COMP3224

Coursework IV

Alberto Tamajo

Student ID: 30696844

May 2022
Electronics and Computer Science Department
University of Southampton

Preface

I believe that if someone had asked me about the main take-home message of this Causal reasoning module prior to the development of this report, I would have answered as follows:

"Causal questions cannot be answered directly from the observed data without the postulation of some causal assumptions".

In contrast, if the same question were posed to me now, I would give a slightly different answer:

"If we truly want to benefit our society, then we need to leverage causality; otherwise, we may be actually harming ourselves without even realising it".

Indeed, my personal opinion is that there is nothing nobler than leveraging science to make the world a better place where to live, and the Counterfactual fairness paper has enlightened me about how tools developed in causal inference can be used to benefit our society.

The Counterfactual fairness paper has also provided me with the first hands-on application of the causal reasoning tools discussed during the semester. I strongly believe that this has been an invaluable experience that has provided me with a set of skills and a breadth of perspective that cannot be gained through typical classroom teaching. Indeed, drawing upon my so-far-modest experience, I can argue that while classroom teaching aids the learning process, self-studies and research allow an individual to achieve a whole new level of understanding.

Lastly, the Counterfactual fairness paper has heavily assisted me in understanding counterfactuals. Indeed, I reckon that the fairness problem tackled by the paper is a perfect example that can be used in classrooms to explain counterfactuals in a simple and practical way.

Word count

	Number of words
Main body	3.784
Appendices and preface	2.150
Total	5.934

Table 1: Report length

Contents

1 Introduction

Nowadays, machine learning algorithms are leveraged in fields as diverse as insurance, lending, hiring, and predictive policing. Due to the ethical and legal implications of the decisions made in these areas, it is of paramount importance to design machine learning predictors that do not discriminate against certain subpopulations, such as people of a certain race, gender, or sexual orientation. Furthermore, due to the nature of our society, training data can contain historical prejudices against some demographic groups. Thus, modellers must account for this when developing predictors so that to avoid the perpetuation of unfairly biased decisions.

As a consequence, given the relevance of these issues, a large number of definitions have been proposed in the literature to formalise fairness and consequently develop fair machine learning predictors. However, depending on the relationship between the protected attributes (attributes that cannot be discriminated against) and the data, most of these fairness definitions actually increase discrimination. Indeed, they fail to account for the causal mechanism behind the generation process of the data. For this reason, the authors of [1] introduce the first explicitly causal approach, named Counterfactual fairness, that leverages the causal framework developed by Pearl [2] to address fairness. However, even though it can be argued that modelling causality is necessary to tackle fairness, the proposal in [1] is also flawed as it is based on an argument that can lead to unfairly biased decisions in specific scenarios. On the grounds of this, this report proposes a modification to the Counterfactual fairness definition, dubbed Total Counterfactual fairness, which instead guarantees fair decisions in all possible circumstances.

In the sections that follow, multiple non-causal approaches to fairness will be analysed, and, whenever applicable, causal reasoning will be leveraged to expose their flaws intuitively. The review of these non-causal approaches will also highlight the importance of causality in fairness problems. Afterwards, the definition of Counterfactual fairness will be inspected. Due to limits on the number of words for this report, it will not be possible to give an account of counterfactuals (the causal inference tool leveraged by the definition of Counterfactual fairness); as such, the reader with no previous background knowledge of this topic is referred to Chapter 4 of [3]. Finally, the new proposal of this report, Total Counterfactual fairness, will be introduced.

2 Notation

Throughout this report, the following notation will be used. Let A, X and U denote the set of observable protected attributes, the set of observable non-protected attributes and the set of latent attributes of an individual, respectively.

The set of observable protected attributes A contains variables that must not be discriminated against, and the decision of whether a variable is protected or not is taken as a primitive in any given problem. The set of observable non-protected attributes X is defined as the union of the sets $X_{\prec A}$ and $X_{\nprec A}$, i.e. $X = X_{\prec A} \cup X_{\nprec A}$. While $X_{\prec A}$ denotes the set of the observable descendants of A, $X_{\nprec A}$ is the set of the observable non-descendants of A. Also, let $\hat{Y}$ be a random variable generated from (A, X, U) by a machine learning algorithm as a prediction of the ground truth decision Y. Notice that the latter, Y, might contain historical prejudices.

3 Fairness Through Unawareness

Fairness Through Unawareness(FTU)[4] claims that an algorithm is fair if its decision-making process does not explicitly use the set of protected attributes A.

