Contrastive Fairness in Machine Learning

Tapabrata Chakraborti *(Member, IEEE,)* Arijit Patra, and J. Alison Noble
Department of Engineering Science, University of Oxford, UK

Abstract—Was it fair that Harry was hired but not Barry? Was it fair that Pam was fired instead of Sam? How can one ensure fairness when an intelligent algorithm takes these decisions instead of a human? How can one ensure that the decisions were taken based on merit and not on protected attributes like race or sex? These are the questions that must be answered now that many decisions in real life can be made through machine learning. However research in fairness of algorithms has focused on the counterfactual questions "what if?" or "why?", whereas in real life most subjective questions of consequence are contrastive: "why this but not that?". We introduce concepts and mathematical tools using causal inference to address contrastive fairness in algorithmic decision-making with illustrative examples.

Index Terms—Causal Inference, Counterfactual Logic, Algorithmic Fairness, Machine Learning, Artificial Intelligence.

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) [12] [13]. Machine learning, after all, is a data driven optimal function fitting exercise, thus it has mostly dealt with association, rather than causation [15]. Given the broad use of machine learning algorithms in the modern world, precautions to ensure the fairness of the decision making process of such algorithms is of great importance.

The algorithm may take decisions partly based on such restricted variables like race, gender, sexual orientation, etc learned from historic data having inherent bias [1][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 to name a few [2][18]. Modern machine learning methods should avoid such unethical discriminatory practice [5]. After all, the efficacy of a decision-making process should be based on both accuracy and fairness.

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 [8][9]. We establish the theoretical and mathematical foundations to answer the contrastive 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 [10]. 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 explanations [17] have 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 in 2018 [4][14][10], but for the first time it is formally introduced to algorithmic fairness in this work. Note that the earlier works are in "algorithmic explainability" and not "algorithmic fairness": they use contrastive concepts to explain the decision-making process, but do not address the issue as to whether the decisions taken are fair. 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 with initial results.

The main contributions of this letter are twofold: 1) We present the mathematical foundations to incorporate the ideas of contrastive causal inference to formally address the question as to whether a decision taken by a learning machine is fair, especially when the decision favours one human over another. 2) We then go on to demonstrate the importance of these formulations both through an illustrative thought experiment as well as a real life scenario and dataset with encouraging initial results.

Manuscript submitted: 25-Dec-2019; Manuscript accepted: 03-July-2020; Final manuscript received: 05-July-2020
Authors acknowledge the financial support of UK EPSRC (grant EP/M013774/1 (Seebibyte)) for TC and JAN, and Rhodes Trust for AP.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LOCS.2020.3007845, IEEE Letters of the Computer Society

2 Background Concepts

The present paper combines two relatively new areas of machine learning research, those of algorithmic fairness and causal inference. We provide in this section, a brief collection of underlying definitions and concepts related to both of these areas.

2.1 Algorithmic Fairness

We first define notations which are used throughout this paper. Further details of these standard symbols in causal inference can be readily found in existing literature [11]. 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. Thus latent attributes are the variables that are part of the causal model, but are not directly measured. 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. Thus protected attributes are restricted and the model equations should not incorporate them, both directly and as proxy. Of course, the intuitive but naive assumption in that case would be that an algorithm may be 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, 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.

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) [6] are the backbone of causal inference methods [16]. 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 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 are 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 $P(Y_{Z \leftarrow z'}(U) \mid Z = z)$. Here with the prior knowledge of the values of the variables Z to be z, we compute the counterfactual probability of output variables Y, had the values of Z instead had been z'.

Counterfactual Fairness. Kusner *et al.* [8] 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 observable 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LOCS.2020.3007845, IEEE Letters of the Computer Society

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

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 the same [9].

Kusner *et al.* in their follow-up work [7] address some limitations of their counterfactual formulation. It is shown that counterfactual fairness is susceptible to unfairness due to proxies and unmeasured confounding, that is though it takes care of the direct effect of protected attributes, it might still be influenced by other unmeasured variables that are dependent on protected attributes. We show in the experiments section of this paper that our contrastive fairness formulation is more robust to such unmeasured confounding.

