Fall 2021 Capstone Project

Progress Report 1

Causality-Informed Fairness

November 2, 2021

Student Contributors:

Junzhi Ge
Mohammed Aqid Khatkhatay
Oscar Jasklowski
Xue Gu
Yue Wang

In Collaboration With:

Dr. Sanghamitra Dutta
Dr. Naftali Cohen

*Student contributors listed alphabetically

Table of Contents

1. Introduction

1.1. Abstract

Machine Learning as well as Artificial intelligence systems are currently being used all the time. Some of these systems are being used for serious decision-making such as advertising[1], immigration[2], trial[3] etc. These decisions if done in a non-conventional or biased manner can have serious, long lasting implications.
While ML and AI can help reduce human interactions in many places and make our lives easier, they might end up being unfair. This happens when there are inherent biases in the data during the training or learning phase in some sensitive features, which lead to unfavorable decisions. ML and AI Fairness are emerging areas which ensure that the output of models is independent or not favorable to a particular class of individuals or others.

1.2. Project Scope and Problem Statement

This Capstone project is a collaboration between the Data Science Institute at Columbia University and JP Morgan Chase and Co. The objective of this project is to leverage, test, experiment and evaluate existing causal techniques to check for unfairness in AI/ML systems. We plan to present an in-depth understanding of existing causal measures and analyze how each one performs on some real-world data or synthetically generated toy data.

Our main focus and outcome of this project will be to research each causal model for unfairness in detail and to come with insights on its inner workings and performance through evaluation metrics. We also plan to thoroughly investigate which causal techniques and metrics should be used in what cases or types of data, whether the data generation process will be known or unknown. Statistically coming up with a framework for causally detecting bias will also be a key outcome in addition to the above goals and objectives.

1.3. Research Framework and Technical Infrastructure

We sought to develop a research framework that would enable us to efficiently explore the literature on causality-based approaches to fairness. Specifically, we sought to explore the following questions:

1. How does a particular fairness technique perform across a variety of metrics?
2. For a given metric, how do various fairness techniques compare?
3. Do certain fairness techniques lend themselves better to certain real-world problems, datasets, or causal models?

To efficiently explore these questions, we created a modular system allowing us to:

1. Swap in datasets (along with their corresponding causal graphs) corresponding to a variety of real-world problems
2. Swap in machine learning models that attempt to achieve fairness by leveraging causality
3. Swap in causality-inspired fairness metrics.

The diagram below i) visually represents this structure and ii) shows **(in dark text)** the elements we built for this first report (elements in grey text will be explored in subsequent reports). Our github repository can be found here.

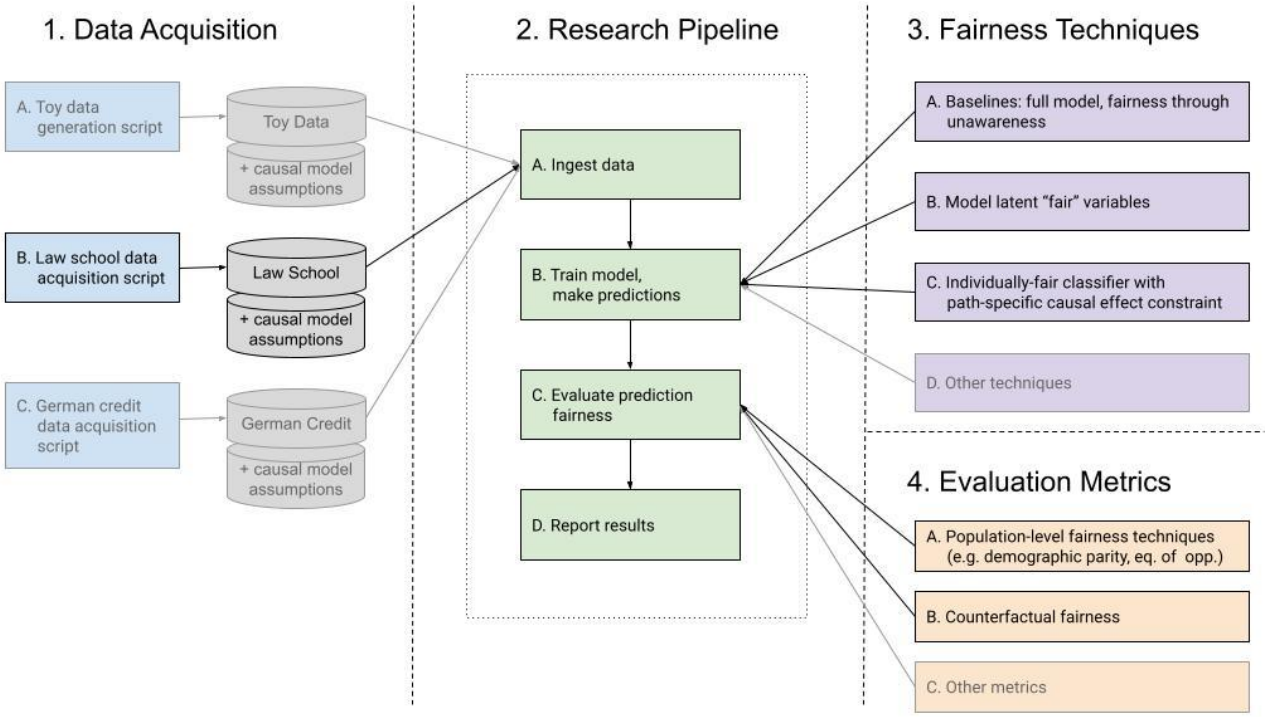

Figure 1: We diagram a modular system that will support our goal of comparing a number of fair inference techniques using a variety of fairness evaluation metrics. Additionally, this modular system allows datasets to be substituted in a straightforward manner, allowing us to ultimately assess fairness techniques for a variety of real-world use cases.

