Counterfactual Fairness

Matt Kusner *
The Alan Turing Institute and
University of Warwick
mkusner@turing.ac.uk

Joshua Loftus *
New York University
loftus@nyu.edu

Chris Russell *
The Alan Turing Institute and
University of Surrey
crussell@turing.ac.uk

Ricardo Silva
The Alan Turing Institute and
University College London
ricardo@stats.ucl.ac.uk

Abstract

Machine learning can impact people with legal or ethical consequences when it is used to automate decisions in areas such as insurance, lending, hiring, and predictive policing. In many of these scenarios, previous decisions have been made that are unfairly biased against certain subpopulations, for example those of a particular race, gender, or sexual orientation. Since this past data may be biased, machine learning predictors must account for this to avoid perpetuating or creating discriminatory practices. In this paper, we develop a framework for modeling fairness using tools from causal inference. Our definition of *counterfactual fairness* captures the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group. We demonstrate our framework on a real-world problem of fair prediction of success in law school.

1 Contribution

Machine learning has spread to fields as diverse as credit scoring [20], crime prediction [5], and loan assessment [25]. Decisions in these areas may have ethical or legal implications, so it is necessary for the modeler to think beyond the objective of maximizing prediction accuracy and consider the societal impact of their work. For many of these applications, it is crucial to ask if the predictions of a model are *fair*. Training data can contain unfairness for reasons having to do with historical prejudices or other factors outside an individual's control. In 2016, the Obama administration released a report[2] which urged data scientists to analyze "how technologies can deliberately or inadvertently perpetuate, exacerbate, or mask discrimination."

There has been much recent interest in designing algorithms that make fair predictions [4, 6, 10, 12, 14, 16–19, 22, 24, 36–39]. In large part, the literature has focused on formalizing fairness into quantitative definitions and using them to solve a discrimination problem in a certain dataset. Unfortunately, for a practitioner, law-maker, judge, or anyone else who is interested in implementing algorithms that control for discrimination, it can be difficult to decide *which* definition of fairness to choose for the task at hand. Indeed, we demonstrate that depending on the relationship between a protected attribute and the data, certain definitions of fairness can actually *increase discrimination*.

*Equal contribution. This work was done while JL was a Research Fellow at the Alan Turing Institute.

[2] https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

In this paper, we introduce the first explicitly causal approach to address fairness. Specifically, we leverage the causal framework of Pearl [30] to model the relationship between protected attributes and data. We describe how techniques from causal inference can be effective tools for designing fair algorithms and argue, as in DeDeo [9], that it is essential to properly address causality in fairness. In perhaps the most closely related prior work, Johnson et al. [15] make similar arguments but from a non-causal perspective. An alternative use of causal modeling in the context of fairness is introduced independently by [21].

In Section 2, we provide a summary of basic concepts in fairness and causal modeling. In Section 3, we provide the formal definition of *counterfactual fairness*, which enforces that a distribution over possible predictions for an individual should remain unchanged in a world where an individual's protected attributes had been different in a causal sense. In Section 4, we describe an algorithm to implement this definition, while distinguishing it from existing approaches. In Section 5, we illustrate the algorithm with a case of fair assessment of law school success.

2 Background

This section provides a basic account of two separate areas of research in machine learning, which are formally unified in this paper. We suggest Berk et al. [1] and Pearl et al. [29] as references. Throughout this paper, we will use the following notation. Let A denote the set of *protected attributes* of an individual, variables that must not be discriminated against in a formal sense defined differently by each notion of fairness discussed. The decision of whether an attribute is protected or not is taken as a primitive in any given problem, regardless of the definition of fairness adopted. Moreover, let X denote the other observable attributes of any particular individual, U the set of relevant latent attributes which are not observed, and let Y denote the outcome to be predicted, which itself might be contaminated with historical biases. Finally, $\hat{Y}$ is the *predictor*, a random variable that depends on A, X and U, and which is produced by a machine learning algorithm as a prediction of Y.

2.1 Fairness

There has been much recent work on fair algorithms. These include fairness through unawareness [12], individual fairness [10, 16, 24, 38], demographic parity/disparate impact [36], and equality of opportunity [14, 37]. For simplicity we often assume A is encoded as a binary attribute, but this can be generalized.

