Path-Specific Counterfactual Fairness

Silvia Chiappa
csilvia@google.com
DeepMind London

Abstract

We consider the problem of learning fair decision systems from data in which a sensitive attribute might affect the decision along both fair and unfair pathways. We introduce a counterfactual approach to disregard effects along unfair pathways that does not incur in the same loss of individual-specific information as previous approaches. Our method corrects observations adversely affected by the sensitive attribute, and uses these to form a decision. We leverage recent developments in deep learning and approximate inference to develop a VAE-type method that is widely applicable to complex non-linear models.

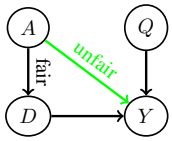
Introduction

Machine learning is increasingly being used to take decisions that can severely affect people's lives, *e.g.* in policing, education, hiring, lending, and criminal risk assessment (Hoffman, Kahn, and Li 2015; Dieterich, Mendoza, and Brennan 2016). This phenomenon has been accompanied by an increase in concern about disparate treatment caused by model errors and bias in the data.

In response to calls from governments and institutions, researchers have started to study how to ensure that learned models do no take decisions that are *unfair* with respect to *sensitive attributes* (*e.g.* race and gender) using different approaches. Among them, the causal framework (Pearl 2000; Dawid 2007; Pearl, Glymour, and Jewell 2016; Peters, Janzing, and Schölkopf 2017) offers an intuitive and powerful way of reasoning about fairness, by viewing unfairness as the presence of an unfair *causal effect* of the sensitive attribute on the decision (Qureshi et al. 2016; Bonchi et al. 2017; Kilbertus et al. 2017; Kusner et al. 2017; Russell et al. 2017; Zhang and Wu 2017; Zhang, Wu, and Wu 2017; Nabi and Shpitser 2018; Zhang and Bareinboim 2018).

Kusner et al. recently introduced a causal, individual-level, definition of fairness, called *counterfactual fairness*, which states that a decision is fair toward an individual if it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute were different. Counterfactual fairness assumes that the entire effect of the sensitive attribute on the decision is problematic. This is restrictive for scenarios in which the sensitive attribute might affect the decision along both fair and unfair pathways.

For example, in the case of Berkeley's alleged sex bias in graduate admission (Pearl 2000), female applicants were rejected more often than male applicants as they were more often applying to departments with lower admission rates. Such an effect of gender through department choice is not unfair as far as the college is concerned. What would be inadmissible is if the college treated male and female applicants with the same qualifications and applying to the same departments differently because of gender. This complex scenario can be represented by the graphical causal model depicted above. In this model, A, Q, D, and Y are random variables representing respectively gender, qualification, department choice, and admission decision, $A \rightarrow D \rightarrow Y$ is a causal path representing the influence of gender A on admission decision Y through department choice D, and $A \rightarrow Y$ is a causal path representing the direct influence of A on Y.



To deal with such scenarios, we propose a novel definition of fairness called *path-specific counterfactual fairness*, which states that a decision is fair toward an individual if it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute *along the unfair pathways* were different. In the Berkeley example, this would mean that an admission decision would be fair toward a female candidate if it would remain the same when pretending that the candidate were male along $A \rightarrow Y$.

We propose an approach that implements path-specific counterfactual fairness by *correcting* the observations corresponding to variables that are descendants of the sensitive attribute along unfair causal pathways. The correction aims at eliminating the unfair information contained in the observations while retaining fair information. Furthermore, we introduce a latent-variable method that, by leveraging recent developments in deep learning and approximate inference, allows to apply this correction approach to complex non-linear models. Our correction procedure allows to retain more individual-specific information than previous approaches to path-specific fairness based on constraining the learning of the model parameters to eliminate or reduce unfair effects (Kilbertus et al. 2017; Nabi and Shpitser 2018).

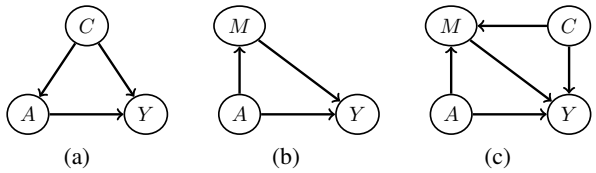


Figure 1: (a): GCM with a confounder C for the causal effect of A on Y. (b): GCM with one direct path and one indirect causal path from A to Y. (c): GCM with a confounder C for the causal effect of M on Y.

Background on Causality

Causal relationships among random variables can visually be expressed using *graphical causal models* (GCMs). A GCM is a special case of a graphical model (see Chiappa for a quick introduction) that captures both independence and causal relations. In this work, we restrict ourselves to *directed acyclic graphs*, *i.e.* graphs in which a node cannot be an *ancestor* of itself. In a directed acyclic graph, the joint distribution over all nodes $p(X_1, \ldots, X_I)$ is given by the product of the conditional distributions of each node X_i given its *parents* $\text{pa}(X_i)$, *i.e.* $p(X_1, \ldots, X_I) = \prod_{i=1}^{I} p(X_i|\text{pa}(X_i))$.

GCMs 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*. The causal effect of A on Y can be seen as the information that A sends to Y through causal paths, or as the conditional distribution of Y given A restricted to causal paths.

This implies that if there exist at least one *open* non-causal 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 GCM $\mathcal{G}$ of Fig. 1(a): the variable C is said to be a *confounder* for the effect of A on Y. In this case, the causal effect of $A = a$ on Y can be seen as the conditional distribution $p_{\rightarrow A=a}(Y|A = a)$ on the modified GCM $\mathcal{G}_{\rightarrow A=a}$, resulting from *intervening* on A by replacing $p(A|C)$ with a 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.

