y_a|A = \bar{a})$. This distribution measures the information travelling from A to Y along all open paths, when A is set to a along causal paths and to $\bar{a}$ along non-causal paths. The *effect of treatment on treated* (ETT) of $A = a$ with respect to $A = \bar{a}$ is defined as $\text{ETT}_{\bar{a}a} = \langle Y_a \rangle_{p(Y_a|A=\bar{a})} - \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}}|A=\bar{a})} = \langle Y_a \rangle_{p(Y_a|A=\bar{a})} - \langle Y \rangle_{p(Y|A=\bar{a})}$. As the PSE, the ETT measures difference in flow of information from A to Y when A takes different values along different paths. However, the PSE considers only causal paths and different values for A along different causal paths, whilst the ETT considers all open paths and different values for A along causal and non-causal paths respectively. Notice that, similarly to $\text{ATE}_{\bar{a}a}$, $\text{ETT}_{\bar{a}a}$ for the CBN of Fig. 4(b) can be expressed as

$$\text{ETT}_{\bar{a}a} = \underbrace{\langle Y_a(M_{\bar{a}}) \rangle - \langle Y_{\bar{a}} \rangle}_{\text{ADE}_{\bar{a}a|\bar{a}}} - \underbrace{(\langle Y_a(M_{\bar{a}}) \rangle - \langle Y_a \rangle)}_{\text{AIE}_{a\bar{a}|\bar{a}}} \,.$$

Notice that, if we define difference in flow of non-causal (along the open back-door paths) information from A to Y when $A = a$ with respect to when $A = \bar{a}$ as $\text{NCI}_{\bar{a}a} = \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}}|A=a)} - \langle Y \rangle_{p(Y|A=\bar{a})}$, we obtain

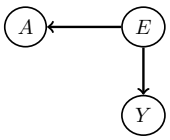
$$\begin{aligned}
\langle Y \rangle_{p(Y|A=a)} - \langle Y \rangle_{p(Y|A=\bar{a})} &= \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}}|A=a)} - \langle Y \rangle_{p(Y|A=\bar{a})} \\
&\quad - (\langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}}|A=a)} - \langle Y \rangle_{p(Y|A=a)}) \\
&= \text{NCI}_{\bar{a}a} - \text{ETT}_{a\bar{a}} = \text{NCI}_{\bar{a}a} - \text{ADE}_{a\bar{a}|a} + \text{AIE}_{\bar{a}a|a} \,.
\end{aligned}$$

4 Fairness Considerations using CBNs

Equipped with the background on CBNs from §3, in this section we further investigate unfairness in a dataset $\Delta = \{a^n, x^n, y^n\}_{n=1}^N$ and issues around building a decision system from it, revisiting and extending examples and material from [11, 32, 46].

Back-door Paths from A to Y

In §2 we have introduced a graphical interpretation of unfairness in a dataset Δ as the presence of an unfair causal path from A to $\mathcal{X}$ or Y. More specifically, we have shown through several examples that unfairness can be due to an unfair link emerging (a) from A or (b) from a subsequent variable in a causal path from A to Y (*e.g.* $D \to Y$ in the college admission example). Our discussion did not mention paths from A to Y with non emerging links from A, namely back-door paths. This is because such paths are not problematic.



An example with a back-door path from A to Y is given by the hiring scenario described on the left, where A represents religious belief and E represents educational background of the applicant, which influences religious participation ($E \to A$). Whilst $Y \not\perp\!\!\!\perp A$, the hiring decision Y is only based on E.

Opening Closed Unfair Paths from A to Y

In §2, we have seen that, in order to reason about fairness of $\hat{Y}$, it is necessary to question and understand unfairness in Δ. In this section, we explain that another crucial element needs to be considered in the fairness discussion around $\hat{Y}$, namely

(i) The subset of variables used to form $\hat{Y}$ could project into $\hat{Y}$ unfair patterns in $\mathcal{X}$ that do not concern Y.