2. Literature Survey

In this section we review six prominent research papers in the field of causality-informed fairness. For each paper, we summarize the techniques and/or metrics discussed.

2.1. Counterfactual Fairness

This paper titled Counterfactual Fairness [4] gives a general introduction and approach towards Counterfactual Fairness. The paper discussed Fairness Through Unawareness (FTU), which means we need to analyze the effect of protected variables on outcome after we remove it. This way we would expect to achieve some fairness. The paper also provided a few approaches for achieving a fair model by manipulating data, attributes or introducing new relationship.It defined counterfactual fairness carefully with theorems, equations and graphs. It also provides the metric that allows us to evaluate counterfactual fairness.

This paper had a few advantages. One was that the basic approach generalized that unfairness can be easily done by removing variables. The paper introduced a general approach, and suggested several ways(in paper) to achieve fairness. It also had a general algorithm. The paper had a few disadvantages. One of them was that some approaches may not be realistic and counterintuitive. In a few cases the algorithm may sacrifice accuracy by achieving fairness. The paper uses the Law School data set. The paper also has an official Implementation. NYC stop-and-frisk data was also used, but just for a visualization example.

2.2. Path-specific Counterfactual Fairness

The paper titled Path Specific Counterfactual Fairness [5] uses the notion of directed edges as a means of identifying the unfair pathways. This paper captures the intuition that a decision is fair on an individual level if it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute along the unfair pathways were different.The paper proposes that unfairness in the indirect unfair pathways could be avoided by eliminating the unfair information contained in the descendants induced by the sensitive attribute, while retaining the remaining fair information.

With restrictive functional assumptions in the data generating process, the proposed path-specific counterfactual fairness (PSCF) method guarantees individual level fairness. Specifically, the latent variable approach allows for more individual specific information to be retained. However, it was only applicable to complex non-linear models and could not be used for general structural causal models.

The paper used the Berkeley Admission dataset as an example, and conducted experiments using UCI adult dataset and UCI German Credit dataset. No official or unofficial implementation of PSCF was found at this point.

2.3. Fair Inference on Outcomes

The paper titled “Fair Inference on Outcomes” [6] is majorly based on fair outcomes of statistical outcomes. It can be summarized in the following quote, “The presence of discrimination can be formalized in a sensible way as the presence of an effect of a sensitive co-variate on the outcome along certain causal pathways”. It presented a technique based on mediation inference. Fairness is somewhat related to reducing unfairness through the unfair pathways.

This can be formulated as a maximization problem. It also considers fairness as a domain specific issue.

The paper had the advantage that it simplified fairness as a maximization problem. However, there was a disadvantage that in cases where the path specific was not easily identifiable.
The fair inference approach was demonstrated on the UCI adult dataset and the Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS dataset.

2.4. Avoiding Discrimination through Causal Reasoning.

The paper titled “Avoiding Discrimination through Causal Reasoning” [7]. This paper establishes that discrimination is causal and can be utilized to draw inference from causal techniques. Causal graphical models can be used extensively for detecting and eliminating unfairness. Due to limitation of the observational criteria in avoiding discrimination in the system, the paper defines the notions of unresolved discrimination and proxy discrimination, and develops separate step-by-step procedures to avoid them.

The paper had a few good advantages. One of them is that there was clear formalization on previous impasse on observation criteria. The paper also had enough generalization that allows for natural removal procedures, without tedious evaluation of path-specific causal effects. However, the proposed algorithm requires very specific linearity assumptions about the underlying causal model generating the data. And it also fails to highlight the correlation between fairness and an individual proxy variable.
The paper lacked implementation and experimentation.

2.5. Fairness in Decision-Making — The Causal Explanation Formula

The paper titled “Fairness in Decision-Making-The Causal Explanation Formula” [8] introduced a framework for causal inference. It also proved the causal explanation formula that quantified the effect of various causal techniques with respect to statistical disparities. It measured the trade-off between procedural fairness and outcome fairness. Furthermore, it also measured how much each mechanism was contributing to the total variation.

This paper had the following advantages. The first one being that the approach can be applied to nonlinear systems, which was a drawback in some other papers. It also helped researchers design preparatory policies. Finally, the paper was also applicable to any kinds of discrimination and types of variables. The paper however had a few disadvantages. One was that it generalized or relaxed ways of decomposing total effects. It also only tells one how variation in one attribute affects another. However, there was no controlled experiment understanding the effects of X. Finally, the paper lacked implementation and did not mention the official dataset that was used for experimentation.

2.6. Individually Fair Path Specific Causal Effect

The paper titled “Learning Individually Fair Classifier with Path-Specific Causal-Effect Constraint” [9] defines probability of individual *un*fairness (PIU) and its upper bound, estimated from data, to impose a fairness constraint in prediction.