The rules of do-calculus (Pearl 2000; Pearl, Glymour, and Jewell 2016) indicate if and how the conditional distribution in the intervened graph can be estimated using observations from $\mathcal{G}$: if C is observed $p_{\rightarrow A=a}(Y|A = a) = \sum_C p(Y|A = a, C)p(C)$, whilst if C is unobserved estimating the conditional distribution using only observations from $\mathcal{G}$ is not possible – in this case the effect is said to be *non-identifiable*.

We define $Y_{A=a}$ to be the random variable with distribution $p(Y_{A=a}) = p_{\rightarrow A=a}(Y|A = a)$. $Y_{A=a}$ is called *potential outcome* variable and we will refer to it with the shorthand Y_a.

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. The simplest cases are the isolation of the contributions along the direct path $A \rightarrow Y$ (*direct effect*) and along the indirect causal paths $A \rightarrow \ldots \rightarrow Y$ (*indirect effect*).

Suppose that the GCM contains only one indirect causal path through a variable M, as in Fig. 1(b). We define $Y_a(M_{a'})$ to be the random variable that results from the interventions $A = a$ along $A \rightarrow Y$ and $A = a'$ along $A \rightarrow M \rightarrow Y$.

The *average direct effect* (ADE) and the *average indirect effect* (AIE) of $A = a$ with respect to $A = a'$ are given by[1]

$$\text{ADE} = \langle Y_a(M_{a'})\rangle - \langle Y_{a'}\rangle, \;\; \text{AIE} = \langle Y_{a'}(M_a)\rangle - \langle Y_{a'}\rangle\,,$$

where, *e.g.*, $\langle Y_a\rangle = \int_{Y_a} Y_a p(Y_a)$.

More generally, the ADE of $A = a$ with respect to $A = a'$ can be estimated by computing the difference between 1) the average effect of $A = a$ along the direct path $A \rightarrow Y$ and $A = a'$ along the indirect causal paths $A \rightarrow \ldots \rightarrow Y$ and 2) the average effect of $A = a'$ along all causal paths.

Similarly, the AIE of $A = a$ with respect to $A = a'$ can be estimated by computing the difference between 1) the average effect of $A = a'$ along the direct path $A \rightarrow Y$ and $A = a$ along the indirect causal paths $A \rightarrow \ldots \rightarrow Y$ and 2) the average effect of $A = a'$ along all causal paths.

Under the independence assumption $Y_{a,m} \perp\!\!\!\perp M_{a'}$ (*sequential ignorability*), $p(Y_a(M_{a'}))$ can be estimated as

$$\begin{aligned} p(Y_a(M_{a'})) &= \int_m p(Y_a(M_{a'})|M_{a'} = m)p(M_{a'} = m) \\ &= \int_m p(Y_{a,m}|M_{a'} = m)p(M_{a'} = m) \\ &= \int_m p(Y_{a,m})p(M_{a'} = m)\,, \end{aligned} \tag{1}$$

where to obtain the second line we have used the *consistency* property (Pearl, Glymour, and Jewell 2016). As there are no confounders, intervening coincides with conditioning, *i.e.* $p(Y_{a,m}) = p(Y|A = a, M = m)$ and $p(M_{a'}) = p(M|A = a')$.

If the GCM contains a confounder for the effect of either A or M on Y, such as C in Fig. 1(c), then $p(Y_{a,m}) \neq p(Y|A = a, M = m)$. In this case, by following similar arguments as the ones used in Eq. (1) but conditioning on C (and therefore assuming $Y_{a,m} \perp\!\!\!\perp M_{a'}|C$), we obtain[2]

$$p(Y_a(M_{a'})) = \int_{m,c} p(Y|a, m, c)p(m|a', c)p(c)\,.$$

If C is unobserved, the effect is non-identifiable.

Path-Specific Effect. In the more complex case in which, rather than computing the direct and indirect effects, we want to isolate the contribution of the effect along a specific group of paths, we can generalize the formulas for the ADE

[1] In this paper, we consider the *natural* effect, which generally differs from the *controlled* effect; the latter corresponds to intervening on M.

[2] We use the notation $p(Y|a, m, c)$ as a shorthand for $p(Y|A = a, M = m, C = c)$.

and AIE by using in the first term the variable resulting from performing the intervention $A = a$ along the group of interest and $A = a'$ along the remaining causal paths.

For example, consider the GCM of Fig. 2 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 green and dashed green-black links. The *path-specific effect* (PSE) of $A = a$ with respect to $A = a'$ for this group of paths is given by

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

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

$$\int_{c,m,l} p(Y|a, c, m, l)p(l|a', c, m)p(m|a, c)p(c) .$$

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

$$\begin{aligned} &A \sim \text{Bernoulli}(\pi),\ C = \epsilon_c , \\ &M = \theta^m + \theta^m_a A + \theta^m_c C + \epsilon_m, \\ &L = \theta^l + \theta^l_a A + \theta^l_c C + \theta^l_m M + \epsilon_l , \\ &Y = \theta^y + \theta^y_a A + \theta^y_c C + \theta^y_m M + \theta^y_l L + \epsilon_y , \end{aligned} \tag{2}$$

where ϵ_c, ϵ_m, ϵ_l and ϵ_y are unobserved independent zero-mean Gaussian terms, we have

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

The PSE is therefore given by

$$\theta^y_a(a - a') + \theta^y_m \theta^m_a (a - a') + \theta^y_l \theta^l_m \theta^m_a (a - a') . \tag{3}$$

Shpitser gives a recursive rule for obtaining the variable of interest for computing the PSE, and a graphical method for understanding whether the PSE is identifiable in the presence of unobserved confounders.