Definition 1 (Fairness Through Unawareness (FTU))**.** *An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process.*

Any mapping $\hat{Y} : X \to Y$ that excludes A satisfies this. Initially proposed as a baseline, the approach has found favor recently with more general approaches such as Grgic-Hlaca et al. [12]. Despite its compelling simplicity, FTU has a clear shortcoming as elements of X can contain discriminatory information analogous to A that may not be obvious at first. The need for expert knowledge in assessing the relationship between A and X was highlighted in the work on individual fairness:

Definition 2 (Individual Fairness (IF))**.** *An algorithm is fair if it gives similar predictions to similar individuals. Formally, given a metric $d(\cdot, \cdot)$, if individuals i and j are similar under this metric (i.e., $d(i, j)$ is small) then their predictions should be similar: $\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)})$.*

As described in [10], the metric $d(\cdot, \cdot)$ must be carefully chosen, requiring an understanding of the domain at hand beyond black-box statistical modeling. This can also be contrasted against population level criteria such as

Definition 3 (Demographic Parity (DP))**.** *A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$.*

Definition 4 (Equality of Opportunity (EO))**.** *A predictor $\hat{Y}$ satisfies equality of opportunity if $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$.*

These criteria can be incompatible in general, as discussed in [1, 7, 22]. Following the motivation of IF and [15], we propose that knowledge about relationships between all attributes should be taken into consideration, even if strong assumptions are necessary. Moreover, it is not immediately clear

for any of these approaches in which ways historical biases can be tackled. We approach such issues from an explicit causal modeling perspective.

2.2 Causal Models and Counterfactuals

We follow Pearl [28], and define a causal model as a triple (U, V, F) of sets such that

- U is a set of latent **background** variables,which are factors not caused by any variable in the set V of **observable** variables;
- F is a set of functions $\{f_1, \ldots, f_n\}$, one for each $V_i \in V$, such that $V_i = f_i(pa_i, U_{pa_i})$, $pa_i \subseteq V \backslash \{V_i\}$ and $U_{pa_i} \subseteq U$. Such equations are also known as **structural equations** [2].

The notation "pa_i" refers to the "parents" of V_i and is motivated by the assumption that the model factorizes as a directed graph, here assumed to be a directed acyclic graph (DAG). The model is causal in that, given a distribution $P(U)$ over the background variables U, we can derive the distribution of a subset $Z \subseteq V$ following an **intervention** on $V \setminus Z$. An intervention on variable V_i is the substitution of equation $V_i = f_i(pa_i, U_{pa_i})$ with the equation $V_i = v$ for some v. This captures the idea of an agent, external to the system, modifying it by forcefully assigning value v to V_i, for example as in a randomized experiment.

The specification of F is a strong assumption but allows for the calculation of **counterfactual** quantities. In brief, consider the following counterfactual statement, "the value of Y if Z had taken value z", for two observable variables Z and Y. By assumption, the state of any observable variable is fully determined by the background variables and structural equations. The counterfactual is modeled as the solution for Y for a given $U = u$ where the equations for Z are replaced with $Z = z$. We denote it by $Y_{Z \leftarrow z}(u)$ [28], and sometimes as Y_z if the context of the notation is clear.

Counterfactual inference, as specified by a causal model (U, V, F) given evidence W, is the computation of probabilities $P(Y_{Z \leftarrow z}(U) \mid W = w)$, where W, Z and Y are subsets of V. Inference proceeds in three steps, as explained in more detail in Chapter 4 of Pearl et al. [29]: 1. **Abduction**: for a given prior on U, compute the posterior distribution of U given the evidence $W = w$; 2. **Action**: substitute the equations for Z with the interventional values z, resulting in the modified set of equations F_z; 3. **Prediction**: compute the implied distribution on the remaining elements of V using F_z and the posterior $P(U \mid W = w)$.

3 Counterfactual Fairness

Given a predictive problem with fairness considerations, where A, X and Y represent the protected attributes, remaining attributes, and output of interest respectively, let us assume that we are given a causal model (U, V, F), where $V \equiv A \cup X$. We postulate the following criterion for predictors of Y.

Definition 5 (Counterfactual fairness)**.** *Predictor $\hat{Y}$ is* **counterfactually fair** *if under any context $X = x$ and $A = a$,*

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \tag{1}$$

for all y and for any value a' attainable by A.

This notion is closely related to **actual causes** [13], or token causality in the sense that, to be fair, A should not be a cause of $\hat{Y}$ in any individual instance. In other words, changing A while holding things which are not causally dependent on A constant will not change the distribution of $\hat{Y}$. We also emphasize that counterfactual fairness is an individual-level definition. This is substantially different from comparing different individuals that happen to share the same "treatment" $A = a$ and coincide on the values of X, as discussed in Section 4.3.1 of [29] and the Supplementary Material. Differences between X_a and $X_{a'}$ must be caused by variations on A only. Notice also that this definition is agnostic with respect to how good a predictor $\hat{Y}$ is, which we discuss in Section 4.

