# Contrastive Algorithmic Fairness: Part 1 (Theory)

**Tapabrata Chakraborti, Arijit Patra, Alison Noble**
Department of Engineering Science
University of Oxford, Oxfordshire, UK
tapabrata.chakraborty; arijit.patra; alison.noble @eng.ox.ac.uk

## Abstract

We present contrastive fairness, a new direction in causal inference applied to algorithmic fairness and argue that just asking the traditional counterfactual "what if?" question might not be enough. We establish the theoretical and mathematical implications of the contrastive question "why this and not that?" in context of algorithmic fairness in machine learning. This is essential to defend the fairness of algorithmic decisions in tasks where a person or sub-group of people is chosen over another (eg. job recruitment, university admission, company layovers, etc). This development is also helpful to institutions to ensure or defend the fairness of their automated decision making processes. This work focuses on laying down the overarching theoretical and mathematical aspects with illustrative thought examples, with the aim of deploying them to solve relevant applications in future.

## 1 Introduction

Machine learning based decision systems have achieved near human performance in many tasks in recent times. But as these algorithms have grown more powerful, they have become more complex (with numerous parameters) and hence more opaque (the decision making process is not easily explainable) [1]. Machine learning after all is a data driven optimal function fitting exercise, thus it has always dealt with association, rather than causation [2]. Given, the broad use of machine learning algorithms in the modern world, precautions to make sure the fairness of the decision making process of such algorithms is of great importance [13].

The machine may take decisions partly based on protected variables (race, gender, sexual orientation, etc) learned from historic data having inherent bias [3]. Then there is the possibility of such bias getting perpetuated with significant social consequence for such tasks like job recruitment, university admission, insurance/lending, preemptive criminal profiling, etc [4, 20]. Modern machine learning methods should avoid such unethical discriminatory practice [18]. After all, the efficacy of a decision making process should be based on both efficiency and ethics [15].

We present *contrastive fairness*, a new direction in causal inference applied to algorithmic fairness. Earlier causal inferential methods in algorithmic fairness dealt with the "what if?" question [5]. We establish the theoretical and mathematical foundations to answer the question "why this and not that?". This is essential to defend the fairness of algorithmic decisions in tasks where a person or sub-group of people is chosen over another (job recruitment, university admission, etc). At its core, any question of fairness is a comparison, because equality is not absolute in society [6]. Some discrimination is part of the process itself (say employee recruitment), what must be ensured therefore, is that the discrimination is on fair grounds. Hence the question of why a certain person was chosen and not another, is of utmost pertinence. Contrastive questions and their explanation [19] has been around for quite some time but not within the purview of artificial intelligence and machine learning.Contrastive explanation in artificial intelligence has only recently been discussed by Miller [7], but for the first time it is formally introduced to algorithmic fairness in this work. It is to be noted that the current paper is meant to lay theoretical and mathematical foundations of contrastive logic in the realm of

algorithmic fairness. Data driven applications and experimental evaluations are to be pursued in future work.

## 2 Background Concepts

The present paper combines two distincSilva, R. and Evans, R. Causal inference through a witness protection program. Journal of Machine Learning Research, 17(56):1–53, 2016.t areas of research, those of algorithmic fairness and causal inference, both comparatively niche areas in machine learning. We provide in this section for the reader, a brief collection of underlying definitions and concepts related to both of these areas.

### 2.1 Algorithmic Fairness

We first define a few notations which are used throughout this paper. Let $Y$ be the expected outcome and $\hat{Y}$ be the predicted outcome. $X$ is the set of observable attributes and $U$ is the set of latent attributes of an individual. $A$ is the set of protected attributes of the individual on which the algorithm should not base its prediction on, in order to be fair. Of course, the intuitive but naive assumption in that case would be that an algorithm maybe considered to be fair if $\hat{Y}$ is only dependent on $X$ and not $A$. However, this amounts to "fairness through unawareness" as there may be attribute(s) in $X$ that are analogous to attribute(s) in $A$, though not explicitly the same. This makes it necessary to devise more strict rules to ensure algorithmic fairness.