Path-Specific Counterfactual Fairness

We are interested in complex scenarios in which the sensitive attribute A might affect the decision variable Y along both fair and unfair causal pathways. We assume that A can only take two values a and a', and that a' is a baseline value.

Kilbertus et al. and Nabi and Shpitser propose to deal with such scenarios by constraining the learning of the model parameters such that the average of the unfair effect is eliminated or reduced. More specifically, Nabi and Shpitser suggest to perform model training by constraining the unfair PSE of A on Y to lie in a small range. The main limitation of this approach is that, at test time, it requires averaging over all variables that are descendants of the sensitive attribute through the unfair causal pathways. This can negatively impact the system's predictive accuracy, as individual-specific information about those descendants is disregarded. Kilbertus et al. propose to directly identify a set of constraints on the conditional distribution of the decision variable that eliminate the unfair effect. This can easily be done in linear models, but it is unclear how to identify the constraints in more complex non-linear scenarios. Furthermore, this approach also unnecessarily removes information from problematic descendants.

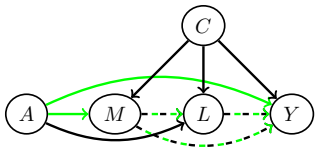


Figure 2: GCM corresponding to Eq. (2).

In contrast, we propose to simply *correct* at test time the decisions of individuals for which $A = a$ by making sure that they coincide with the one that would have been taken in a counterfactual world in which the sensitive attribute along the unfair pathways were set to the baseline. This requires correcting the observations corresponding to variables that are descendants of the sensitive attribute through unfair pathways, by removing the unfair information induced by the sensitive attribute while retaining the remaining fair information. We achieve this through a generalization of the abduction-action-prediction method for counterfactual reasoning (Pearl 2000). We generally refer to our approach as *path-specific counterfactual fairness* (PSCF). For the Berkeley alleged sex bias case for example, PSCF would ensure that the admission decision of a female applicant coincides with the one that would have been taken in a counterfactual world in which her gender a were male a' along the direct path $A \to Y$, by taking a decision based on the intervention $A = a'$ along $A \to Y$.

To highlight its relation with the approaches of Kilbertus et al. and of Nabi and Shpitser, we first explain PSCF for the case in which the data-generation mechanism is given by the linear model of Eq. (2) (Fig. 2). Assume that the direct effect of A on Y and the effect through M are considered unfair. PSCF corrects the decision of an individual for which $A = a$ by performing the intervention $A = a'$ along the direct path $A \to Y$ and the paths passing through M, $A \to M \to, \ldots, \to Y$, namely along the green and dashed green-black links of Fig. 2. (Notice that the dashed green-black links differ fundamentally from the green links; they contain unfairness only as a consequence of $A \to M$, corresponding to the parameter θ^m_a, being unfair.) More precisely, assuming that $a' = 0$ is the baseline value of A, given an instance $\{a^n = a = 1, c^n, m^n, l^n\}$, the PSCF approach computes a fair prediction y^n_{PSCF} of y^n as the mean of $p(Y_{a'}(M_{a'}, L_a(M_{a'}))|a, c^n, m^n, l^n)$. This is achieved by first computing ϵ^n_m and ϵ^n_l from a^n, c^n, m^n, l^n and the model equations (*abduction*), *i.e.*

$$\begin{aligned} \epsilon^n_m &= m^n - \theta^m - \theta^m_a - \theta^m_c c^n , \\ \epsilon^n_l &= \theta^l - \theta^l_a - \theta^l_c c^n - \theta^l_m m^n . \end{aligned}$$

Then fair transformations of m^n and l^n, m^n_{PSCF} and l^n_{PSCF}, and the fair prediction y^n_{PSCF} are obtained by substituting ϵ^n_m and ϵ^n_l into the model equations with the problematic terms θ^m_a and θ^l_a removed (this corresponds to the intervention $A = a'$ along the direct path $A \to Y$ and the paths passing through

$M, A \to M \to, \ldots, \to Y$), *i.e.*

$$
\begin{aligned}
m^n_{\text{PSCF}} &= \theta^m + \cancel{\theta^m_a} + \theta^m_c c^n + \epsilon^n_m \,, \\
l^n_{\text{PSCF}} &= \theta^l + \theta^l_a + \theta^l_c c^n + \theta^l_m m^n_{\text{PSCF}} + \epsilon^n_l \,, \\
y^n_{\text{PSCF}} &= \theta^y + \cancel{\theta^y_a} + \theta^y_c c^n + \theta^y_m m^n_{\text{PSCF}} + \theta^y_l l^n_{\text{PSCF}} \,. \qquad (4)
\end{aligned}
$$

This approach can be seen as performing a correction on the decision through a correction on all the variables that are descendants of the sensitive attribute along unfair pathways (UP), namely M and L in this case.