This could happen, for example, if a closed unfair path from A to Y is opened when conditioning on the variables used to form $\hat{Y}$.

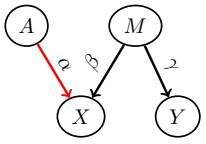


Fig. 6. CBN underlying a music degree scenario.

As an example, assume the CBN in Fig. 6 representing the data generation process underlying a music degree scenario, where A corresponds to gender, M to music aptitude (unobserved, *i.e.* $M \notin \Delta$), X to the score obtained from an ability test taken at the beginning of the degree, and Y to the score obtained from an ability test taken at the end of the degree. Individuals with higher music aptitude M are more likely to obtain higher initial and final scores ($M \to X$, $M \to Y$). Due to discrimination occurring at the initial testing, women are assigned a lower initial score than men for the same aptitude level ($A \to X$). The only path from A to Y, $A \to X \leftarrow M \to Y$, is closed as X is a collider on this path. Therefore the unfair influence of A on X does not reach Y ($Y \perp\!\!\!\perp A$).

Nevertheless, as $Y \not\perp\!\!\!\perp A|X$, a prediction $\hat{Y}$ based on the initial score X only will contain the unfair influence of A on X.

For example, assume the following linear model: $Y = \gamma M, \quad X = \alpha A + \beta M$, with $Var(A) = 1$, $Var(M) = 1$, $Cov(A, M) = 0$. A linear predictor of the form $\hat{Y} = \theta_X X$ would have optimal parameters $\theta_X = \frac{1}{Var(X)} Cov(X, Y) = \frac{1}{\alpha^2+\beta^2}\gamma\beta$, giving $\hat{Y} = \frac{\gamma\beta}{\alpha^2+\beta^2}(\alpha A + \beta M)$, *i.e.* $\hat{Y} \not\perp\!\!\!\perp A$. Therefore, this predictor would be using the sensitive attribute to form a decision, although implicitly rather than explicitly. Instead, a predictor explicitly using the sensitive attribute, $\hat{Y} = \theta_X X + \theta_A A$, would have optimal parameters

$$\begin{pmatrix} \theta_X \\ \theta_A \end{pmatrix} = \begin{pmatrix} \alpha^2 + \beta^2 & \alpha \\ \alpha & 1 \end{pmatrix}^{-1} \begin{pmatrix} \gamma\beta \\ 0 \end{pmatrix} = \begin{pmatrix} \gamma/\beta \\ -\alpha\gamma/\beta \end{pmatrix},$$

i.e. $\hat{Y} = \frac{\gamma}{\beta}(\alpha A + \beta M) - \frac{\alpha\gamma}{\beta} A = \gamma M$. Therefore, this predictor would be fair. From the CBN we can see that explicit use of A can be of help in retrieving the latent cause of Y, M. Indeed, since $M \not\perp\!\!\!\perp A|X$, using A in addition to X can give information about M. In general (*e.g.* in a nonlinear setting), it is not clear that using A would be enough to ensure $\hat{Y} \perp\!\!\!\perp A$. Nevertheless, this example shows how explicit use of the sensitive attribute in a model can ensure fairness rather than lead to unfairness.

This observation is relevant to one of the simplest fairness definition, motivated by legal requirements, *Fairness through Unawareness*, which states that $\hat{Y}$ is fair as long as it does not make explicit use of the sensitive attribute A. Whilst this fairness definition is often indicated as problematic as some of the variables used to form $\hat{Y}$ could be a proxy for A (such as neighborhood for race), the example above shows a more subtle issue with this definition.

Path-Specific Population-level Measure of Unfairness

In this section, we show how the path-specific analysis of §3 can be used in complex unfairness scenarios concerning Δ.

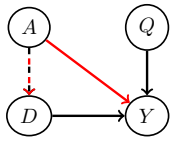


Fig. 7. CBN underlying a college admission scenario.