3 Proposed Method: Contrastive Fairness

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 questions 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 to be incorporated. We still use the same counterfactual logic but expand it to fit contrastive cases.

When comparing decisions between two individuals however, we need to make further assumptions. Not only must the decision making processes 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:

$$P(\hat{Y}_{A_i\leftarrow a_i}\,(U_i) = d \mid X_i = x_i, A_i = a_i) = P(\hat{Y}_{A_i\leftarrow a_i'}(U_i) = d \mid X_i = x_i, A_i = a_i) \quad (2)$$

$$P(\hat{Y}_{A_j\leftarrow a_j}\,(U_j) = d \mid X_j = x_j, A_j = a_j) = P(\hat{Y}_{A_j\leftarrow a_j'}(U_j) = d \mid X_j = x_j, A_j = a_j) \quad (3)$$

Next, even if the decision making process itself is fair, for the decision to "make sense", for one individual the decision taken should have a 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:

$$P(\hat{Y}(U_i) = d \mid X_i = x_i, A_i = a_i) > P(\hat{Y}(U_i) = d^{'} \mid X_i = x_i, A_i = a_i) \quad (4)$$

$$P(\hat{Y}(U_j) = d' \mid X_j = x_j, A_j = a_j) > P(\hat{Y}(U_j) = d \mid X_j = x_j, A_j = a_j) \quad (5)$$

Lastly, one 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 presented below.

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

$$P(\hat{Y}_{A_j\leftarrow a_i}\,(U_j) = d' \mid X_j = x_j, A_j = a_j) > P(\hat{Y}_{A_j\leftarrow a_i}\,(U_j) = d \mid X_j = x_j, A_j = a_j) \quad (7)$$

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

3.1 An Illustrative Example of Contrastive Fairness

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 other locations elsewhere in 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. This decision should not be based on such protected attributes like race, sex, religion, etc. It is of great importance to the organisation to be able to justify the decision fairness of the "HR algorithm" and this is where the equations 2 to 7 come into play.

4 Test Case: Law School Success Revisited

Contrastive fairness builds on the principles of causal inference and counterfactual logic. So in order to highlight the contribution of the proposed method, we use the same test case as used in counterfactual fairness [8], that of student success in law school.

4.1 Population Level Counterfactual Fairness

Dataset. The Law School Admission Council dataset [19] 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 3 levels/types of graphical diagrams to model the problem with some assumptions. Of these, we only consider the highest level (called Level 3 in [8]) in this work, which claims to ensure counterfactual fairness under some strong assumptions. The model is presented in Fig 1 (left diagram). The corresponding structural equations have GPA, LSAT and FYA as functions of race (R), sex (S) [5] and independent error terms as follows:

$$\begin{aligned} \text{GPA} &= b_G + w_G^R R + w_G^S S + \epsilon_G, \quad \epsilon_G \sim p(\epsilon_G) \\ \text{LSAT} &= b_L + w_L^R R + w_L^S S + \epsilon_L, \quad \epsilon_L \sim p(\epsilon_L) \\ \text{FYA} &= b_F + w_F^R R + w_F^S S + \epsilon_F, \quad \epsilon_F \sim p(\epsilon_F) \end{aligned} \tag{8}$$

Since R and S are the protected attributes, the independent error terms in the linear model (ϵ_G, ϵ_L, ϵ_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. ϵ_G may be calculated as:

$$\begin{aligned} \hat{\text{GPA}} &= b_G + w_G^R R + w_G^S S, \quad \epsilon_G = \text{GPA} - \hat{\text{GPA}} \\ \hat{\text{LSAT}} &= b_L + w_L^R R + w_L^S S, \quad \epsilon_L = \text{LSAT} - \hat{\text{LSAT}} \end{aligned} \tag{9}$$

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

$$\hat{\text{FYA}} = b'_F + w_F^G \epsilon_G + w_F^L \epsilon_L \tag{10}$$

The authors [8] 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.