To understand the relation with the *fair inference on outcomes* (FIO) method suggested by Nabi and Shpitser, the PSE for this model (Eq. (3)) with $a = 1$ and $a' = 0$ takes the form

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

FIO consists in performing a constrained learning of the model parameters θ such that the PSE lies in a small range. After training, a prediction y^n_{FIO} of y^n for an instance $\{a^n, c^n, m^n, l^n\}$ can be obtained as $y^n_{\text{FIO}} = \langle Y \rangle_{p(Y|a^n, c^n)}$, where $p(Y|a^n, c^n)$ is given by

$$\int_{m,l} p(Y|a^n, c^n, m, l) p(l|a^n, c^n, m) p(m|a^n, c^n) \,,$$

i.e. as

$$
\begin{aligned}
m^n_{\text{FIO}} &= \hat{\theta}^m + \hat{\theta}^m_a a^n + \hat{\theta}^m_c c^n \,, \\
l^n_{\text{FIO}} &= \hat{\theta}^l + \hat{\theta}^l_a a^n + \hat{\theta}^l_c c^n + \hat{\theta}^l_m m^n_{\text{FIO}} \,, \\
y^n_{\text{FIO}} &= \hat{\theta}^y + \hat{\theta}^y_a a^n + \hat{\theta}^y_c c^n + \hat{\theta}^y_m m^n_{\text{FIO}} + \hat{\theta}^y_l l^n_{\text{FIO}} \,,
\end{aligned}
$$

where $\hat{\theta}$ indicate the learned model parameters.

Assume that, at the end of training, $\hat{\theta}$ for both PSCF and FIO coincide with the true underlying parameters θ, except for $\hat{\theta}^m_a$ and $\hat{\theta}^y_a$ in FIO which are assigned zero values to satisfy the constraint PSE $= 0$. Then, given an instance $\{a^n = a = 1, c^n, m^n, l^n\}$, we can express y^n_{PSCF} as $y^n_{\text{PSCF}} = \langle Y \rangle_{p(Y|a^n, c^n, m^n, l^n)} - \text{PSE}$, since

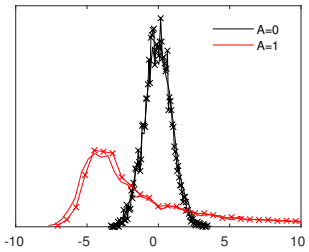
$$y^n_{\text{PSCF}} = \theta^y + \theta^y_c c^n + \theta^y_m m^n + \theta^y_l l^n - \theta^m_a(\theta^y_m + \theta^y_l \theta^l_m) \,;$$

and y^n_{FIO} as $y^n_{\text{FIO}} = \langle Y \rangle_{p(Y|a^n, c^n)} - \text{PSE}$, since

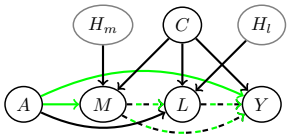
$$y^n_{\text{FIO}} = \theta^y + \theta^y_c c^n + \theta^y_m \bar{m}^n + \theta^y_l \bar{l}^n - \theta^m_a(\theta^y_m + \theta^y_l \theta^l_m) \,,$$

where $\bar{m}^n = \langle M \rangle_{p(M|a^n, c^n)} = \theta^m + \theta^m_a + \theta^m_c c^n$. This formulation highlights the disadvantage of FIO over PSCF in disregarding specific information about the individual, ϵ^n_m and ϵ^n_l, through the use of $\bar{m}^n$ and $\bar{l}^n$. As the constraint PSE $= 0$ is not necessarily achieved by assigning zero values to $\hat{\theta}^m_a$ and $\hat{\theta}^y_a$, this correspondence does not generally hold.

As the reason for averaging over M and L, Nabi and Shpitser indicate the need to account for the constraints that are potentially imposed on $\hat{\theta}^m_a$ and $\hat{\theta}^l_m$. If a constraint is imposed on a parameter, then the corresponding variable needs indeed to be integrated out to ensure that such a constraint is taken into account in the prediction. For any model, the PSE would contain the parameters corresponding to the UP descendants of A, which means that FIO would always require integrating out the UP descendants. However, even if we a priori identify a set of constraints that give PSE $= 0$, the UP descendants must be integrated out or corrected from unfairness even if no constraints are imposed on the corresponding parameters. Consider the case discussed above, where we achieve PSE $= 0$ by setting $\hat{\theta}^m_a$ and $\hat{\theta}^y_a$ to zero values. This does not constrain $\hat{\theta}^l_m$. However, to form a prediction of y^n, we would still need to integrate over L, as the observation l^n contains the problematic term $\theta^y_l \theta^l_m \theta^m_a$, corresponding to the unfair part of the effect of A on L.



(a)



(b)

Figure 3: (a): Empirical distribution of the estimate of ϵ^n_m for the case in which m^n is generated by Eq. (2) with an extra non-linear term $f(A, C)$ (continuous lines). Histograms of $\tilde{p}(H_m|A)$ (crossed lines). (b): GCM with an explicit latent variable for each UP descendant of A.

In this simple case, we could avoid having to integrate over M and L by a priori imposing the constraints $\hat{\theta}^y_a = 0$ and $\hat{\theta}^y_m = -\hat{\theta}^y_l \hat{\theta}^l_m$, *i.e.* by constraining the conditional distribution used to form a prediction of y^n, $p(Y|A, C, M, L)$. This coincides with the constraint proposed by Kilbertus et al. to avoid proxy discrimination. However, this approach achieves removal of the problematic unfairness in m^n and l^n by cancelling out the entire m^n from the prediction. This is also suboptimal, as all information within m^n is disregarded. Furthermore, it is not clear how to extend this approach to more complex scenarios.

In conclusion, the main advantage of our approach is that it allows to retain fair individual-specific information contained in the UP descendants. This is achieved by leaving unaltered the underlying data-generation mechanism during training.

Model-Observations Mismatch. Whilst offering several advantages over previous approaches to path-specific fairness, in the presence of a strong mismatch between the assumed and actual data-generation mechanisms, the PSCF approach described above would most likely not remove unfairness completely.