In other words, an algorithm is fair as long as it is unaware of the set of protected attributes A. Clearly, any algorithm based on the function $\hat{Y} : X \rightarrow Y$ satisfies the FTU condition. However, despite its compelling simplicity, Fairness Through Unawareness does not prevent machine learning predictors from creating or perpetuating discriminatory practices in certain scenarios. Indeed, in some decision problems, even though the set X does not contain protected attributes, its elements may contain discriminatory information that is affected by A in a not so obvious way.

Therefore, prior to applying FTU, expert knowledge is necessary to create a graph causal model G of the decision problem at hand and evaluate the relationship between A and X. If some nodes in X are descendants of some nodes in A in the graph causal model G, Fairness Through Unawareness does not guarantee the fairness of the machine learning predictor. In contrast, if none of the nodes in X descends from A in the graph causal model G, the fairness of an algorithm satisfying the FTU condition is assured.

The chocolate bar scenario in Appendix A shows through a practical example how predictions of a machine learning algorithm satisfying FTU can be unfairly biased when discriminatory information spreads from some nodes in A to X.

4 Individual Fairness

Individual Fairness(IF)[5, 6, 7, 8] states that an algorithm is fair as long as it makes similar decisions for similar individuals. More formally, let $d(\cdot,\cdot)$ be a metric, for all individuals i and j, if i and j are similar under d then the predictions of a fair algorithm need to be similar, i.e. $\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)})$.

Obviously, the definition of Individual Fairness is agnostic with respect to its notion of similarity metric. On one side, this is an advantage as the IF definition attains generality and can be applied to a wide variety of sensitive decision problems. Conversely, this is also a weakness because there are no unified ways of defining similarity in fairness and a deep understanding of the domain at hand is necessary for a correct design of the similarity metric d.

In summary, Individual Fairness is not a tool that can be used out of the box to develop fair algorithms. Rather, it provides guidance on how to measure fairness in an algorithm's predictions through a fair similarity metric. The latter, however, requires deep knowledge of the domain at hand and accordingly, its design is left to the end-user.

5 Demographic Parity

A predictor $\hat{Y}$ is said to satisfy Demographic Parity (DP)[9, 10] if $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$.

In other words, a machine-learning algorithm satisfies Demographic Parity if the distribution of predictions for any two demographic groups is identical. Thus, Demographic Parity enforces algorithms to give predictions on a par among demographic groups.

It can be argued that, although it seems reasonable to treat different demographic groups on the same level, in specific scenarios, predictors satisfying DP actually create unfair bias in an otherwise fair world. Indeed, let us suppose as a matter of example that individuals belonging to a given demographic group $A = 1$ are much more likely to be qualified for a certain job than individuals of a demographic group $A = 0$. A predictor $\hat{Y}$ satisfying the DP condition would hire the same proportion of employees from both $A = 1$ and $A = 0$. This, however, does not provide equality of opportunity. Indeed, given that people belonging to $A = 0$ are less likely to be qualified than individuals in $A = 1$, then inevitably, a qualified individual belonging to $A = 0$ is more likely to be hired as opposed to a qualified person in $A = 1$, creating inequality of opportunities.

6 Equality of Opportunity

A predictor $\hat{Y}$ is said to satisfy Equality of Opportunity(EO)[11] if $P(\hat{Y}|A=0, Y=1) = P(\hat{Y}|A=1, Y=1)$.

That is, the distribution of the random variable $\hat{Y}$ should remain unchanged for individuals that have been assigned with the same decision on the training dataset, regardless of their protected attributes' value.

An example will make the definition of Equality of Opportunity clearer. Let us suppose that A denotes the gender of individuals and $A = 0$ corresponds to the female sex while $A = 1$ denotes the male sex. Also, let Y be a binary variable such that $Y = 1$ means "hired", and $Y = 0$ denotes "not hired". A predictor $\hat{Y}$ satisfying the EO condition gives the same proportion of "hired" and "not hired" decisions for male and female individuals that have been actually hired. This is quite different from the definition of Demographic Parity. Indeed, the latter would enforce an algorithm to predict the same proportion of "hired" and "not hired" decisions for the male and female groups irrespectively of whether they have been actually hired or not. Therefore, Equality of Opportunity can provide enhanced fairness in the decision-making process as it can overcome the potential unfairness of Demographic Parity that has been outlined in the previous section.

However, Equality of Opportunity relies on a basic assumption: individuals in the training dataset need to have had equal opportunity. In fact, if individuals in the training data have not already had equal opportunity, algorithms enforcing EO will not remedy such unfairness. The Football scouting scenario in Appendix B gives an example of this fact.