Let's reconsider the college admission example discussed in §1 (Fig. 7), and let's first assume that the path $A \to D \to Y$ is fair. In this case, $\text{PSE}_{\bar{a}a}$ along the direct path $A \to Y$, given by

$$\langle Y_a(D_{\bar{a}}) \rangle_{p(Y_a(D_{\bar{a}}))} - \langle Y_{\bar{a}} \rangle_{p(Y_{\bar{a}})} = p(Y_a(D_{\bar{a}})) - p(Y_{\bar{a}}) \,,$$

would provide a measure of the strength of discrimination on Y overall population. If instead $A \to D$, and therefore $A \to D \to Y$, is unfair, a measure of strength of discrimination could be given by $\text{ATE}_{\bar{a}a} = p(Y_a) - p(Y_{\bar{a}})$, which coincides with $p(Y|A = a) - p(Y|A = \bar{a})$ measuring how far Y is from being independent from A. Notice that, whilst $p(Y = 1|A = \bar{a})$ can be estimated as

$$p(Y = 1|A = \bar{a}) \approx \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{\hat{y}^n = 1, a^n = \bar{a}} \,,$$

computing the PSE requires knowledge of the CBN underlying the data. This is the price to pay to be able to compute information that travels along a subset of paths, rather than full conditioning. If the causal structure is not known or estimating its conditional distributions is challenging, the resulting PSE estimate could be imprecise.

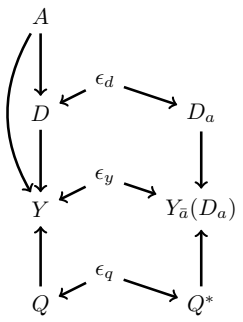
Path-Specific Individual-level Measure of Unfairness

In this section, we show that CBNs can be used to answer fairness questions for specific individuals.

In the college admission scenario of Fig. 7 in which the path $A \to D \to Y$ is considered fair, we might want to ask whether a rejected female applicant $\{a^n = a = 1, q^n, d^n, y^n = 0\}$ was treated unfairly, by asking if she would have been admitted had she been male ($A = \bar{a} = 0$) along the direct path $A \to Y$. This can be done by comparing $p(Y_{\bar{a}}(D_a)|A = a, Q = q^n, D = d^n)$ with the outcome in the actual world (corresponding to $p(Y_a(D_a)|A = a, Q = q^n, D = d^n) = \mathbb{1}_{Y_a(D_a)=y^n}$).

To understand how this can be achieved, consider the following linear model associated to a CBN with the same structure as the one in Fig. 7

$$A \sim \text{Bern}(\pi), Q = \theta^q + \epsilon_q, D = \theta^d + \theta^d_a A + \epsilon_d, Y = \theta^y + \theta^y_a A + \theta^y_q Q + \theta^y_d D + \epsilon_y \,.$$



The relationships between A, Q, D, Y and $Y_{\bar{a}}(D_a)$ in this model can be inferred from the *twin Bayesian network* [36] on the left resulting from the intervention $A = a$ along $A \to D$ and $A = \bar{a}$ along $A \to Y$: in addition to A, Q, D, Y, the network contains the variables Q^*, D_a and $Y_{\bar{a}}(D_a)$ corresponding to the counterfactual world in which $A = \bar{a}$ along $A \to Y$. The two groups of variables are connected through $\epsilon_d, \epsilon_q, \epsilon_y$, indicating that the factual and counterfactual worlds share the same unobserved randomness. From this network, we can deduce that $Y_{\bar{a}}(D_a) \perp\!\!\!\perp \{A, Q, D\}|\{\epsilon_q, \epsilon_d\}$[5], and therefore that we can express $p(Y_{\bar{a}}(D_a)|A = a, Q = q^n, D = d^n)$ as

$$p(Y_{\bar{a}}(D_a)|a, q^n, d^n) = \int_{\epsilon_q, \epsilon_d} p(Y_{\bar{a}}(D_a)|\epsilon_q, \epsilon_d, \cancel{a}, \cancel{q^n}, \cancel{d^n}) p(\epsilon_q, \epsilon_d, |a, q^n, d^n) \,. \quad (4)$$

As $\epsilon^n_q = q^n - \theta^q$, $\epsilon^n_d = d^n - \theta^d - \theta^d_a$, we obtain[6] $\langle Y_{\bar{a}}(D_a) \rangle_{p(Y_{\bar{a}}(D_a)|A=a,Q=q^n,D=d^n)} = \theta^y + \theta^y_q q^n + \theta^y_d d^n$.