Indeed, in this case the estimates of ϵ^n_m and ϵ^n_l would not be independent from the sensitive attribute A. Consider, for example, the case in which we assume the data-generation process of Eq. (2), but the observed m^n, $n = 1, \ldots, N$, are generated from a modified version of Eq. (2) containing an extra non-linear term $f(A, C)$. The learned model parameters $\hat{\theta}$ would not be able to describe this non-linear term, which would therefore be absorbed into the estimate of ϵ^n_m, making

it dependent on A, as shown in Fig. 3(a) (continuous lines).

To solve this issue, we propose to decompose ϵ_m into two components, *i.e.* $\epsilon_m = H_m + \eta_m$, and to adopt a training procedure in which $\tilde{p}(H_m|A=a)$, defined as

$$\tilde{p}(H_m|A=a) = \frac{1}{N_a}\sum_{n=1}^{N_a} p(H_m|a^n=a, c^n, m^n, l^n) \quad (5)$$

where N_a indicates the number of observations for which $a^n = a$, is encouraged to have small dependence on A. We can then use, *e.g.*, the mean of $p(H_m|a^n, c^n, m^n, l^n)$, rather than the estimate of ϵ_m^n. In other words, we make sure that, when estimating the latent randomness associated with an individual, we only pick up the part that does not depend on A, and only use this part to perform the prediction.

Encouraging independence on A is necessary, as otherwise the estimated $\tilde{p}(H_m^n|A)$ would be close to the estimate of ϵ_m^n. This is shown by the histograms of $\tilde{p}(H_m|A)$ in Fig. 3(a) (crossed lines), obtained by assuming a Gaussian distribution for $p(H_m)$ and by learning the model parameters using an expectation maximization approach.

To more generally ensure that the abduction procedure will not end up with estimates that depend on the sensitive variable, we need to encourage latent independence on A for each descendant of A that needs to be corrected, namely for each UP descendant, and therefore introduce another latent variable for L, H_l (see Fig. 3(b)).

We propose a way to encourage independence on A together with a method that generalizes the PSCF approach described above to complex non-linear models in the next section.

PSCF-VAE

Consider more general equations for the GCM of Fig. 3(b), given by

$$\begin{aligned}
&A \sim \text{Bernoulli}(\pi),\ C \sim p_\theta(C)\,,\\
&H_m \sim p_\theta(H_m),\ M \sim p_\theta(M|A,C,H_m)\,,\\
&H_l \sim p_\theta(H_l),\ L \sim p_\theta(L|A,C,M,H_l)\,,\\
&Y \sim p_\theta(Y|A,C,M,L)\,, \qquad (6)
\end{aligned}$$

where if M is categorical we assume $p_\theta(M|A,C,H_m) = f_\theta(A,C,H_m)$, where $f_\theta(A,C,H_m)$ can be any function (*e.g.* a neural network); whilst if M is continuous we assume that $p_\theta(M|A,C,H_m)$ is Gaussian with mean $f_\theta(A,C,H_m)$.

The model likelihood $p_\theta(A,C,M,L,Y)$, and the posterior distributions $p_\theta(H_m|A,C,M,L)$ and $p_\theta(H_l|A,C,M,L)$ required to form fair predictions, are generally intractable. We address this issue with a variational approach that computes Gaussian approximations $q_\phi(H_m|A,C,L,M)$ and $q_\phi(H_l|A,C,L,M)$ of $p_\theta(H_m|A,C,M,L)$ and $p_\theta(H_l|A,C,M,L)$ respectively, parametrized by ϕ, as discussed in detail below.

After learning θ and ϕ, analogously to Eq. (4), we compute a fair prediction y^n_{PSCF} for an instance $\{a^n=a, c^n, m^n, l^n\}$ as $\langle Y_{a'}(M_{a'}, L_a(M_{a'}))\rangle_{p(Y_{a'}(M_{a'},L_a(M_{a'}))|a,c^n,m^n,l^n)}$, estimated using a Monte-Carlo approach. Specifically, we first draw samples $h_m^{n,i} \sim q_\phi(H_m|a,c^n,m^n,l^n)$ and $h_l^{n,i} \sim q_\phi(H_l|a,c^n,m^n,l^n)$, for $i=1,\dots,I$, and then form

$$\begin{aligned}
&m^{n,i}_{\text{PSCF}} \sim p_\theta(M|a', c^n, h_m^{n,i}),\\
&l^{n,i}_{\text{PSCF}} \sim p_\theta(L|a, c^n, m^{n,i}_{\text{PSCF}}, h_l^{n,i})\,,\\
&y^n_{\text{PSCF}} = \frac{1}{I}\sum_{i=1}^{I}\langle Y\rangle_{p_\theta(Y|a',c^n,m^{n,i}_{\text{PSCF}},l^{n,i}_{\text{PSCF}})}\,. \qquad (7)
\end{aligned}$$

In the experiments, we used $I = 500$.