Most earlier notions of algorithmic fairness were global, that is true for the population. The two most popular among these are Demographic Parity and Equality of Opportunity. Demographic parity holds if $P(\hat{Y} \mid A = 0) = P(\hat{Y} \mid A = 1)$, that is, we get the same prediction, irrespective of the value to which the protected attributes are set at. Note that this does not take into account the expected outcome $Y$ which means it ensures equality of result over the population, instead of any calibration using expected outcomes in sub-populations. Equality of opportunity does exactly that, that is it only seeks to ensure a certain prediction if the expected outcome supports that prediction for the sub-population in question. Equality of opportunity holds if $P(\hat{Y} = 1 \mid A = 0, Y = 1) = P(\hat{Y} = 1 \mid A = 1, Y = 1)$. It has been shown that these two criteria can never simultaneously hold true [12].

This brings to light the need for individual level fairness criterion, besides the above population level ones. If individual $i$ and $j$ are similar, that is some distance metric $d(i, j)$ is less than a small threshold, then individual fairness holds if $\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)})$. Of course, this introduces the constraint that the metric $d(i, j)$ should be properly chosen, which requires some domain knowledge expertise.

### 2.2 Causal Inference

Structural Causal Models (SCM) [8] are the backbone of causal inference methods [9]. These consist of three major interacting elements: causal diagrams, structural equations, and counterfactual/intervention logic. These together make up the triple of sets $(U, X, F)$ which constitute the SCM.

1. Causal graphical diagrams, which are basically directed acyclic graphs (DAG). The nodes of the diagram are the variables and the directed arrow between them specify the flow of causal relations between the variables. There are two types of variables: $U$ is the set of latent background variables and $X$ are observable variables.

2. Structural equations are a set of functions $\{f_1, \ldots, f_n\} \in F$ corresponding to the variables $\{X_1, \ldots, X_n\} \in X$ such that $X_i = f_i(p_i, U_{p_i})$, $p_i \subseteq X \backslash \{X_i\}$ and $U_{p_i} \subseteq U$.

3. For causal inferential analysis, counterfactual and interventional logic is carried out using a set of rules called *do-calculus*.

Since causal diagrams are essentially directed acyclic graphs, each observable variable $X_i$ will be connected to its parent variables $p_i$, where $X_i \subseteq X$ and $p_i \subseteq X \backslash X_i$. Thus we see above that the value of the observable variable $X_i$ depends on its parent variables as well as the latent variables $U$, through the function $f_i$.

**Intervention logic.** As seen above, the value of a measurable variable $X_i$ is given by $X_i = f_i(p_i, U_{p_i})$. Now if an external agent deliberately sets the value of $X_i = x$, then that is called an intervention (eg. randomised control trials). So assuming that we know the probability distribution $P(U)$ of latent variables $U$, we can perform an intervention on $Z$ variables belonging to $X$ (that is $Z \subseteq X$), and then compute the resulting probability distribution of the remaining variables in $X$ other than $Z$, that is $X \setminus Z$.

**Counterfactual logic.** This then also helps us to do counterfactual calculations, where we essentially compute $P(Y_{Z \leftarrow z}(U) \mid W = w)$. Here $Y$ are those variables belonging to the set of observable variables $X$, that we want to measure the effect of, so essentially output variables. $Z$ are those variables that we have intervened on by setting to value $z$, and $W$ are all other variables with known probability distribution.

## 2.3 Counterfactual Fairness

Kusner *et al.* [5] present the notion of counterfactual fairness. For a given problem of algorithmic fairness, let the causal model be given as usual by the set tuple $(U, V, F)$, where $V \equiv A \cup X$. $A$ are the protected variables and $U$ are the latent variables. $X$ are the observables variables other than $A$, so that they together make up the total set of observable variables $V$. $\hat{Y}$ is a fair predictor of the output variables $Y$ if

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

This condition of counterfactual fairness should be fair for any $x, a, a'$ and for all $y$. The equation essentially enforces the condition that the probability distribution of $Y$ should not be affected if any of the protected variables are intervened on keeping other conditions same [14].

# 3 Contrastive Fairness

In this Section we present the idea of contrastive fairness in details which is the main contribution of this paper. First we present several contrastive fairness questions and then formally formulate them. We take initial cues from the work on contrastive explanation by Miller [7].

## 3.1 Why do we need Contrastive Fairness? Why Counterfactual Fairness is not enough?

Counterfactual fairness formalised the use of causal inference in ensuring fairness of machine learning algorithms. However, the criterion is population based, whereas many real life fairness question compare how two individuals are treated, and whether the difference in decision for them was fair? Why was this decision taken for an individual and not some other decision? All these are contrastive cases of individual fairness, which requires some further considerations in order to incorporate. We still use the same counterfactual logic but expand it to fit contrastive cases.