Eq. (4) suggests that, in more complex scenarios (*e.g.* in which variables are not linearly related), we can obtain a Monte-Carlo estimate of $p(Y_{\bar{a}}(D_a)|a, q^n, d^n)$

[5] Notice that $Y_{\bar{a}}(D_a) \perp\!\!\!\perp A$, but $Y_{\bar{a}}(D_a) \not\perp\!\!\!\perp A|D$.

[6] Notice that $\langle Y_{\bar{a}}(D_a) \rangle_{p(Y_{\bar{a}}(D_a)|A=a,Q=q^n,D=d^n)} = \langle Y \rangle_{p(Y|A=a,Q=q^n,D=d^n)} - \text{PSE}^{\not a}_{\bar{a}a}$. Indeed $\langle Y \rangle_{p(Y|A=a,Q=q^n,D=d^n)} = \theta^y + \theta^y_a + \theta^y_q q^n + \theta^y_d d^n$ and $\text{PSE}_{\bar{a}a} = \theta^y_a$. This equivalence does not hold in the non-linear setting.

by sampling ϵ_q and ϵ_d from $p(\epsilon_q, \epsilon_d, |A = a, Q = q^n, D = d^n)$. This approach in used in [11] to introduce a general method to obtain a prediction $\hat{Y}$ of Y such that the two distributions $p(\hat{Y}_{\bar{a}}(D_a)|A = a, Q = q^n, D = d^n)$ and $p(\hat{Y}_a(D_a)|A = a, Q = q^n, D = d^n)$ coincide (*path-specific counterfactual fairness*).

Notice that $p(Y_{\bar{a}}(D_a)|A = a) = \int_{\mathcal{X}} p(Y_{\bar{a}}(D_a)|A = a, Q, D)p(Q, D|A = a)$ and when A does not have any incoming link, then $p(Y_{\bar{a}}(D_a)|A = a) = p(Y_{\bar{a}}(D_a))$ which illustrates the connection with the previous section.

5 Conclusions

We used causal Bayesian networks to provide a graphical interpretation of unfairness in a dataset as the presence of an unfair causal effect of a sensitive attribute. We used this viewpoint to revisit the recent debate surrounding the COMPAS pretrial risk assessment tool and, more generally, to point out that fairness evaluation on a model requires careful considerations on the patterns of unfairness underlying the training data. We then showed that, by being able to represent complex unfairness scenarios, causal Bayesian networks provide us with a powerful tool to measure unfairness and to design techniques for alleviating model unfairness.

Our discussion did not cover difficulties in making reasonable assumptions on the structure of the causal Bayesian network underlying a dataset, nor on the estimations of the associated conditional distributions or of other quantities of interest. These are obstacles that need to be carefully considered to avoid improper use of this framework.

Acknowledgements

The authors would like to thank Ray Jiang, Christina Heinze-Deml, Tom Stepleton, and Tom Everitt for useful discussions.

Appendix A Bayesian Networks

A *graph* is a collection of nodes and links connecting pairs of nodes. The links may be directed or undirected, giving rise to *directed* or *undirected graphs* respectively.

A *path* from node X_i to node X_j is a sequence of linked nodes starting at X_i and ending at X_j. A *directed path* is a path whose links are directed and pointing from preceding towards following nodes in the sequence.

A *directed acyclic graph* (DAG) is a directed graph with no directed paths starting and ending at the same node. For example, the directed graph in Fig. 8(a) is acyclic. The addition of a link from X_4 to X_1 gives rise to a cyclic graph (Fig. 8(b)).