Specifically, this approach places an upper limit on individual *un*fairness and constrains it to be near zero, which is the approach to achieving individual fairness. This method also makes no restrictive functional assumptions on the data generating process, making it more practical than PSCF to use in real-world problems. However, the method requires a specified causal graph with no non-convex functions for optimization purposes. Another drawback is that there is an inherent tradeoff between individual fairness and prediction accuracy.

The paper was implemented officially and experimented on German Credit dataset and Adult dataset.

2.7. Literature Summary

Table 1: Survey of in-scope papers. We summarize whether the aforementioned papers describe metrics or techniques, whether an implementation is available, and the pros and cons of each.

Paper	Advantages / Disadvantages	Metrics or Techniques?	Implementation available?
Counterfactual Fairness (2.1)	A: Basic approach can be easily done by removing variables. Give specific definitions and examples. It has a general approach, and suggests several ways(in paper) to achieve fairness. Has a general algorithm D: Some approaches may not be realistic and counterintuitive. May sacrifice accuracy by achieving fairness.	Metrics + Technique (latent variable inference)	No (implemented this from scratch)
Path-Specific Counterfactual Fairness (2.2)	A: Latent variable approach allows for more individual specific information to be retained. D: Applicable to complex Non-Linear models. Requires causal model underlying the data generation process.	Technique	No

Fair Inference on Outcomes (2.3)	A: “It is possible to construct examples, where some causal paths from a sensitive variable to the outcome are intuitively discriminatory.” D: Fails for cases where path specific was not easily identifiable	Technique	Yes (see here)
Avoiding Discrimination through Causal Reasoning (2.4)	A: Clear formalization on previous impasse on observation criteria Enough generalization that allows for natural removal procedures, without tedious evaluation of path-specific causal effects D: Algorithms require very specific linearity assumptions about the underlying causal model generating the data	Technique	No
Fairness in Decision-Making: The Causal Explanation Formula (2.5)	A: Can be applied to nonlinear systems. Can help design preparatory policies. Applicable to any kinds of discrimination and types of variables D: A generalized or relaxed way of decomposing total effects. Only tells how variation in one attribute affects another. Not a controlled experiment, understanding effects of X	Metrics	No
Individually-Fair Path Specific Causal Effect (2.6)	A: Individual level fairness is guaranteed. The paper uses Stochastic gradient Descent, which works perfectly for convex functions hence convergence is guaranteed. D: Estimating the upper and lower bounds without a causal graph is not possible. Stochastic Gradient Descent was used, which fails to converge for non-convex functions. Individual fairness occurs at the cost of prediction.	Techniques	Yes (see here)

3. Exploratory Data Analysis and Visualization

3.1. Law School Admissions Data Set

3.1.1 Overview

The Law School data set that we use is obtained by a survey conducted by the Law School Admission Council (LSAC) across 163 law schools in the United States. It contains data on admissions of 21791 law school students along with their GPA (Grade Point Average), race, sex, LSAT (Law School Admission Test) scores, region_first, ZFYA, sander_index and first_pf. The columns and their values can be summarized in the following table.

3.1.2. Features

Table 2: Law School Dataset Feature Summary. Below are a variety of attributes for each column in the law school dataset.

Features	Meaning	Data Type	Feature used in analysis?
race	Race of the applicant consisting of White, Hispanic, Asian, Black Other, Mexican, Puertorican or Amerindian	categorical	Yes
sex	1: Female 2: Male	enum	Yes
LSAT	Law School Admission Test Score	float	Yes
UGPA	Undergraduate Grade Point Average (on a scale of 5)	float	Yes
region_first	The state from where the applicant comes. In our data the following values: GL, MS, NE, NG, SE, Mt, FW, SC, MW, NW and PO	float	No
ZFYA	First Year Average Grade (Standardized)	float	Yes

sander_index	Indicates academic index gap	float	No
first_pf	Undefined in resource	bool	No

3.1.3. Heatmap

Figure 2 below shows a heat map of all the continuous variables in the dataset including the target variable ZFYA.