7 Counterfactual Fairness

As outlined in the previous sections, the fairness definitions mentioned so far have some drawbacks. Also, they fail to tackle the problem of learning a projection of Y into the domain of fair decisions so that to disregard any historical prejudices in the training datasets. For this reason, the authors of [1] propose a novel definition, dubbed Counterfactual fairness, which addresses fairness from a causal perspective. Indeed, the paper in question argues that, even though causal model assumptions are generally unfalsifiable, strong assumptions regarding the causal mechanisms of the world's phenomena are necessary to correctly model fairness.

7.1 Definition

The definition of Counterfactual fairness is based on counterfactuals, and it states that given a structural causal model $M = (U, V, F)$, where $V = A \cup X$, a predictor $\hat{Y}$ is said to be counterfactually fair if under any context $X = X$ and $A = a$,

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

for all values of y and for any value a' obtainable by A.

Therefore, informally, a machine learning predictor is counterfactually fair if it enforces the distribution over the possible predictions for an individual to remain identical in the case the individual's protected attributes change in a causal sense. This is based on the common sense that a decision is fair towards an individual if it is the same in the actual world and in a counterfactual world where the individual belongs to a different demographic group.

As a matter of example, let us assume that the distribution over the possible predictions for a given individual I with protected attribute $A =$ "*white*" and context attributes $X = x$ is $P(\hat{Y}_I)$. For a predictor to be considered counterfactually fair, the original distribution $P(\hat{Y}_I)$ needs to remain unchanged for the same individual I with context attributes $X = x$ but a different protected attribute $A =$ "*black*".

7.2 How to achieve counterfactual fairness

The definition of Counterfactual fairness implies that the set of protected attributes A should not be a cause of an algorithm's predictions. Otherwise, changing A while holding things which are not causally dependent on A constant will likely change the distribution of $\hat{Y}$, giving different decisions for different demographic groups.

As a result, the most straightforward way to achieve counterfactual fairness involves the use of the non-descendants of A in the predictors $\hat{Y}$. This is stated more formally in what follows:

Lemma 1. Let $M = (U, V, F)$ be the structural causal model of a decision problem. If a predictor $\hat{Y}$ is a function of the non-descendants of A then $\hat{Y}$ is guaranteed to be counterfactually fair.

The proof of Lemma 1 is straightforward. Let $\hat{Y} : X_{\nsucc A} \to Y$ be a prediction function such that $X_{\nsucc A}$ is the set of the observable non-descendants of A. Hence, $\hat{Y}$ is invariant with respect to the counterfactual values of A as, by definition, $X_{\nsucc A_{A \leftarrow a}}(U) = X_{\nsucc A_{A \leftarrow a'}}(U)$.

Simply put, Lemma 1 ensures that changing A in a counterfactual sense will not alter the distribution of predictions $\hat{Y}$ as discriminatory information is never involved in the decision-making process.

However, it is essential to notice that if some ancestors of the members in A are not in A, Lemma 1 allows for their use in the decision-making process. Even though this might seem rather counterintuitive, as ancestral information of discriminatory information might be itself prejudiced, the problem lies in the definition of A in the first place. Indeed, intuitively, the set of protected attributes A should be closed under ancestral relationships.

Finally, although counterfactually fair predictors are subject to the restrictions imposed by Lemma 1, they can be modelled under three levels of assumptions of increasing strength: Level 1, 2 and 3. More details regarding these assumption levels are provided in Appendix C.

7.3 How to train counterfactually fair predictors

A predictor $\hat{Y}$ is somewhat useless if it does not relate to the true predictions Y. Hence, it is necessary to minimise the predictive error of $\hat{Y}$ with respect to Y. Substantially, given that $\hat{Y}$ is counterfactually fair, a projection of Y into the space of fair decisions, disregarding historical biases, needs to be learnt.

Formally, let $\hat{Y} = g_\theta(\cdot)$ be a predictor parameterised by θ, such as a neural network or logistic regression. In order to minimise the predictive error of $\hat{Y}$ with respect to a training dataset $D = \{(A^{(i)}, X^{(i)}, Y^{(i)}) | i = 1, 2, ..., N\}$, an empirical loss $L(\theta)$ needs to be minimised with respect to θ. The choice of $L(\theta)$ depends on the decision problem at hand and, as such, could range from a simple mean-squared-error loss to a log-likelihood loss.