Relation to individual fairness. IF is agnostic with respect to its notion of similarity metric, which is both a strength (generality) and a weakness (no unified way of defining similarity). Counterfactuals and similarities are related, as in the classical notion of distances between "worlds" corresponding to different counterfactuals [23]. If $\hat{Y}$ is a deterministic function of $W \subset A \cup X \cup U$, as in several of

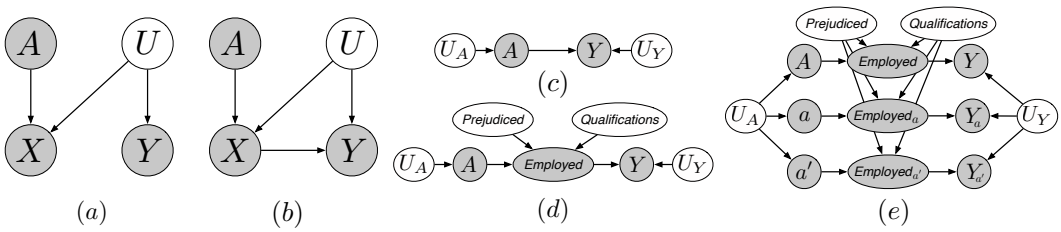

Figure 1: (a), (b) Two causal models for different real-world fair prediction scenarios. See Section 3.1 for discussion. (c) The graph corresponding to a causal model with A being the protected attribute and Y some outcome of interest, with background variables assumed to be independent. (d) Expanding the model to include an intermediate variable indicating whether the individual is employed with two (latent) background variables **Prejudiced** (if the person offering the job is prejudiced) and **Qualifications** (a measure of the individual's qualifications). (e) A twin network representation of this system [28] under two different counterfactual levels for A. This is created by copying nodes descending from A, which inherit unaffected parents from the factual world.

our examples to follow, then IF can be defined by treating equally two individuals with the same W in a way that is also counterfactually fair.

Relation to Pearl et al. [29]. In Example 4.4.4 of [29], the authors condition instead on X, A, and the observed realization of $\hat{Y}$, and calculate the probability of the counterfactual realization $\hat{Y}_{A \leftarrow a'}$ differing from the factual. This example conflates the predictor $\hat{Y}$ with the outcome Y, of which we remain agnostic in our definition but which is used in the construction of $\hat{Y}$ as in Section 4. Our framing makes the connection to machine learning more explicit.

3.1 Examples

To provide an intuition for counterfactual fairness, we will consider two real-world fair prediction scenarios: **insurance pricing** and **crime prediction**. Each of these correspond to one of the two causal graphs in Figure 1(a),(b). The Supplementary Material provides a more mathematical discussion of these examples with more detailed insights.

Scenario 1: The Red Car. A car insurance company wishes to price insurance for car owners by predicting their accident rate Y. They assume there is an unobserved factor corresponding to aggressive driving U, that (a) causes drivers to be more likely have an accident, and (b) causes individuals to prefer red cars (the observed variable X). Moreover, individuals belonging to a certain race A are more likely to drive red cars. However, these individuals are no more likely to be aggressive or to get in accidents than any one else. We show this in Figure 1(a). Thus, using the red car feature X to predict accident rate Y would seem to be an unfair prediction because it may charge individuals of a certain race more than others, even though no race is more likely to have an accident. Counterfactual fairness agrees with this notion: changing A while holding U fixed will also change X and, consequently, $\hat{Y}$. Interestingly, we can show (Supplementary Material) that in a linear model, regressing Y on A and X is equivalent to regressing on U, so off-the-shelf regression here is counterfactually fair. Regressing Y on X alone obeys the FTU criterion but is not counterfactually fair, so *omitting A (FTU) may introduce unfairness into an otherwise fair world.*

Scenario 2: High Crime Regions. A city government wants to estimate crime rates by neighborhood to allocate policing resources. Its analyst constructed training data by merging (1) a registry of residents containing their neighborhood X and race A, with (2) police records of arrests, giving each resident a binary label with $Y = 1$ indicating a criminal arrest record. Due to historically segregated housing, the location X depends on A. Locations X with more police resources have larger numbers of arrests Y. And finally, U represents the totality of socioeconomic factors and policing practices that both influence where an individual may live and how likely they are to be arrested and charged. This can all be seen in Figure 1(b).