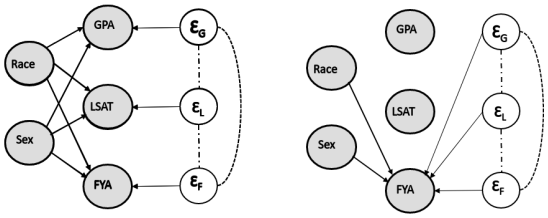

Fig. 1: Left: Causal graphical model of the law school success test case [8]; race and sex are protected variables, ϵ_* are latent variables, GPA, LSAT and FYA are measured variables [9]. Right: Even when the output variable is directly modeled by latent variables, it might still have non-linear dependance on protected attributes especially by proxy.

4.2 Confounding: Counterfactual vs. Contrastive

Even though FYA is modeled using the latent variables in GPA and LSAT, it might have indirect 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 $F\hat{Y}A$ is the predicted value. Now we claim that due to the above discussed effects, the residual ϵ_F may still be complexly dependent on protected attributes at higher orders:

$$\epsilon_F = \text{FYA} - \hat{\text{FYA}}, \quad \epsilon_F = b_F'' + f(R,S) + \epsilon_F'' \tag{11}$$

Here, $f(R,S)$ is some unknown complex higher order function of the protected variables and ϵ_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.

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 $f(R,S)$

TABLE 1: Average accuracy (%) using logistic regression.

Full	Unaware	Counterfactual	Contrastive
0.873	0.894	0.918	0.937

with sufficient abstraction [20]. 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 $f_i^{'}$.

$$\hat{\text{FYA}}_i = f_i(R, S) + f_i^{'} \tag{12}$$

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 cost function:

$$\begin{aligned} \hat{\text{FYA}}_i - \hat{\text{FYA}}'_i &= f_i(R, S) - f_{i_{A_i \leftarrow \{a_i'\}}}(R, S) + f_i^{'} - f_i^{'} \\ &= f_i(R, S) - f_{i_{A_i \leftarrow \{a_i'\}}}(R, S) \end{aligned} \tag{13}$$

We use the new formulations to conduct the same experiments on law school data as done in the original counterfactual fairness paper [8]. The results are presented in Table 1. Note the "Full" means taking all features available and hence has the highest bias. "Unaware" means that the features that are not protected attributes are taken but the fact that they themselves may be biased by proxy by protected attributes is not considered.

Here the cost function needs to be minimised with the protected attributes being intervened counterfactually. Since the other terms are independent of race and sex they get cancelled. 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 can be expected to mitigate the earlier problems to a large extent.

5 Conclusion

We adopt causal inferential logic to address the question of contrastive fairness in machine learning. We lay out the mathematical foundations to achieve this with counterfactual logic at its core. Contrastive 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, 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). We illustrate the idea and also present initial experimental results on real world data.

References

1. Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. (2016) Machine Bias. *ProPublica*.

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

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

4. Dhurandhar, A; Chen, P-Y; Luss, R.; Tu, C-C; Ting, P.; Shanmugam, K.; and Das, P. (2018) Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives, *NeurIPS'18*.

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

6. 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.

7. Kilbertus, N.; Ball, P. J.; Kusner, M. J.; Weller, A.; and Silva, R. (2019) The Sensitivity of Counterfactual Fairness to Unmeasured Confounding. *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 213.

8. Kusner M. J.; Loftus, J. R.; Russell C.; and Silva R. (2017) Counterfactual Fairness. *NeurIPS*, pp. 4069-4079.

9. Kusner, M. J.; Russell C.; Loftus, J. R.; and Silva R. (2018) Causal Interventions for Fairness. *arXiv:1806.02380*.

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

11. Loftus, J. R.; Russell C.; Kusner, M. J.; and Silva R. (2018) Causal Reasoning for Algorithmic Fairness. *arXiv:1805.05859*.

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

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

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

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

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

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

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

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

20. Zemel, R. S; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. (2013) Learning fair representations. *ICML'13*, 28(3):325–333.