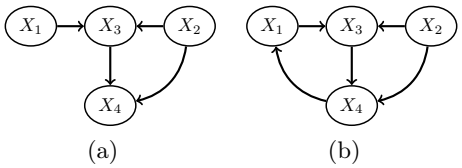


Fig. 8. Directed (a) acyclic and (b) cyclic graph.

A node X_i with a directed link to X_j is called *parent* of X_j. In this case, X_j is called *child* of X_i.

A node is a *collider* on a path if it has (at least) two parents on that path. Notice that a node can be a collider on a path and a non-collider on another path. For example, in Fig. 8(a) X_3 is a collider on the path $X_1 \to X_3 \leftarrow X_2$ and a non-collider on the path $X_2 \to X_3 \to X_4$.

A node X_i is an *ancestor* of a node X_j if there exists a directed path from X_i to X_j. In this case, X_j is a *descendant* of X_i.

A *Bayesian network* is a DAG in which nodes represent random variables and links express statistical relationships between the variables. Each node X_i in the graph is associated with the conditional distribution $p(X_i|\text{pa}(X_i))$, where $\text{pa}(X_i)$ is the set of parents of X_i. The joint distribution of all nodes, $p(X_{1:I} \equiv X_1, \ldots, X_I)$, is given by the product of all conditional distributions, *i.e.* $p(X_{1:I}) = \prod_{i=1}^{I} p(X_i|\text{pa}(X_i))$.

In a Bayesian network, the sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are statistically independent given $\mathcal{Z}$ ($\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \,|\, \mathcal{Z}$) if all paths from any element of $\mathcal{X}$ to any element of $\mathcal{Y}$ are *closed* (or *blocked*). A path is closed if at least one of the following conditions is satisfied:

(a) There is a non-collider on the path which belongs to the conditioning set $\mathcal{Z}$.
(b) There is a collider on the path such that neither the collider nor any of its descendants belong to the conditioning set $\mathcal{Z}$.

Appendix B EFPRs/EFNRs and Calibration

Assume that EFPRs and EFNRs are satisfied, *i.e.* $p(\hat{Y}=1|A=0,Y=1) = p(\hat{Y}=1|A=1,Y=1) \equiv p_{\hat{Y}_1|Y_1}$, and $p(\hat{Y}=1|A=0,Y=0) = p(\hat{Y}=1|A=1,Y=0) \equiv p_{\hat{Y}_1|Y_0}$. From

$$p(Y=1|A=0,\hat{Y}=1) = \frac{p_{\hat{Y}_1|Y_1}\overbrace{p(Y=1|A=0)}^{p_{Y_1|A_0}}}{p_{\hat{Y}_1|Y_1}p_{Y_1|A_0} + p_{\hat{Y}_1|Y_0}(1-p_{Y_1|A_0})},$$
$$p(Y=1|A=1,\hat{Y}=1) = \frac{p_{\hat{Y}_1|Y_1}p_{Y_1|A_1}}{p_{\hat{Y}_1|Y_1}p_{Y_1|A_1} + p_{\hat{Y}_1|Y_0}(1-p_{Y_1|A_1})},$$

we see that, to also satisfy $p(Y=1|A=0,\hat{Y}=1) = p(Y=1|A=1,\hat{Y}=1)$, we need $\big(\cancel{p_{\hat{Y}_1|Y_1}p_{Y_1|A_1}} + \cancel{p_{\hat{Y}_1|Y_0}}(1-\cancel{p_{Y_1|A_1}})\big)p_{Y_1|A_0} = \big(\cancel{p_{\hat{Y}_1|Y_1}p_{Y_1|A_0}} + \cancel{p_{\hat{Y}_1|Y_0}}(1-\cancel{p_{Y_1|A_0}})\big)p_{Y_1|A_1}$, *i.e.* $p_{Y_1|A_0} = p_{Y_1|A_1}$. Therefore, EFPRs/EFNRs and Calibration cannot be satisfied at the same time unless $\hat{Y} = Y$ or $Y \perp\!\!\!\perp A$.