In this example, higher observed arrest rates in some neighborhoods are due to greater policing there, not because people of different races are any more or less likely to break the law. The label $Y = 0$

does not mean someone has never committed a crime, but rather that they have not been caught. *If individuals in the training data have not already had equal opportunity, algorithms enforcing EO will not remedy such unfairness*. In contrast, a counterfactually fair approach would model differential enforcement rates using U and base predictions on this information rather than on X directly.

In general, we need a multistage procedure in which we first derive latent variables U, and then based on them we minimize some loss with respect to Y. This is the core of the algorithm discussed next.

3.2 Implications

One simple but important implication of the definition of counterfactual fairness is the following:

Lemma 1. *Let $\mathcal{G}$ be the causal graph of the given model (U, V, F). Then $\hat{Y}$ will be counterfactually fair if it is a function of the non-descendants of A.*

Proof. Let W be any non-descendant of A in $\mathcal{G}$. Then $W_{A\leftarrow a}(U)$ and $W_{A\leftarrow a'}(U)$ have the same distribution by the three inferential steps in Section 2.2. Hence, the distribution of any function $\hat{Y}$ of the non-descendants of A is invariant with respect to the counterfactual values of A. □

This does not exclude using a descendant W of A as a possible input to $\hat{Y}$. However, this will only be possible in the case where the overall dependence of $\hat{Y}$ on A disappears, which will not happen in general. Hence, Lemma 1 provides the most straightforward way to achieve counterfactual fairness. In some scenarios, it is desirable to define path-specific variations of counterfactual fairness that allow for the inclusion of some descendants of A, as discussed by [21, 27] and the Supplementary Material.

Ancestral closure of protected attributes. Suppose that a parent of a member of A is not in A. Counterfactual fairness allows for the use of it in the definition of $\hat{Y}$. If this seems counterintuitive, then we argue that the fault should be at the postulated set of protected attributes rather than with the definition of counterfactual fairness, and that typically we should expect set A to be closed under ancestral relationships given by the causal graph. For instance, if *Race* is a protected attribute, and *Mother's race* is a parent of *Race*, then it should also be in A.

Dealing with historical biases and an existing fairness paradox. The explicit difference between $\hat{Y}$ and Y allows us to tackle historical biases. For instance, let Y be an indicator of whether a client defaults on a loan, while $\hat{Y}$ is the actual decision of giving the loan. Consider the DAG $A \rightarrow Y$, shown in Figure 1(c) with the explicit inclusion of set U of independent background variables. Y is the objectively ideal measure for decision making, the binary indicator of the event that the individual defaults on a loan. If A is postulated to be a protected attribute, then the predictor $\hat{Y} = Y = f_Y(A, U)$ is not counterfactually fair, with the arrow $A \rightarrow Y$ being (for instance) the result of a world that punishes individuals in a way that is out of their control. Figure 1(d) shows a finer-grained model, where the path is mediated by a measure of whether the person is employed, which is itself caused by two background factors: one representing whether the person hiring is prejudiced, and the other the employee's qualifications. In this world, A is a cause of defaulting, even if mediated by other variables[3]. The counterfactual fairness principle however forbids us from using Y: using the twin network [4] of Pearl [28], we see in Figure 1(e) that Y_a and $Y_{a'}$ need not be identically distributed given the background variables.

In contrast, any function of variables not descendants of A can be used a basis for fair decision making. This means that any variable $\hat{Y}$ defined by $\hat{Y} = g(U)$ will be counterfactually fair for any function $g(\cdot)$. Hence, given a causal model, the functional defined by the function $g(\cdot)$ minimizing some predictive error for Y will satisfy the criterion, as proposed in Section 4.1. We are essentially learning a projection of Y into the space of fair decisions, removing historical biases as a by-product.

Counterfactual fairness also provides an answer to some problems on the incompatibility of fairness criteria. In particular, consider the following problem raised independently by different authors (e.g.,

[3] For example, if the function determining employment $f_E(A, P, Q) \equiv I_{(Q>0, P=0 \text{ or } A\neq a)}$ then an individual with sufficient qualifications and prejudiced potential employer may have a different counterfactual employment value for $A = a$ compared to $A = a'$, and a different chance of default.

[4] In a nutshell, this is a graph that simultaneously depicts "multiple worlds" parallel to the factual realizations. In this graph, all multiple worlds share the same background variables, but with different consequences in the remaining variables depending on which counterfactual assignments are provided.