## 3.2 Posing Contrastive Questions

First we list the main contrastive causal questions from existing literature. Then we modify them into contrastive questions pertaining to algorithmic decision making. Lastly we modify them to form fairness criteria.

Van Bouwel and Weber [10] mention 3 kinds of **contrastive causal questions** that may be posed.

- *P-contrast:* Why does object $X$ have property $P$, rather than property $Q$?
- *O-contrast:* Why does object $X$ have property $P$, but object $Y$ has property $Q$?
- *T-contrast:* Why does object $X$ have property $P$ at time $t$, but property $Q$ at time $t'$?

This defines three types of contrast: within an object (P-contrast), between objects themselves (O-contrast), and within an object over time (T-contrast).

We modify these questions to ask **algorithmic decision questions** as follows:

- *D-contrast:* Why does individual $I$ receive decision $D$, rather than decision $D'$?
- *I-contrast:* Why does individual $I$ receive decision $D$ but individual $J$ receives decision $D'$?
- *T-contrast:* Why does individual $I$ receive decision $D$ at time $t$, but decision $D'$ at time $t'$?

Again, note the three types of contrast: for one individual (D-contrast), between two individuals (I-contrast), and for one individual over time (T-contrast).

Then we convert these into **contrastive fairness questions**, quite simply by rephrasing as follows:

- *D-contrast:* Is it fair to make decision $D$ for individual $I$, instead of decision $D'$?
- *I-contrast:* Is it fair to make decision $D$ for individual $I$, while make $D'$ for individual $J$?
- *T-contrast:* Is it fair to make decision $D$ for individual $I$ at time $t$, but make $D'$ at time $t'$?

These three questions framed here give the main groundwork of ensuring contrastive fairness of algorithmic decision making pertaining to an individual. Next we mathematically formulate these criteria.

### 3.3 D-Contrast: Is it fair to make decision $D$ for individual $I$, instead of decision $D'$?

This basically boils down to counterfactual fairness but for a particular individual. This is because if the decision making process is counterfactually fair for that individual for the entire decision space, then it should be fair when contrasting between any two decision made for that individual. Thus the the decision making algorithm is fair for a particular individual $i$ if for any valid decision value $d$, the following holds.

$$P(\hat{Y}_{A_i \leftarrow a}(U_i) = d \mid X_i = x, A_i = a) = P(\hat{Y}_{A_i \leftarrow a'}(U_i) = d \mid X_i = x, A_i = a) \tag{2}$$

Other symbols have same meaning as in eqn. 1. It should be noted that though the above equation ensures fairness of the decision making process, it does not comment on the fairness of the two competing decisions themselves. To show that the decision taken, say $d$, is better than an alternative decision say $d'$, the predicted probability score of the former should be greater than the latter as shown below.

$$P(\hat{Y}(U_i) = d \mid X_i = x_i, A_i = a_i) > P(\hat{Y}(U_i) = d' \mid X_i = x_i, A_i = a_i) \tag{3}$$

### 3.4 I-Contrast: Is it fair to make decision $D$ for individual $I$, but make $D'$ for individual $J$?

When comparing decisions between two individuals however, we need to make further assumptions. Not only the decision making processes must be separately fair for both individuals, but also the difference in decision should be "sensible", that is the probability values generated by the predictor should support that.

First we establish the fairness for the two individuals for the entire decision space as follows:

$$P(\hat{Y}_{A_i \leftarrow a_i}(U_i) = d \mid X_i = x_i, A_i = a_i) = P(\hat{Y}_{A_i \leftarrow a_i'}(U_i) = d \mid X_i = x_i, A_i = a_i)$$
$$P(\hat{Y}_{A_j \leftarrow a_j}(U_j) = d \mid X_j = x_j, A_j = a_j) = P(\hat{Y}_{A_j \leftarrow a_j'}(U_j) = d \mid X_j = x_j, A_j = a_j) \tag{4}$$

Next, even if the decision making process itself is fair, for the decision to "make sense", for one individual the decision taken should have higher probability score assigned by the predictor than the alternative decision, while the opposite should hold true for the other individual. This is presented mathematically as follows:

$$P(\hat{Y}(U_i) = d \mid X_i = x_i, A_i = a_i) > P(\hat{Y}(U_i) = d' \mid X_i = x_i, A_i = a_i)$$
$$P(\hat{Y}(U_j) = d' \mid X_j = x_j, A_j = a_j) > P(\hat{Y}(U_j) = d \mid X_j = x_j, A_j = a_j) \tag{5}$$