Bibliography

[1] AI Now Institute. Litigating algorithms: Challenging government use of algorithmic decision systems, 2018.

[2] M. Alexander. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, 2012.

[3] D. A. Andrews and J. Bonta. *Level of Service Inventory – Revised*. Multi-Health Systems Toronto, 2000.

[4] D. A. Andrews, J. Bonta, and J. S. Wormith. The recent past and near future of risk and/or need assessment. *Crime & Delinquency*, 52(1):7–27, 2006.

[5] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks., May 2016. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[6] D. Arnold, W. Dobbie, and C. S. Yang. Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133:1885–1932, 2018.

[7] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

[8] M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. Technical report, Upturn, 2018.

[9] T. Brennan, W. Dieterich, and B. Ehret. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009.

[10] A. Byanjankar, M. Heikkilä, and J. Mezei. Predicting credit risk in peer-to-peer lending: A neural network approach. In *IEEE Symposium Series on Computational Intelligence*, pages 719–725, 2015.

[11] S. Chiappa. Path-specific counterfactual fairness. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

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

[13] A. Chouldechova, E. Putnam-Hornstein, D. Benavides-Prado, O. Fialko, and R. Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research*, 81:134–148, 2018.

[14] S. Corbett-Davies, E. Pierson, A. Feller, and S. Goel. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 797–806, 2017.

[15] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear., October 2016. `https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/`

17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.8c6e8c1cfbdf.

[16] S. Corbett-Davies and S.Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

[17] P. Dawid. Fundamentals of statistical causality. Technical report, University College London, 2007.

[18] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.

[19] W. Dieterich, C. Mendoza, and T. Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity, 2016.

[20] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[21] L. Eckhouse, K. Lum, C. Conti-Cook, and J. Ciccolini. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, 2018.

[22] V. Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[23] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[24] A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.". *Federal Probation*, 80(2):38–46, 2016.

[25] Harvard Law School. Note: Bail reform and risk assessment: The cautionary tale of federal sentencing. *Harvard Law Review*, 131(4):1125–1146, 2018.

[26] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pages 1–9, 2014.

[27] M. Hoffman, L. B. Kahn, and D. Li. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

[28] W. S. Isaac. Hope, hype, and fear: The promise and potential pitfalls of artificial intelligence in criminal justice. *Ohio State Journal of Criminal Law*, 15(2):543–558, 2017.

[29] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference*, pages 43:1–43:23, 2016.

[30] J. L. Koepke and D. G. Robinson. Danger ahead: Risk assessment and the future of bail reform. *Washington Law Review*, 93, 2017.
[31] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.
[32] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*, pages 4069–4079, 2017.
[33] K. Lum. Limitations of mitigating judicial bias with machine learning. *Nature Human Behaviour*, 1(7), 2017.
[34] K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.
[35] S. G. Mayson. Bias in, bias out. *Yale Law Journal*, 2019.
[36] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
[37] J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.
[38] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine learning*, 95(1):103–127, 2014.
[39] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.
[40] M. Rosenberg and R. Levinson. Trump's catch-and-detain policy snares many who call the U.S. home. `https://www.reuters.com/investigates/special-report/usa-immigration-court`, June 2018.
[41] A. D. Selbst. Disparate impact in big data policing. *Georgia Law Review*, 52:109–195, 2017.
[42] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, Prediction, and Search*. MIT Press, 2000.
[43] M. T. Stevenson. Assessing risk assessment in action. *Minnesota Law Review*, 103, 2017.
[44] M. Malekipirbazari V. and Aksakalli. Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10):4621–4631, 2015.
[45] R. Vaithianathan, T. Maloney, E. Putnam-Hornstein, and N. Jiang. Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *American Journal of Preventive Medicine*, 45(3):354–359, 2013.
[46] J. Zhang and E. Bareinboim. Fairness in decision-making – the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.