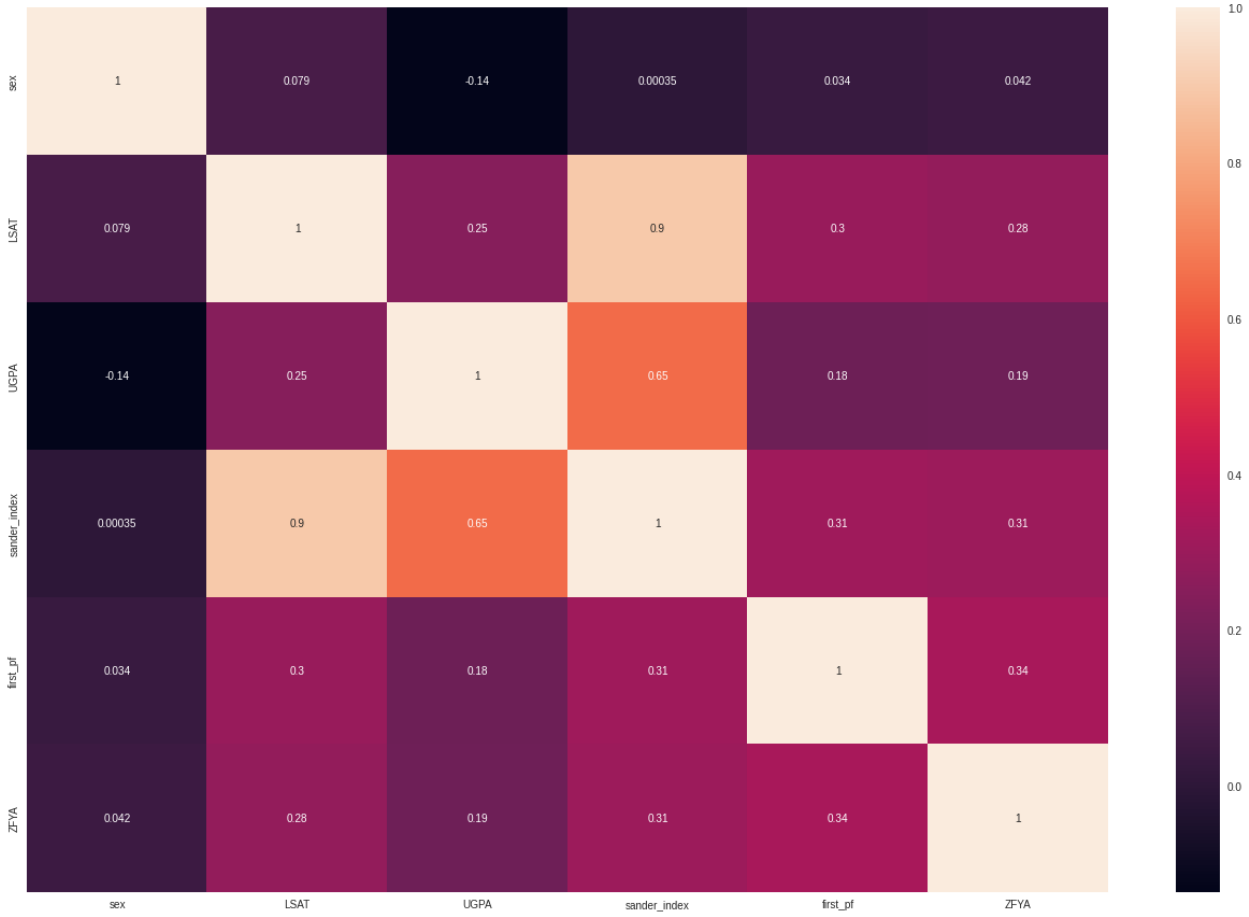

Figure 2: Showing heatmap of all continuous variables in the dataset

It can be seen from the heatmap in Fig 2 that most features in our dataset are positively correlated with each other, but not with the target variable ZFYA which is standardized.

3.1.4. Distribution Plots

The following figure highlights the distribution of sander_index, UGPA, LSAT and ZFYA using three plots.