It is crucial to notice that in the presence of a Level 2 counterfactually fair predictor $\hat{Y}$, the loss function needs to be augmented with an expectation term. This holds because a deconvolution approach, which extracts the latent sources U from the observed values of the descendants of A and pipelines them to $\hat{Y}$ via $P(U|x, a)$, is used in Level 2 predictors. Consequently, the empirical loss is defined as

$$L(\theta) = \frac{\sum_{i=1}^{N} \mathbb{E}[l(y^i, g_\theta(U^{(i)}, x^{(i)}_{\nsucc A})) | x^{(i)}, a^{(i)}]}{N}$$

where $l(\cdot, \cdot)$ is again a loss function such as log-likelihood or squared loss, and $x^{(i)}_{\nsucc A}$ is the value of the non-descendants of A for a given instance $i \in D$.

Each expectation is taken with respect to random variable $U^{(i)} \sim P_{\mathcal{M}}(U|x^{(i)}, a^{(i)})$, where $P_{\mathcal{M}}(U|x^{(i)}, a^{(i)})$ is the conditional distribution of the latent variables given the observed values of a given instance $i \in D$ and the underlying causal model $\mathcal{M}$. As shown in the following algorithm, dubbed FairLearning, if the expectations cannot be computed analytically, Markov Chain Monte Carlo (MCMC) can be used to approximate them.

```
procedure FairLearning(D,$\mathcal{M}$)
    For each instance $i \in D$, sample $m$ MCMC samples $U_1^{(i)}, U_2^{(i)}, ..., U_m^{(i)} \sim P_{\mathcal{M}}(U|x^{(i)}, a^{(i)})$

    Create augmented dataset $D'$ such that each point $(a^{(i)}, x^{(i)}, y^{(i)}) \in D$ is replaced with the corresponding $m$ data points $\{(a^{(i)}, x^{(i)}, y^{(i)}, u_j^{(i)})\}_{j=1}^{m}$

    $\hat{\theta} \leftarrow argmin_\theta \frac{\sum_{i' \in D'} l(y^{(i')}, g_\theta(U^{(i')}, x_{\nsucc A}^{(i')}))}{N \times m}$
end procedure
```

7.4 Experiments

In order to prove the validity of their approach, the Counterfactual fairness authors have conducted experiments on some practical problems that require fairness. These experiments show that, in an unfairly biased word, unfair prediction algorithms perform better than their counterfactually fair counterparts. However, this was expected as counterfactually fair models necessarily need to trade-off some accuracy in favour of fairness.

Nevertheless, counterfactually fair predictors, especially the Level 3 predictors, can still be leveraged in real-world scenarios as they achieve reasonable performance. The interested reader is referred to Appendix D for an extensive analysis of the experiments conducted by the authors in the *law school success prediction problem.*

8 Total Counterfactual fairness

8.1 Overview

Even though it is common sense that a decision is fair towards an individual if it is the same in the actual world and in a counterfactual world where the individual belongs to a different demographic group, this report argues that this reasoning can lead to unfairly biased decisions. That is, enforcing the distribution over the possible predictions for an individual to remain identical in the case the individual's protected attributes change in a causal sense is potentially unfair in some scenarios.

The rationale behind this claim stems from the observation that a change in a protected attribute value in the counterfactual world might cause some of the protected attribute's descendants to vary in such a manner that the decision in the counterfactual world should differ from its factual counterpart. Note that an unfair decision in the counterfactual world can translate into an unfair decision in the factual world for a given individual that shares the same attributes.

It can be argued, however, that this problem should not exist in the first place as, following Lemma 1, the decision-making process should not be influenced by the value of the protected attributes descendants. Nonetheless, this report claims that this practice constitutes an obstacle to the achievement of fairness in some scenarios. Indeed, besides having a discriminatory effect on some of its descendants, a protected variable might affect its other descendants in a non-discriminatory way. As such, the prejudice-free descendants of a protected variable should actually be used as input to a fair predictor as they can provide significant information to the decision-making process. Without this additional information, the predictions of counterfactually fair algorithms are not guaranteed to be unbiased.

In what follows, a toy decision problem is provided in order to show through a practical example how the definition of counterfactual fairness fails to achieve fairness in certain scenarios. Furthermore, this example clearly shows that the use of the prejudice-free descendants is a necessary condition for the predictions not to discriminate against some individuals.

8.2 Causaland

Let us suppose that a new grocery store starts its business in Causaland (a fictional country) and consequently wishes to hire new employees for the position of shop assistant. Its hiring policy is straightforward: an applicant shall be hired if and only if he/she is in possession of a college degree. However, due to gender bias during the recruitment process, some male applicants without a college degree are also hired.