4

Lastly, once must make sure that even if the protected variable values of the two individuals were to be same counterfactually, then also decision $D$ would have higher value than decision $D'$ for individual $I$ and decision $D'$ would have higher value than decision $D$ for individual $J$. This is present below.

$$P(\hat{Y}_{A_i \leftarrow a_j}(U_i) = d \mid X_i = x_i, A_i = a_i) > P(\hat{Y}_{A_i \leftarrow a_j}(U_i) = d' \mid X_i = x_i, A_i = a_i)$$
$$P(\hat{Y}_{A_j \leftarrow a_i}(U_j) = d' \mid X_j = x_j, A_j = a_j) > P(\hat{Y}_{A_j \leftarrow a_i}(U_j) = d \mid X_j = x_j, A_j = a_j) \quad (6)$$

If these equations are satisfied then one can surmise that the contrast in decision made between these two individuals is fair.

### 3.5 T-Contrast

The main question being asked here is that if over time the decision made for an individual changes, is that change fair or not. To ensure, this one has to first verify that the process itself is fair for all valid decisions $d$ that can be made over all time points $t$. This is shown here:

$$P(\hat{Y}_{A_i \leftarrow a_i}(U_i) = d \mid X_i = x_i(t), A_i = a_i) = P(\hat{Y}_{A_i \leftarrow a_i'}(U_i) = d \mid X_i = x_i(t), A_i = a_i) \quad (7)$$

Now if the original decision was $d$ at time $t$, and became some other $d'$ at time $t'$, then it must also hold that at time $t$, decision $d$ had higher prediction than $d'$, whereas at time $t'$, the opposite is true. This is formulated below:

$$P(\hat{Y}(U_i) = d \mid X_i = x_i(t), A_i = a_i) > P(\hat{Y}(U_i) = d' \mid X_i = x_i(t), A_i = a_i)$$
$$P(\hat{Y}(U_i) = d' \mid X_i = x_i(t'), A_i = a_i) > P(\hat{Y}(U_i) = d \mid X_i = x_i(t'), A_i = a_i) \quad (8)$$

### 3.6 Illustrative thought example: Fairness of Employee Job Location

Consider two employees $P$ and $Q$ having the same job duties and responsibilities in the same organisation. The organisation has office locations in London and different other locations of the UK. Employees might have a preference of working in London, so if they are assigned to a different office location, the decision making process should be fair. This becomes even more significant for the organisation, if a contrastive allocation of location between two employees is challenged and needs to be defended, especially if the decision is taken by an algorithm based on employee background data and performance statistics. We consider all the three scenarios of contrastive fairness for this test case and discuss what needs to be shown, in order to prove fairness.

Suppose employee $P$ is joining the organisation and in his/her application form had indicated that he/she would prefer to be located in the London headquarters, but is being located elsewhere due to more employee requirements in the remote office, and this decision is being taken by a machine learning enabled "HR algorithm". This decision should not be based on such protected attributes like sex, religion, ethnicity, etc. This is an example of *D-contrast* and to ensure fairness, the algorithm must satisfy equations 2 and 3. Also consider the case of the same employee $P$ being first located in London, and then at a later point of time being relocated to another office. In that case, the algorithm must satisfy the conditions of *T-contrast* via equations 7 and 8. Lastly consider employee $P$ being assigned a London office and employee $Q$ being assigned a remote office and employee $Q$ counters this decision in the belief that the decision was biased by ethnicity. It is of great importance to the organisation to be able to justify the decision fairness of the "HR algorithm" and that can be achieved using the *I-contrast* equations 4, 5 and 6.

### 3.7 Why make contrastive decision instead of making same decision for both individuals?

A core notion underpinning contrastive fairness is that one decision is more "desirable" than another and hence the need to justify the difference in decision between two individuals. In that case, a natural follow-up question that arises is why not make the "desirable" decision for both individuals, why at

all go for the less desirable alternative decision? For example, revisiting the illustrative example in Section 3.2, why not locate both employees $P$ and $Q$ in the London office, with the assumption that this is the more desirable decision, that is both employees prefer to work in London.