If we group the observed and latent variables as $V = \{A,C,M,L,Y\}$ and $H = \{H_m, H_l\}$ respectively, the variational approximation $q_\phi(H|V)$ to the intractable posterior $p_\theta(H|V)$ is obtained by finding the variational parameters ϕ that minimize the Kullback-Leibler divergence $\text{KL}(q_\phi(H|V)||p_\theta(H|V))$. This is equivalent to maximizing a lower bound $\mathcal{F}_{\theta,\phi}$ on the log of the marginal likelihood $\log p_\theta(V) \geq \mathcal{F}_{\theta,\phi}$ with

$$\mathcal{F}_{\theta,\phi} = -\langle \log q_\phi(H|V)\rangle_{q_\phi(H|V)} + \langle \log p_\theta(V,H)\rangle_{q_\phi(H|V)}\,,$$

where, *e.g.*,

$$\langle \log q_\phi(H|V)\rangle_{q_\phi(H|V)} = \int_H q_\phi(H|V)\log q_\phi(H|V).$$

In our case, rather than $q_\phi(H|V)$, we use $q_\phi(H|V^* \equiv V\backslash Y)$. Our approach is therefore to learn simultaneously the latent embedding and predictive distributions in Eq. (7). This could be preferable to other causal latent variable approaches such as the FairLearning algorithm proposed by Kusner et al., which separately learns a predictor of Y using samples from the previously inferred latent variables and from the non-descendants of A.

In order for $\mathcal{F}_{\theta,\phi}$ to be tractable conjugacy is required, which heavily restricts the family of models that can be used. This issue can be addressed with a Monte-Carlo approximation known as variational auto-encoding (VAE) (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014). This approach represents H as a non-linear transformation $H = f_\phi(\mathcal{E})$ of a random variable $\mathcal{E}$ from a parameter free distribution q_ϵ. As we choose q to be Gaussian, $H = \mu_\phi + \sigma_\phi\mathcal{E}$ with $q_\epsilon = \mathcal{N}(0,1)$ for the univariate case. This enables us to rewrite the bound as

$$\mathcal{F}_{\theta,\phi} = -\langle \log q_\phi(H = f_\phi(\mathcal{E})) + \log p_\theta(V, H = f_\phi(\mathcal{E}))\rangle_{q_\epsilon}\,.$$

The first part of the gradient of $\mathcal{F}_{\theta,\phi}$ with respect to ϕ, $\nabla_\phi\mathcal{F}_{\theta,\phi}$, can be computed analytically, whilst the second part is approximated by

$$\langle \nabla_\phi \log p_\theta(V, H = f_\phi(\mathcal{E}))\rangle_{q_\epsilon} \approx \frac{1}{I}\sum_{i=1}^{I}\nabla_\phi \log p_\theta(V, h^i = f_\phi(\epsilon^i)),\ \epsilon^i \sim q_\epsilon\,.$$

In the experiments, we used $I = 1$, as commonly done in the VAE literature. The variational parameters ϕ are parametrized by a neural network taking as input V^*.

Independence on A. In order to ensure that $\tilde{q}_\phi(H|A)$, defined similarly to Eq. (5), does not depend on A, we experimented with an adversary approach (Edwards and Storkey

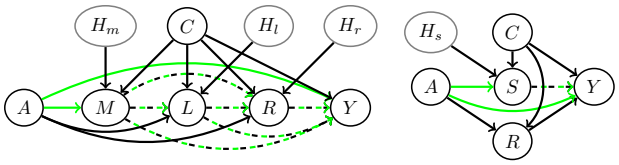


Figure 4: (a): GCM for the UCI Adult dataset. (b): GCM for the UCI German Credit dataset.

2016) and with a *maximum mean discrepancy* (MMD) penalization approach (Gretton et al. 2012; Louizos et al. 2016), which gave similar but more stable results. The MMD approach adds a penalty term to the bound $\mathcal{F}_{\theta,\phi}$,

$$-\beta \mathcal{L}_{\text{MMD}}(a, a') \,,$$

where β is a weighting factor that determines the degree of independence, and therefore might correspond to different levels of fairness. $\mathcal{L}_{\text{MMD}}(a, a')$ is the sum of several terms, one for each latent variable, where *e.g.* the term for H_m is given by

$$\mathcal{L}^m_{\text{MMD}}(a, a') = \frac{1}{N_a^2} \sum_{i=1}^{N_a} \sum_{j=1}^{N_a} k(h_m^{a,i}, h_m^{a,j})$$

$$+ \frac{1}{N_{a'}^2} \sum_{i=1}^{N_{a'}} \sum_{j=1}^{N_{a'}} k(h_m^{a',i}, h_m^{a',j}) - \frac{2}{N_a N_{a'}} \sum_{i=1}^{N_a} \sum_{j=1}^{N_{a'}} k(h_m^{a,i}, h_m^{a',j}) \,,$$

where k is a Gaussian kernel, and $h_m^{a,i}$ is a sample from the variational distribution for an individual for which $A = a$.

Experiments

We evaluate the proposed PSCF-VAE method on the UCI Adult and German Credit datasets.

As prior distribution p_θ for each latent variable (Eq. (6)) we used a ten-dimensional Gaussian with diagonal covariance matrix, whilst as f_θ we used a neural network with one linear layer of size 100 with tanh activation, followed by a linear layer (the outputs were Gaussian means for continuous variables and logits for categorical variables). As variational distribution q_ϕ we used a ten-dimensional Gaussian with diagonal covariance, with means and log variances obtained as the outputs of a neural network with two linear layers of size 20 and tanh activation, followed by a linear layer. Training was achieved with the Adam optimizer (Kingma and Ba 2015) with learning rate 0.01, mini-batch size 128, and default values $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e\text{-}8$. Training was stopped after 20,000 steps.