Furthermore, in Causaland, a much larger proportion of male individuals get a college degree as opposed to the female population. This happens because women tend to choose more competitive courses (those with higher rejection rates) than men when applying. Thus, male individuals are much more likely to be in possession of a college degree than females.

Even though the gender of individuals is assumed to be a protected attribute, it is of paramount importance to notice that only the path Gender $\rightarrow Y$ is considered discriminatory. This holds because some males are hired even without a college degree. In contrast, the path Gender $\rightarrow$ College degree $\rightarrow Y$ is not believed to be biased because, although girls are much less likely to get a college degree, the reason why this occurs is under their control. The graph causal model of this toy decision problem is illustrated in Figure 1.

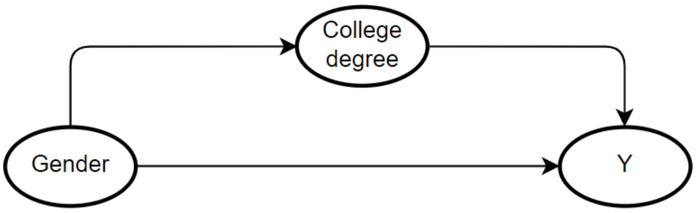

Figure 1: Graph causal model of the Causaland toy decision problem

Now, let us assume that a female individual without a college degree applies for the position of shop assistant. Hence, in line with the company's recruitment policy, this applicant should not be hired. Accordingly, a counterfactually fair predictor should output the same decision. Also, by definition, the counterfactually fair predictor should give the same decision in the case the gender of the applicant changes in a causal sense.

However, due to the underlying causal mechanism of the toy decision problem, the same applicant, had she been a male, would be much more likely to get a college degree and consequently be hired. Therefore, the decision of the counterfactually fair predictor would be considered unfair in this case as it violates the company's recruitment policy. Note that the decision of the predictor would not be unfair towards the counterfactual applicant if the college degree were to be leveraged during the decision-making process. However, this would violate the counterfactual fairness definition as the factual and counterfactual applicants get different decisions.

8.3 Definition

As clearly shown by the previous example, modelling the prejudice-free descendants of the protected attributes is a necessary condition to achieve fairness in some scenarios. However, at the same time, the Counterfactual fairness definition would be violated because the decisions in the factual and counterfactual worlds may differ. As a consequence, a new definition of fairness, which allows for the use of the prejudice-free descendants in the decision-making process but also guarantees equality of decisions in the factual and counterfactual worlds, is necessary. The solution to this problem is straightforward: it is sufficient to block the value of the prejudice-free descendants when computing counterfactual decisions. This suggests the following simple modification of the definition of Counterfactual fairness.

Let G be the graph causal model of a decision problem and let X_{DF} denote the set of prejudice-free descendants of the set of protected attributes A. A predictor $\hat{Y}$ is said to be totally counterfactually fair if under any context $X = x$, protected attributes $A = a$ and prejudice-free descendants $X_{DF} = x_{DF}$

$$P(\hat{Y}_{A\leftarrow a, X_{DF}\leftarrow x_{DF}}(U) = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a', X_{DF}\leftarrow x_{DF}}(U) = y|X = x, A = a)$$

It is important to note that the above definition of fairness, dubbed Total Counterfactual fairness, was proposed in the Counterfactual fairness paper first under the name of Path-specific Counterfactual fairness. However, while it was formulated with the only aim of giving freedom of use to the prejudice-free descendant in the Counterfactual fairness paper, this report has derived its formulation from a different perspective. Indeed, the Counterfactual fairness authors did not diagnose the potential unfairness of the Counterfactual fairness definition and, as such, did not recognise that the Path-specific Counterfactual fairness actually solves this issue.

Notice that Total Counterfactual fairness clearly solves the problem encountered in the Causaland example. Indeed, had the female applicant been a male, the decision would still be "not hired" as her male counterpart would not hold a college degree either. This holds because the definition of Total Counterfactual fairness blocks the value of the "College degree" attribute in the counterfactual world.

9 Conclusion

In conclusion, this report has provided a brief literature review of the most relevant fairness definitions and has highlighted the importance of modelling the causal structure of the world when dealing with fairness-related problems.

Furthermore, drawing upon the definition of Counterfactual fairness, a novel fairness definition, named Total Counterfactual fairness, has been proposed in order to guarantee equality of decisions in all possible circumstances, leveraging the causal framework of counterfactuals.

References

[1] Matt J Kusner et al. "Counterfactual fairness". In: *Advances in neural information processing systems* 30 (2017).