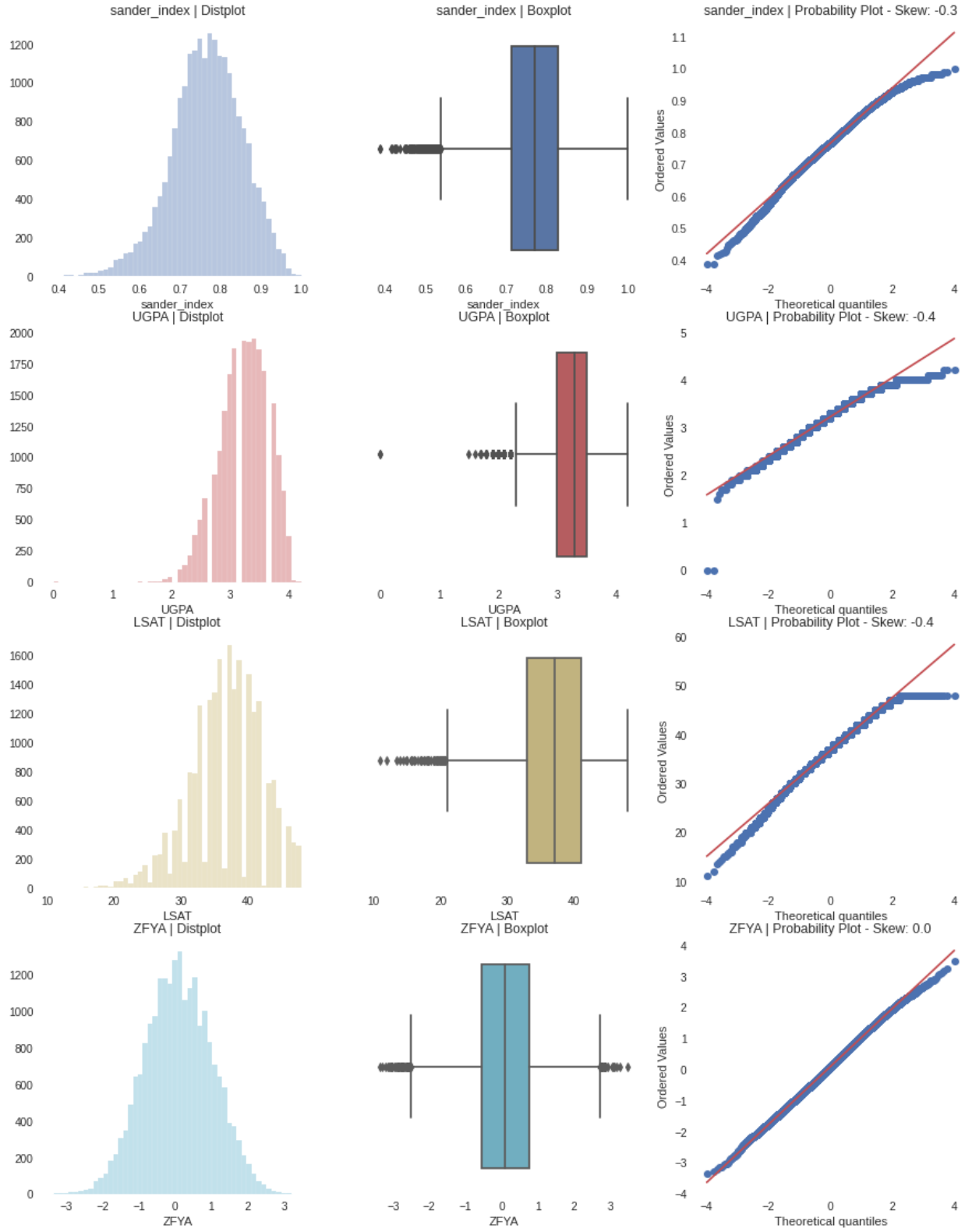

Figure 3: Distribution Plots of Law School Dataset

It can be seen from Fig 3 that ZFYA is standardized around 0. There is hardly any skewness of the ZFYA values, unlike the other terms being analyzed. The box plots help to show the quantile distribution of the attributes present. These plots reiterate the presence of some dependence in the data generation process.

3.1.5. Continuous Variables over categorical

The following figure highlights the variation of continuous variables with respect to the categorical variables on the target variable.

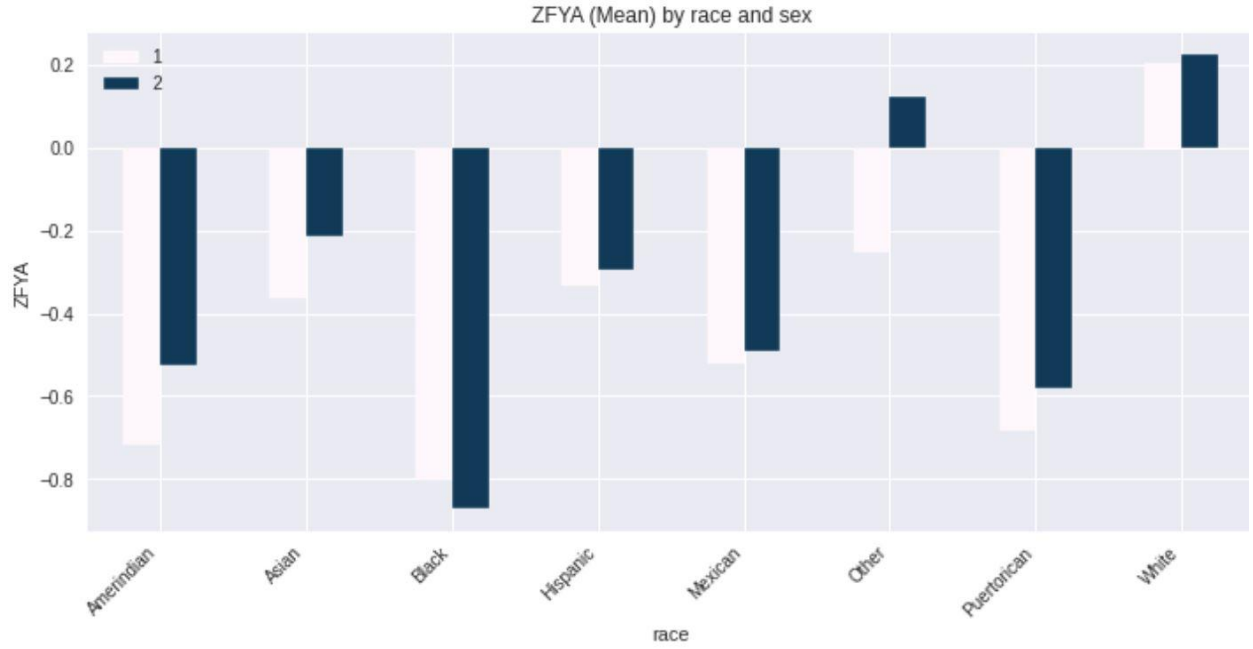

Figure 4: Mean ZYFA by race and sex (1=Female, 2= Male)

Observationally, it is evident that females have a negative ZFYA score compared to males. Regionally mean ZFYA was also plotted to see if any categorical variables were affecting it.