To justify the contrastive decision in such a situation, one has to first show that the preferable decision $d$ had higher probability score for individual $I$ than individual $J$ by enough margin, even when the protected attributes are swapped or made same as shown below:

$$P(\hat{Y}_{A_i \leftarrow \{a_i, a_j\}}(U_i) = d \mid X_i = x_i, A_i = a_i) - P(\hat{Y}_{A_j \leftarrow \{a_i, a_j\}}(U_j) = d \mid X_j = x_j, A_j = a_j) > \lambda_{ij} \tag{9}$$

Besides this, it helps further to justify the decision making process if similar conditions can be shown for $d'$ but with individual $J$ having higher probability than individual $I$, but this is not a necessary condition. Note that combinations of protected attributes are tried by counterfactual logic to verify invariance and hence fairness.

$$P(\hat{Y}_{A_j \leftarrow \{a_i, a_j\}}(U_j) = d' \mid X_j = x_j, A_j = a_j) - P(\hat{Y}_{A_i \leftarrow \{a_i, a_j\}}(U_i) = d' \mid X_i = x_i, A_i = a_i) > \lambda_{ij} \tag{10}$$

Though fairness of algorithmic decision making is the main objective of the problem at hand, it should ideally be achieved with non-significant loss in algorithmic performance. If the performance error is the difference between the predicted output ($\hat{Y}$) and expected output ($Y$), then this can be used along with the fairness criteria as a multi-objective pareto front optimisation, with higher priority on fairness in general.

## 4 Test Case: Law School Success Revisited

Though we have formalized contrastive reasoning in the context of algorithmic fairness for the first time in this work, it falls under the purview of causal inference mehtods, at the heart of which lies interventional and counterfactual logic always. So in order to demonstrate how contrastive logic in fairness builds on the core counterfactual logic, we revisit the same experimental design on law school success as used in [5]. We first point out how the counterfactual fairness evaluation as described in [5] might not be adequate under some conditions. After that we move on to explain how contrastive equations can be utilised to answer further questions using the same basic framework.

### 4.1 Population Level Counterfactual Fairness [5]

**Dataset.** The Law School Admission Council dataset [11] has data on 21,790 law students across 163 United States law schools. For each student, it has information like pre-entrance grade-point average (GPA) score, law school entrance examination score (LSAT), post-entrance law school first year grade point average (FYA). The dataset also has some social attributes recorded for students, like race and sex.

**Problem.** Predict the FYA with sufficient accuracy based on LSAT and GPA while ensuring it is not biased by protected attributes like race and sex.

**Model.** The authors propose a couple of different graphical diagrams to model the problem with some assumptions. Of these, we only consider the highest level (called Level 3 in [5]) in this work, which claims to ensure counterfactual fairness under some strong assumptions. The model is presented in Fig 1(a). The corresponding structural equations have GPA, LSAT and FYA as functions of race (R), sex (S) [16] and independent error terms as follows:

$$\begin{aligned} \text{GPA} &= b_G + w_G^R R + w_G^S S + \epsilon_G, \quad \epsilon_G \sim p(\epsilon_G) \\ \text{LSAT} &= b_L + w_L^R R + w_L^S S + \epsilon_L, \quad \epsilon_L \sim p(\epsilon_L) \\ \text{FYA} &= b_F + w_F^R R + w_F^S S + \epsilon_F, \quad \epsilon_F \sim p(\epsilon_F) \end{aligned} \tag{11}$$
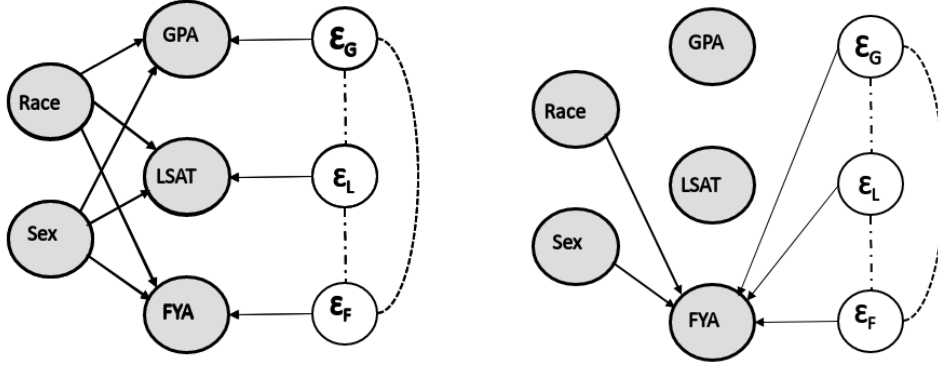
Figure 1: (a) Causal graphical model of the law school success test case [5]; race and sex are protected variables, $\epsilon_*$ are latent variables, GPA, LSAT and FYA are emasured variables. [5] (b) Even when the output variable is directly modeled by latent variables, it might still have non-linear dependance on protected attributes especially by proxy.