[2] Judea Pearl. "Causal inference in statistics: An overview". In: *Statistics surveys* 3 (2009), pp. 96–146.

[3] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer.* John Wiley & Sons, 2016.

[4] Nina Grgic-Hlaca et al. "The case for process fairness in learning: Feature selection for fair decision making". In: *NIPS symposium on machine learning and the law.* Vol. 1. 2016, p. 2.

[5] Cynthia Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference.* 2012, pp. 214–226.

[6] Matthew Joseph et al. "Fair algorithms for infinite and contextual bandits". In: *arXiv preprint arXiv:1610.09559* (2016).

[7] Christos Louizos et al. "The variational fair autoencoder". In: *arXiv preprint arXiv:1511.00830* (2015).

[8] Rich Zemel et al. "Learning fair representations". In: *International conference on machine learning.* PMLR. 2013, pp. 325–333.

[9] Muhammad Bilal Zafar et al. "Fairness constraints: Mechanisms for fair classification". In: *Artificial Intelligence and Statistics.* PMLR. 2017, pp. 962–970.

[10] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29 (2016).

[11] Muhammad Bilal Zafar et al. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". In: *Proceedings of the 26th international conference on world wide web.* 2017, pp. 1171–1180.

[12] DT Stan. *RStan: the R interface to Stan.* 2016.

Appendices

A Chocolate bar scenario

A company wishes to hire new employees by predicting their workplace productivity Y. They assume that an unobserved factor U exists that causes employees to be more productive in the workplace and, at the same time, induces the consumption of a large number of chocolate bars. The consumption of a large number of chocolate bars is the observed variable X. Besides, people belonging to a certain demographic group A are more likely to consume more chocolate as opposed to other demographic groups. However, the chocolate-greedy group is not more likely to exhibit the unobserved factor U or be more productive in the workplace. Figure 2 shows the graph causal model of the so-far described process.

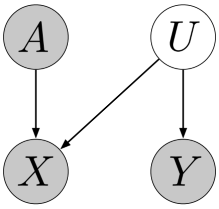

Figure 2: Graph causal model of the chocolate bar scenario.

Let F be a machine learning predictor based on the mapping $\hat{Y} : X \to Y$; clearly, F satisfies the FTU criterion. However, even though no demographic group is more likely to be productive in the workplace, individuals in the chocolate-greedy group are more likely to be hired as they are more prone to consume large quantities of chocolate. Therefore, machine learning predictors satisfying the FTU criterion might introduce unfairness into a fair world.

B Football scouting scenario

A professional football club is willing to estimate scout recommendation rates by amateur football teams through a machine learning algorithm F satisfying EO. In order to achieve this aim, F is trained on a dataset based on past data. In this data, A denotes a football player's race, X is a football player's team, and Y is a binary variable indicating whether a football player was scouted or not. Finally, U is assumed to be an unobserved variable representing the totality of a player's football skills and the scouting practices. The football player's team X is assumed to depend on his race A. Also, scouts are more likely to attend the matches of certain amateur football teams as opposed to others. Consequently, football players of some amateur teams are more likely to be scouted. This whole process is clearly illustrated in the graph causal model of Figure 3.

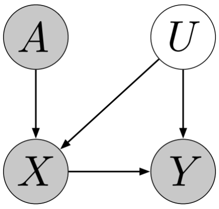

Figure 3: Graph causal model of the football scouting scenario.

Even though F satisfies the Equality of Opportunity condition, it can be argued that F would still give unfairly biased decisions. This holds true as higher scouting recommendation rates for some football teams are due to greater scouting attendance, not because players of a certain race are more likely to have better football skills. Therefore, given that the players in the training dataset did not have equal opportunity, F will perpetuate such unfairness in future predictions as it enforces EO on a dataset not satisfying the Equality of Opportunity condition a priori.

C Level 1, Level 2 and Level 3 counterfactually fair predictors

Although counterfactually fair predictors are subject to the restrictions imposed by Lemma 1, they can be modelled under three levels of assumptions of increasing strength. As it is also shown in Appendix D, a rise in the strength of assumptions generally translates into an increase in predictive power. Before delving into further details, it is of paramount importance to emphasise that causal model assumptions are generally unfalsifiable. Therefore, a causal model that passes testable implications can only be considered as a conjecture formulated according to the best of our knowledge.

C.1 Level 1

In Level 1, no further causal assumptions are postulated besides a partial causal ordering of the observable variables involved in the decision problem. In other words, a causal graph of the observable variables is sketched without inferring conditional probabilities, structural equations, and latent variables. Hence, following Lemma 1, Level 1 counterfactually fair predictors are only functions of the observable non-descendants of A.