The UCI Adult Dataset

The Adult dataset from the UCI repository (Lichman 2013) contains 14 attributes including age, working class, education level, marital status, occupation, relationship, race, gender, capital gain and loss, working hours, and nationality for 48,842 individuals; 32,561 and 16,281 for the training and test sets respectively. The goal is to predict whether the individual's annual income is above or below \$50,000. We assumed the GCM of Fig. 4(a) (following Nabi and Shpitser),

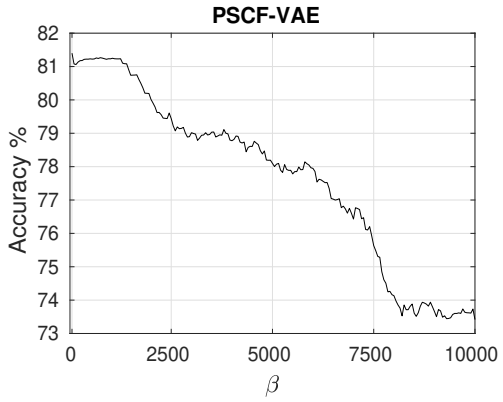


Figure 5: Test accuracy of PSCF-VAE on the UCI Adult dataset for increasing values of β.

where A corresponds to the protected attribute sex, C to the duple age and nationality, M to marital status, L to level of education, R to the triple working class, occupation, and hours per week, and Y to the income class[3]. Age, level of education and hours per week are continuous, whilst sex, nationality, marital status, working class, occupation, and income are categorical. Besides the direct effect $A \to Y$, the effect of A on Y through marital status, namely along the paths $A \to M \to, \ldots, \to Y$, is considered unfair.

Nabi and Shpitser assume that all variables are continuous, except A and Y, and linearly related, except Y for which $p(Y = 1|\text{pa}(Y)) = \pi = \sigma(\theta^y + \sum_{X_i \in \text{pa}(Y)} \theta^y_{x_i} X_i)$ where $\sigma(\cdot)$ is the sigmoid function. With the encoding $A \in \{0, 1\}$, where 0 indicates the male baseline value, and under the approximation $\log(\pi/(1-\pi)) \approx \log \pi$, we can write the PSE in the odds ratio scale as PSE $\approx \exp(\theta^y_a + \theta^y_m \theta^m_a + \theta^y_l \theta^l_m \theta^m_a + \theta^y_r (\theta^r_m \theta^m_a + \theta^r_l \theta^l_m \theta^m_a))$. An instance from the test set $\{a^n, c^n, m^n, l^n, r^n\}$ is classified by using $p(Y|a^n, c^n) = \int_{m,l,r} p(Y|a^n, c^n, m, l, r) \times p(r|a^n, c^n, m, l) p(l|a^n, c^n, m) p(m|a^n, c^n)$.

In Fig. 5, we show the accuracy obtained by PSCF-VAE on the test set for increasing values of β, ranging from $\beta = 0$ (no MMD penalization) to $\beta = 10,000$. As we can see, accuracy decreases from 81.2% to 73.4%. Notice that predictions were formed using samples of H_m, H_l and H_r also for males, even if not required. Also notice that forming predictions from $p_\theta(Y|a^n, c^n, m^n, l^n, r^n)$ gives 82.7% accuracy.

In Fig. 6, we show histograms of two dimensions of $\tilde{q}_\phi(H_m|A)$ (first and second row) and one dimension of $\tilde{q}_\phi(H_l|A)$ (third row) for $\beta = 0$, $\beta = 2,500$, and $\beta = 5,000$ (left to right) after 20,000 training steps for females (red) and males (blue) – these are the only variables that show differences between male and females. As we can see, increasing β induces a reduction in the number of modes in the posterior, which corresponds to information loss. For $\beta = 10,000$ all histograms are unimodal (not shown). For $\beta = 5,000$, for

[3]We omit race, and capital gain and loss (although including capital gain and loss would increase test accuracy from 82.7% to 84.7%) to use the same attributes as Nabi and Shpitser.

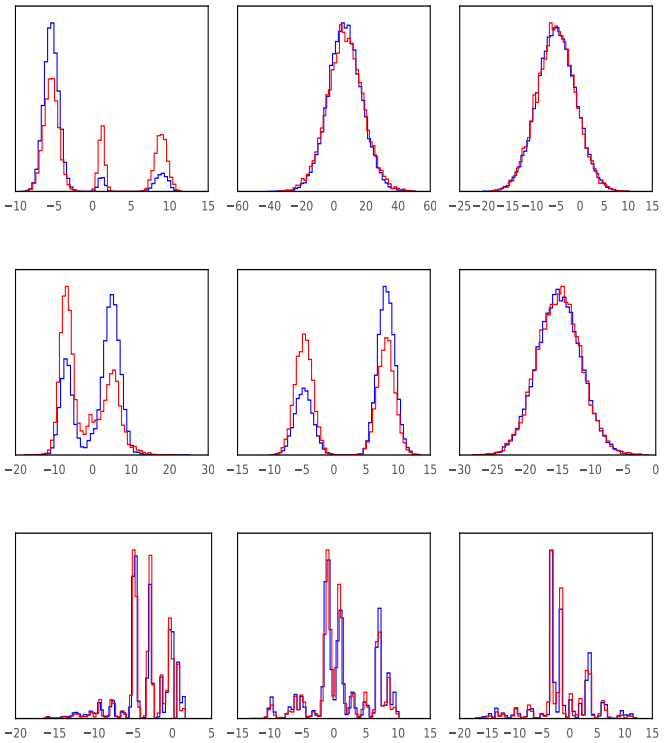


Figure 6: Histograms of two dimensions of $\tilde{q}_\phi(H_m|A)$ (first and second row) and one dimension of $\tilde{q}_\phi(H_l|A)$ (third row) for $\beta = 0$, $\beta = 2,500$, and $\beta = 5,000$ (left to right) after 20,000 training steps for females (red) and males (blue).

which accuracy is around 78%, the histograms for females and males are similar – this can therefore be considered a fair accuracy.