Since $R$ and $S$ are the protected attributes, the independent error terms in the linear model ($\epsilon_G$, $\epsilon_L$, $\epsilon_F$) are considered to be the latent variables ($U$ from previous sections) the probability distribution of which must be estimated. The terms $b_*$ and $w_*^*$ are the constants and weights respectively of the standard linear models that are fitted to the data. The latent variables, though assumed to be independent of protected variables, may have interactions between them, but this is chosen to be neglected for the model formulation.

Assuming GPA and LSAT be the observed variables, and $\hat{\text{GPA}}$ and $\hat{\text{LSAT}}$ be their corresponding predictors, then the independent error terms may be estimated as follows. $\epsilon_G$ may be calculated as:

$$\hat{\text{GPA}} = b_G + w_G^R R + w_G^S S, \;\; \epsilon_G = \text{GPA} - \hat{\text{GPA}}$$
$$\hat{\text{LSAT}} = b_L + w_L^R R + w_L^S S, \;\; \epsilon_L = \text{LSAT} - \hat{\text{LSAT}} \tag{12}$$

Thereafter using the estimated values of $\epsilon_G$ and $\epsilon_L$, the predicted value of $\hat{\text{FYA}}$ is calculated as:

$$\hat{\text{FYA}} = b_F' + w_F^G \epsilon_G + w_L^S \epsilon_L \tag{13}$$

The authors [5] claim that the predictor $\hat{\text{FYA}}$ thus calculated can be taken to be counterfactually fair since it is only a function of the latent variables that are independent of the protected attributes.

### 4.2 Advocating Checking of Individual Level Contrastive Fairness Criterion

**Limitations of only using counterfactual fairness.** The counterfactual fairness model makes quite strong assumptions regarding linear relation of protected variables (race and sex) and independence of non-protected variables. This might hold in many cases at a population or even sub-population level of considerable sample size and randomized representation. However, most causal questions of fairness in real life are contrastive and many contrastive questions are asked at individual level. The assumptions of linearity and independence that may hold true at population level might not be valid per individual.For example, consider the attribute: number of hours of library access after school hours. This might be a latent variable with an effect on LSAT and GPA and over the entire population can be considered to be independent of race and sex. But at an individual level the scenario can be quite different. In fact at an individual level in culturally heterogeneous populations, race might have an effect on which sex is allowed how much mobility after hours and hence affect access to library or other resources.

Furthermore, the model itself might need to be different between GPA, LSAT vs. FYA. Eg. the combined effect of race and sex for an individual on after hours mobility and library access changes before and after school. Generally, in school a person lives with their parents and they move out during college. In the latter case guardians have less say and hence mobility will vary a lot between the two cases. Thus the same variable might have a different structural function in first year of law school than earlier. Thus we see that at an individual level there are complex non-linear interactions, interactions by proxy and hence independence is not guaranteed. In fact, in essence it might get reduced to fairness by unawareness.

This is illustrated in fig. 1(b), where though FYA is modeled using the latent variables in GPA and LSAT, it might have direct dependency to race and sex, specially by proxy. Now these issues can partially be aligned and conceptualised from earlier equations as shown below. FYA is the observed first year grades of the law school students, whereas $\hat{\text{FYA}}$ is the predicted value according to equation 13. Now we claim that due to the above discussed effects, the residual $\epsilon_F$ may still be complexly dependent on protected attributes at higher orders:

$$\epsilon_F = \text{FYA} - \hat{\text{FYA}}, \quad \epsilon_F = b_F^{''} + f(R,S) + \epsilon_F^{''} \tag{14}$$

Here, $f(R,S)$ is some unknown complex higher order function of the protected variables and $\epsilon_F^{''}$ are the truly independent latent variables. The problem is how to deal with $f(R,S)$, and it is difficult to do that with counterfactual fairness at a population level. We show below that contrastive fairness can be used to mitigate these issues to some extent.