3.1.6. Violin Plots

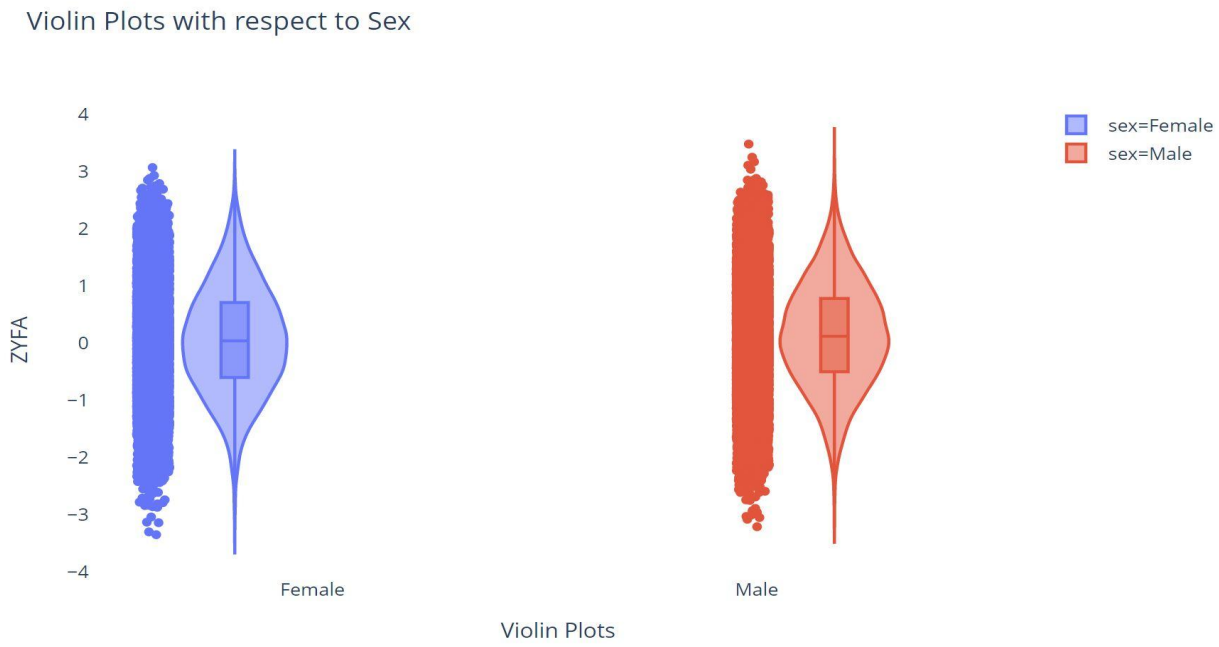

Figure 5: Shows the violin plot comparison for both males and females, around the ZFYA standardized score. We observe that both the plots are similar in style to each other. Males have a few outliers more compared to females when it comes to the ZFYA score.

Figure 6: Shows the violin plot comparison for both race, around the ZFYA standardized score. We observe that both the plots are similar in style to each other. Whites have fewer outliers

compared to other races (including others) when it comes to the ZFYA score. Whites also perform better compared to other races in this survey.

Figure 7: Shows the violin plot comparison for both race with respect to races, around the ZFYA standardized score. We observe that both the plots are quite insightful in style to each other. This graph provides evidence that there is some relationship between sex as a protected attribute rather than race.

4. Implementation Approach and Results

4.1. Causality Assumptions and Graph for the Law School Dataset

The causal model we defined is shown below:

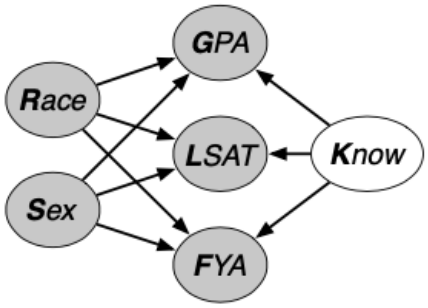

Figure 8: This depicts our assumed causal model for First Year Average (FYA; a measure of a student's success in their first year of law school). We assume that Race and Sex are the two protected attributes that effect GPA, LSAT and FYA, and Knowledge is the latent variable that's not provided in the data but is also effecting the three scores. We would use the available data to learn the latent knowledge attribute (more details in 4.3.1).

4.2. Baseline Model

We will use the following definition in this report. We consider A as the set of protected attributes of a sample, those are variables that should not be considered while we are trying to learn a model, since they may result in discrimination. Thus, protected variables should not affect our final outcomes no matter what our problem setting is. We consider X as the other observable attributes of any individual, U as the set of implicit latent attributes, and Y as the outcome we want to predict. Finally, we consider $\widehat{Y}$ as a predictor, a random variable that depends on A, X and U. We would like to learn this through the model we build.

4.2.1 Full model

Full model method is to use all available attributes to predict the response variables. In our case, the Law school data, X attributes would be race, sex, LSAT, and UGPA. Among all those variables, we consider race and sex as protected attributes, since we don't really want sex and race to have bias on FYA scores. And our response would be FYA, which is the first year average score. We would use linear regressions to train and predict the possibility of success, while each sample has different protected attributes, race and sex.

4.2.2 Unaware model:

When people are trying to predict outcomes Y, it is hard to define what is fair and unfair. As in this problem setting, we use the Law School dataset to give a demonstration on a simple approach for a fair model.

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

When analyznig Law school dataset, as described above, we consider race and sex as protected attributes. We train the samples with linear regression without the protected attributes this time. We take LSAT and UGPA as X and analyze their effect on FYA scores. Furthermore, we train the data, then make prediction for FYA scores, we can compare the effect of protected variables with full models by checking the predicted accuracy of separate protected attributes such as sex.