Therefore, formally, a Level 1 counterfactually fair predictor is defined as a function $\hat{Y} : X_{\nprec A} \rightarrow Y$, where $X_{\nprec A}$ denotes the set of observable non-descendants of A. However, as there are generally only a few such variables in many problems, Level 1 predictors commonly achieve low predictive power.

C.2 Level 2

Level 2 predictors improve upon the previous fair prediction level by assuming that some latent variables U exist that non-deterministically affect the value of the observable descendants of A. Specifically, a causal graph of the problem at hand is built, including the latent variables U, and the conditional probability distribution of each variable given its parents is inferred.

As in Level 1, the values of the observable non-descendants of A can be fed directly into a predictor $\hat{Y}$. However, the same cannot be done for the observable descendants of A; otherwise, Lemma 1 would be violated. In order to overcome this issue, Level 2 fair predictors use a deconvolution approach which extracts the latent sources U from the observed values of the descendants of A and pipelines them to $\hat{Y}$ via $P(U|x, a)$.

Thus, formally, a Level 2 counterfactually fair predictor is defined as a function $\hat{Y} : (U, X_{\nprec A}) \rightarrow Y$, where $X_{\nprec A}$ denotes the set of observable non-descendants of A. Lemma 1 is satisfied in this case as the latent variables U are assumed to be independent of any observable variable. As noticeable, Level 2 predictors increase the information extracted in the data, which translates into a better prediction performance than Level 1.

C.3 Level 3

Level 3 fair predictors maximise the information extracted in the data, achieving the highest predictive power, by postulating the strong assumption of an additive error model. This requires building a structural causal model $M = (U, X_{\nprec A} \cup X_{\prec A} \cup A, F)$ for the problem at hand such that for each variable $X^{(i)}_{\nprec A} \in X$, its function is defined as $F_i = f_i(pa_i) + \epsilon_i$, where pa_i denotes the set of parents of $X^{(i)}_{\nprec A}$ and ϵ_i is an error term.

As in Level 1 predictors, the values of the observable non-descendants of A are fed directly into $\hat{Y}$. However, the same cannot be done for the observable descendants of A; otherwise, Lemma 1 would be violated. In order to overcome this issue, a deconvolution approach is used. However, unlike Level 2, here, the error terms of the observable descendants of A are calculated deterministically from the observed values and are given as input to $\hat{Y}$. This can be done thanks to the assumption of an additive error model.

As such, formally, a Level 3 counterfactually fair predictor is defined as a function $\hat{Y} : (U_{\prec A}, X_{\nprec A}) \rightarrow Y$, where $U_{\prec A}$ denotes the error terms of the observable descendants of A and $X_{\nprec A}$ denotes the set of observable non-descendants of A. Level 3 predictors satisfy Lemma 1 as the error terms are generated by the exogenous variables, which by definition are assumed to be independent of any observable variable. Notice that the error terms of the protected attributes are not fed to $\hat{Y}$; otherwise, the ancestral closure property would not be satisfied.

D Law school success prediction problem

Given information on 21,790 law school students, such as their entrance exam scores (LSAT), their grade-point average (GPA), and their first-year average grade (FYA), a school is willing to predict a new applicant's FYA. Although LSAT, GPA and FYA are commonly known to be unfairly biased due to social factors, the school would like the predictions not to be biased against an applicant's race and sex.

D.1 Models

The paper's authors have built four different models and compared their prediction accuracy and fairness level. All predictors use logistic regression but output predictions from different input variables.

The first predictor, named **Full predictor**, leverages all features of interest, including the sex and race protected attributes, for the decision-making process. The second model, dubbed **Unaware predictor**, satisfies the FTU condition and, as such, uses all features except for the race and sex protected variables.

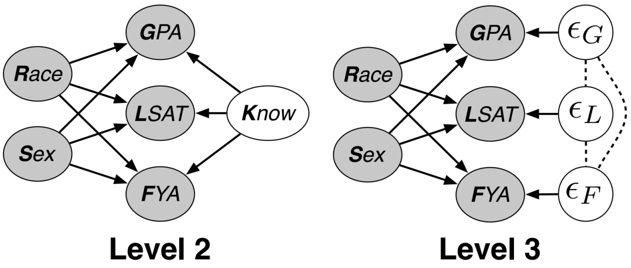

Figure 4: (Level 2)Graph causal model for assumed for the **Fair K predictor**. (Level 3) Graph causal model assumed for the **Fair Add predictor**