**Role of contrastive fairness with deep neural representations.** To minimise the effect of higher order interactions of protected attributes on the predictor, the problem needs to be recast as a cost function that might be minimised preferably by a neural network that can represent $f(R,S)$ with sufficient abstraction [17]. This is much easier to do in contrastive case at individual level due to its inherent difference formulation to perform comparison. For individual $I$, the predicted FYA is for simplicity recast below where all the parts independent of race and sex are clumped into $f_i'$.

$$\hat{\text{FYA}}_i = f_i(R,S) + f_i^{'} \tag{15}$$

Now when using contrastive logic, we make sure that the decision for individual $i$, is not affected by race and sex at a counterfactual level, that is, we have the condition:

$$J(i) = \hat{\text{FYA}}_i - \hat{\text{FYA}}'_i = f_i(R,S) - f_{i_{A_i \leftarrow \{a_i'\}}}(R,S) + f_i^{'} - f_i^{'} = f_i(R,S) - f_{i_{A_i \leftarrow \{a_i'\}}}(R,S) \tag{16}$$

Here J(i) is the cost function that needs to be minimised with the protected attributes being intervened (using properties of do-calculus from causal inference theory) counterfactually. since the other terms are independent of race and sex they get cancelled. Thus now representing this by a neural network of sufficient depth to approximate the higher order function $f$ and then minimising the cost function for different individuals as data points, mitigates the earlier problems to a large extent.

The same logic can be extended when dealing with contrastive predictions between individuals. Given the above discussion, the present authors advocate caution while evaluating population level fairness using counterfactual fairness and suggest that individual level contrastive fairness criteria should also be taken into account with equal importance.

## 5   Conclusion

We adopt contrastive logic from causal inference to solve the question of algorithmic fairness in machine learning. We lay out the mathematical foundations to achieve this with counterfactual logic at its core. Contratstive questions (why this and not that?) have previously been asked in explainable artificial intelligence. But for the first time we propose contrastive criteria (is it fair to take one decision instead of another particularly differing between two individuals?) in the domain of machine learning. These generic rules can be adopted for various tasks (eg. HR decisions like job recruitment, company layovers, etc) which will explored in future application focused research, to expand on the theoretical framework established in this paper.

# References

[1] Miller T. (2019) Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267: 1-38.

[2] Pearl, J. (2018) Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. *arXiv:1801.04016*.

[3] DeDeo, S. (2016) Wrong side of the tracks: Big Data and Protected Categories. *arXiv:1412.4643*.

[4] Angwin, J., Larson, J., Mattu, & S., Kirchner, L. (2016) Machine Bias. *ProPublica*.

[5] Kusner, M. J., Loftus, J. R., Russell C. & Silva R. (2017) Counterfactual Fairness. *NeurIPS*, pp. 4069-4079.

[6] Lipton, P. (1990) Contrastive Explanation. *Royal Institute of Philosophy Supplement*. 27:247–266.

[7] Miller, T. (2018) Contrastive Explanation: A Structural-Model Approach. *arXiv:1811.03163*.

[8] Halpern, J. Y. & and Pearl, J. (2005) Causes and explanations: A structural-model approach. *The British Journal for the Philosophy of Science*. 56(4): 843-911.

[9] Pearl, J. (2009) Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

[10] Bouwel, J. V. & Weber, E. (2002) Remote causes, bad explanations? *Journal for the Theory of Social Behaviour*. 32(4): 437-449.

[11] Wightman, L. F. (1998) Lsac national longitudinal bar passage study. lsac research report series.

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

[13] Miller, T. (2019) "But why?" Understanding explainable artificial intelligence. *ACM Crossroads*, 25(3): 20-25.

[14] Kusner, M. J., Russell C., Loftus, J. R. & Silva R. (2018) Causal Interventions for Fairness. *arXiv:1806.02380*.

[15] Loftus, J. R., Russell C., Kusner, M. J., & Silva R. (2018) Causal Reasoning for Algorithmic Fairness. *arXiv:1805.05859*.

[16] Glymour, C., & Glymour, M. R. (2014) Commentary: Race and sex are causes. *Epidemiology*, 25(4): 488–490.

[17] Zemel, R. S, Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013) Learning fair representations. *ICML*, 28(3):325–333.

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

[19] Ruben, D.H. (1987) Explaining contrastive facts, *Analysis*, 47(1):35–37.

[20] Silva, R., & Evans, R. (2016) Causal inference through a witness protection program. *Journal of Machine Learning Research*, 17(56):1–53.