Concerning protected attributes, we are trying to explore if the model is fair while we change the protected variables, since protected attributes should not affect the actual outcome Y. This

would be a good base model except the fact that there would be some implicit relationship between protected attributes and other attributes, which may not be shown by the data.

4.3. Advanced fairness techniques

4.3.1 “Latent variable inference” model

Would the model be considered fair if we can use some latent variables which are not related to the protected variable? Prediction based on the latent variable is the ‘latent variable inference’ model: we try to find some latent variables which are not descendents of any protected attributes, in this way , we eliminate any effect protected attributes would result. In our approach, we introduce a model using a latent variable called ‘knowledge’. We assume that knowledge is not a descendant of sex or race which means those protected attributes will not affect the latent attribute knowledge we define. At the same time, we assume that knowledge is correlated with scores. Since we have no data about knowledge, our approach is to use the prior data and distribution of scores, sex, and race, we can thus learn a posterior distribution of knowledge. After that, we generate random samples from its distribution. In this way, we learn latent attribute knowledge from available data. Lastly, we try to use this latent variable knowledge, as well as LSAT, UGPA score to train and predict the first-year average score using linear regression.

4.3.2 “Individually fair path-specific” model

While naively removing protected attributes could also remove unfairness in the prediction (i.e. unawareness model), it is usually insufficient in complex cases where multiple unfair pathways are present. An example given in [9] shows that in the hiring process for a physically demanding job, it is unfair to hire candidates based on their sex, which is a direct cause for hiring. It is also unfair to hire candidates based on how many children they have, which is affected by sex. Both pathways are unfair, but removing protected attribute sex cannot remove unfairness through the latter pathway.

In this subsection, we will use a method that can guarantee fairness for each individual via optimizing the upper bound of probability of individual unfairness (PIU) defined in [9] to 0. Although there is a slight cost of prediction accuracy, this method makes no impractical functional assumptions on the data, which makes it very competitive in solving real-world counterfactual fairness problems. Specifically, our method has two main competitors. [5] has a similar goal, but their path-specific counterfactual fairness (PSCF) method achieves fairness by enforcing a restricted data generating process. In contrast, [10] does not require restrictive functional assumptions, but it fails to achieve counterfactual fairness at the individual level.

For unfair pathways π in a causal graph and potential outcomes $Y(A \Leftarrow 0)$, $Y(A \Leftarrow 1)|\pi \in \{0, 1\}$, probability of individual unfairness (PIU) is defined by $P(Y(A \Leftarrow 0) \neq Y(A \Leftarrow 1)|\pi)$. In [9], this definition of PIU allows us to achieve individual level fairness without conditioning on X (feature vector) for each individual. However, we fail to minimize PIU in that it requires the joint distribution of $Y(A \Leftarrow 0)$, $Y(A \Leftarrow 1)|\pi$, which is

impossible to compute without a specified structural equations model. Instead, [9] proposes an upper bounds on PIU which can be estimated from the data set. By minimizing the upper bound on PIU, PIU is in turn minimized. For binary potential outcomes $Y(A \Leftarrow 0)$, $Y(A \Leftarrow 1)|\pi \in \{0, 1\}$, the joint distribution $P(Y(A \Leftarrow 0) \neq Y(A \Leftarrow 1)|\pi)$ is upper bounded by $P(Y(A \Leftarrow 0) \neq Y(A \Leftarrow 1)|\pi) <= 2P'(Y(A \Leftarrow 0) \neq Y(A \Leftarrow 1)|\pi)$, where $P'(Y(A \Leftarrow 0)\, Y(A \Leftarrow 1)|\pi) = P(Y(A \Leftarrow 0))\, P(\, Y(A \Leftarrow 1)|\pi)$. In other words, for any distributions $P(Y(A \Leftarrow 0))$ and $P(\, Y(A \Leftarrow 1)|\pi)$, the resulting PIU is at most twice the PIU for independent joint distribution P'.

To compare with previous models, specifically fairness through unawareness and latent variable model (with constructed “Knowledge” posterior), we will need to understand PIU in the context of the Law School data set and its induced $P(v)$ and causal graph. Unfortunately, due to the insufficient information provided in the implementation repository, we cannot yet apply PIU in the Law School dataset, and thus no direct comparisons about prediction accuracy and fairness is possible at this point. However, for a synthetic dataset with clearly defined $P(v)$ on observed variables and $P(u)$ on unobserved confounders, we are able to obtain both PIU and prediction accuracy.

In our next report, we will try to reach out to the authors for an explanation in the implementation, specifically their experiment dataset, to gain a better understanding of their approach. We will further make the preliminary assumptions, for simplicity, no unobserved confounders U will be considered in the structural causal model G for the Law School data set. In addition, we will only consider “SEX” to be the protected attribute in our modelling, and with “RACE” removed.

4.4. Fairness Evaluation Metrics

There are a number of methods to evaluate the fairness of a prediction technique with respect to sensitive variables. Below, we explore two classes of fairness evaluation methods:

1. Population methods: *Demographic parity* and *equality of opportunity*
2. Individual-level methods: *Counterfactual fairness*

Below, we define each metric, compare their theoretical advantages, and evaluate our prediction techniques on each method.