The unconstrained PSE on this dataset is 3.64. When constraining the PSE to be smaller than 3.7 (thus essentially imposing no constraint), FIO gives 73.8% accuracy, due the information that is lost by integrating out M, L and R. Constraining the PSE to be smaller than 3.6 also gives 73.8% accuracy. Constraining the PSE to be smaller than 1.05, as suggested by Nabi and Shpitser, gives 73.4% accuracy (Nabi and Shpitser report 72%). These results demonstrate that loss in accuracy in FIO is due to integrating out M, L and R, rather than to ensuring fairness.

The UCI German Credit Dataset

The German Credit dataset from the UCI repository contains 20 attributes of 1,000 individuals applying for loans. Each applicant is classified as a good or bad credit risk, *i.e.* as likely or not likely to repay the loan. We assume the GCM in Fig. 4(b), where A corresponds to the protected attribute sex, C to age, S to the triple status of checking account, savings, and housing, and R the duple credit amount and repayment duration. The attributes age, credit amount, and repayment duration are continuous, whilst checking account, savings, and housing are categorical. Besides the direct effect $A \to Y$, we would like to remove the effect of A on Y through S. We

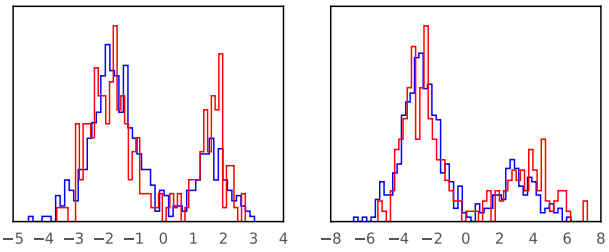


Figure 7: Histograms of $\tilde{q}_\phi(H_s|A)$ for one dimension of the variable housing for $\beta = 0$ and $\beta = 10,000$ after 20,000 training steps for females (red) and males (blue).

only need to introduce a hidden variable H_s for S, as R does not need to be corrected.

We divided the dataset into training and test sets of sizes 700 and 300 respectively. As for the Adult dataset, we varied β from 0 to 10,000. The test accuracy remained 76.0% for all values of β (predictions were formed using samples of H_s also for males). This is same accuracy obtained when forming predictions from $p_\theta(Y|a^n, c^n, s^n, r^n)$.

In Fig. 7, we show $\tilde{q}_\phi(H_s|A)$ for one dimension of the variable housing, which shows the most significant difference between females and males, for $\beta = 0$ and $\beta = 10,000$.

Conclusions

We have proposed a novel intuitive definition of fairness, path-specific counterfactual fairness, which states that a decision is fair toward an individual if it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute along the unfair pathways were different.

We have introduced a latent inference-projection method, PSCF-VAE, that achieves path-specific counterfactual fairness by correcting the variables that are descendants of the sensitive attribute along unfair pathways during testing, leaving unaltered the underlying data-generation mechanism during training. The proposed method is widely applicable to complex non-linear models.

PSCF-VAE requires providing the causal model underlying the data generation process. Future work will consider relaxing this requirement.

Acknowledgements

The author would like to thank Thomas P. S. Gillam for his contribution to this work.

References

[Bonchi et al.] Bonchi, F.; Hajian, S.; Mishra, B.; and Ramazzotti, D. 2017. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics* 3(1):1–21.

[Chiappa] Chiappa, S. 2014. Explicit-duration Markov switching models. *Foundations and Trends in Machine Learning* 7(6):803–886.

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

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

[Edwards and Storkey] Edwards, H., and Storkey, A. 2016. Censoring representations with an adversary. In *4th International Conference on Learning Representations*.

[Gretton et al.] Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13:723–773.

[Hoffman, Kahn, and Li] Hoffman, M.; Kahn, L.; and Li, D. 2015. Discretion in hiring.

[Kilbertus et al.] Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 30*, 656–666.

[Kingma and Ba] Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

[Kingma and Welling] Kingma, D., and Welling, M. 2014. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations*.

[Kusner et al.] Kusner, M.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30*, 4069–4079.

[Lichman] Lichman, M. 2013. UCI machine learning repository.

[Louizos et al.] Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2016. The variational fair autoencoder. In *4th International Conference on Learning Representations*.

[Nabi and Shpitser] Nabi, R., and Shpitser, I. 2018. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Pearl, Glymour, and Jewell] Pearl, J.; Glymour, M.; and Jewell, N. 2016. *Causal Inference in Statistics: A Primer*. Wiley.

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

[Peters, Janzing, and Schölkopf] Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

[Qureshi et al.] Qureshi, B.; Kamiran, F.; Karim, A.; and Ruggieri, S. 2016. Causal discrimination discovery through propensity score analysis. *ArXiv e-prints*.

[Rezende, Mohamed, and Wierstra] Rezende, D.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, 1278–1286.

[Russell et al.] Russell, C.; Kusner, M.; Loftus, J.; and Silva, R. 2017. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems 30*, 6417–6426.

[Shpitser] Shpitser, I. 2013. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science* 37(6):1011–1035.

[Zhang and Bareinboim] Zhang, J., and Bareinboim, E. 2018. Fairness in decision-making – the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[Zhang and Wu] Zhang, L., and Wu, X. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* 1–16.

[Zhang, Wu, and Wu] Zhang, L.; Wu, Y.; and Wu, X. 2017. A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI*, 3929–3935.