The **Fair K predictor**, the third model, is a Level 2 counterfactually fair predictor. Here, it is assumed that a latent variable representing a student's knowledge (K) influences the GPA, LSAT and FYA scores. This can be clearly seen in the graph causal model in Figure 4(Level 2). In order to infer the posterior distribution of K given the observed variables, the probabilistic language Stan [12] is employed on the training dataset, assuming the following distributions for the feature attributes:

$$GPA \sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G) \qquad FYA \sim \mathcal{N}(b_F + w_F^K K + w_F^R R + w_F^S S, 1)$$

$$LSAT \sim Poisson(exp(b_L + w_L^K K + w_L^R R + w_L^S S)) \qquad K \sim \mathcal{N}(0, 1)$$

Finally, the last model, called **Fair add predictor**, is a Level 3 counterfactually fair predictor. Thus, given the assumed graph causal model in Figure 4(Level 3), the GPA, LSAT and FYA are are expressed as follows:

$$GPA = b_G + w_G^R R + w_G^S S + \epsilon_G, \ \epsilon_G \sim p(\epsilon_G)$$
$$LSAT = b_L + w_L^R R + w_L^S S + \epsilon_L, \ \epsilon_L \sim p(\epsilon_L)$$
$$FYA = b_F + w_F^R R + w_F^S S + \epsilon_F, \ \epsilon_F \sim p(\epsilon_F)$$

The error terms $\epsilon_G, \epsilon_L, \epsilon_F$, which are used as input to the **Fair add predictor**, are estimated by fitting three deterministic models that use race and sex to predict GPA, LSAT and FYA, respectively, and then compute the residuals of each model.

D.2 Accuracy results

	Full	Unaware	Fair K	Fair Add
RMSE	0.873	0.894	0.929	0.918

Table 2: Root mean squared error(RMSE) achieved by the four predictors on the test set.

Results in Table 2 show that the unfair models slightly outperform the counterfactually fair predictors in terms of prediction accuracy. However, this was expected as counterfactually fair models necessarily trade-off accuracy in favour of fairness. It is crucial to notice that the **Fair add predictor** achieves better performance accuracy than the **Fair K predictor** due to its stronger underlying assumptions. Also, the reason why the **Full predictor** outperforms the **Unaware predictor** is self-explanatory: it also uses the race and sex attributes for the prediction task.

From these experiments, it is possible to conclude in general terms that, in an unfairly biased word, unfair prediction algorithms perform better than their fair counterparts. Nevertheless, fair predictors can still be used as they achieve reasonable performance.

D.3 Fairness results

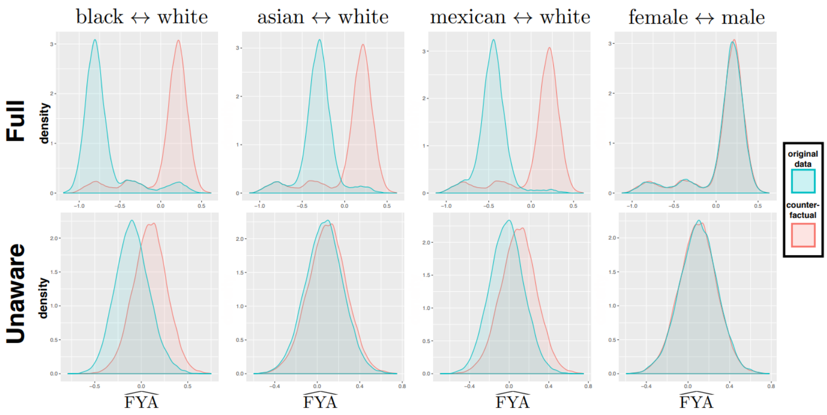

Figure 5: Density plots of predicted FYA_a and $\text{FYA}_{a'}$

The **Fair K** and **Fair add** predictors are counterfactually fair by definition. Here, it is demonstrated, instead, that the **Full** and **Unaware** models give unfairly biased predictions in the law school dataset.

By assuming that the true model of the world is given by Figure 4 (Level 2), the parameters of this model can be fitted using the observed data. Successively, samples can be generated from the model, and consequently, their counterfactual counterparts can be easily computed. If a model is counterfactually fair, the distribution of predicted FYA for the original data (FYA_a) and the density of predicted FYA for the counterfactual data ($\text{FYA}_{a'}$) should lie exactly on top of each other. However, this is not the case for the **Fair K** and **Fair add** predictors, as illustrated in Figure 5. Therefore, the **Fair K** and **Fair add** predictors are proved not to be counterfactually fair when employed in the law school dataset.