4.4.1. Population metrics: Demographic Parity and Equality of Opportunity Definitions

A predictor Y satisfies *Demographic Parity* (with respect to a sensitive attribute A) if:

$$P(\widehat{Y} \mid A = 0) = P(\widehat{Y} \mid A = 1)$$

In practice, we don't expect these quantities to be exactly equal, but we can define our *Demographic Parity* metric as the absolute value of the difference between them, as below:

$$DP = \left| P(\widehat{Y} \mid A = 0) - P(\widehat{Y} \mid A = 1) \right|$$

Similarly, a predictor Y satisfies *Equality of Opportunity* (with respect to a sensitive attribute A) if:

$$P(\widehat{Y} = 1 \mid A = 0, Y = 1) = P(\widehat{Y} = 1 \mid A = 1, Y = 1)$$

As with *Demographic Parity*, we don't expect these quantities to be exactly equal, but we can define our *Equality of Opportunity* metric as the absolute value of the difference between them, as below:

$$EO = \left| P(\widehat{Y} = 1 \mid A = 0, Y = 1) - P(\widehat{Y} = 1 \mid A = 1, Y = 1) \right|$$

4.4.2. Casual Metrics through Counterfactual Fairness

Support attribute A is the protected attributes, and X is all other attributes in the dataset, we define counterfactual probability as the probability of the predictor falling into one class if attribute A would belong to a different class.

$$P(Y_{A = a'}(U) = y \mid X = x, A = a)$$

Given that, counterfactual fairness is defined as: Prediction $\widehat{Y}$ is counterfactually fair if under any context X = x and A = a,

$P(Y_{A = a}(U) = y \mid X = x, A = a)$ = $P(Y_{A = a'}(U) = y \mid X = x, A = a)$ for all y and for any value a' attainable by A [4].

4.4.3. Metrics Tradeoffs

It's important to note that these metrics may be incompatible insofar as, when comparing two classifiers, there is no guarantee that one classifier will outperform the other across all metrics. As a result, we need to make principled decisions in which metric we optimize for based on our real world problem. Table 3 summarizes the tradeoffs between our selected metrics.

Table 3: Metric Tradeoff Summary. Below we outline the inherent advantages and disadvantages of each metric. One can see that each captures a slightly different notion of "fairness," so, for a given dataset and model pair, the choice for which to optimize will have important consequences, outlined below.

Metric	Main Advantage	Main Drawback	Based on

			Causal Inference?
Demographic parity	Because it guarantees that predicted proportions for protected groups mirror the population proportions for protected groups, you **promote a feedback loop that leads to proportional representation**.	Because the metric is independent of the target variable, **in cases where protected groups and unprotected groups have different proportions of qualified candidates** (e.g. in the context of loan approval), **you will necessarily reject the optimal classifier** (i.e. you'll reject some qualified candidates and fail to reject unqualified candidates).	No
Equality of opportunity	This guarantees equality of treatment among subpopulations. In other words, the same proportion of qualified candidates in any subgroup will be selected by a classifier, meaning **it allows for an optimal classifier**.	This will not address unfairness in the long run because it does not deal with possible discrimination outside of the model (e.g. racial biases can result in a lower proportion of qualified individuals). Over time, **optimizing for this metric can exacerbate systemic biases that lead to disproportionate success outcomes in the first place**.	No
Counterfactual fairness	This metric can be computed at the individual level, meaning that **fair inference for individuals can be captured by this metric**. While other techniques can be “fair” on average across a population, the consequences of unfair prediction on an individual may be severe, depending on the application.	**This metric is not trivial to calculate under certain circumstances**. Due to assumptions implicit in our causal graph, we were able to somewhat trivially calculate this metric by swapping the sex (i.e. generating counterfactual test rows) for individuals. Under other circumstances, sophisticated sampling techniques are needed.	Yes

4.5. Results:

Table 4: Performance across all three metrics for all three models, in addition to classification accuracy:

Performance of models with respect to fairness metrics, treating gender as the sensitive attribute.					
Model	**Demographic Parity**	**Equality of Opportunity**	**Counterfactual Fairness**		**Prediction Accuracy**
			Female	**Male**	
Full Model	0.0409	0.027	0.1036	0.0885	63%
Unaware Model	0.0109	0.0205	0.2628	0.2018	60%
Latent Variable Inference Model	0.0068	0.0025	0.0644	0.1976	55%

4.5.1. Analysis of Demographic Parity and Equality of Opportunity with Respect to Gender

With respect to demographic parity (DP) metric, we expect the *full model* to have the worst performance (i.e. greatest value for DP), which is borne out in the results in table 4. The full model is a simple logistic regression classifier that makes no effort to reduce the influence of sex on success. Therefore, if sex directly predicts the success of a student, the relationship will be captured by the full model, and will negatively impact DP. Put another way, we expect a direct tradeoff between DP and prediction accuracy (i.e. full model has the best prediction accuracy and the worst DP score).

By the same logic, we expect the *full model* to perform well relative to the *unaware model* on the equality of opportunity (EO) metric. EO should reward an optimal classifier, and we would expect the full model to be a more optimal classifier than the unaware model (which is unaware of gender). It's surprising that this is not the case (slightly better performance for the unaware model). We hypothesize that there may be an unobserved latent variable that is a descendent of sex and a parent of FYA. If sex and that variable had opposite influences on FYA for females, the removing sex could actually serve to improve EO for females. This type of confounding is possible in real-world datasets where our causal model may be a dramatic oversimplification of reality. To mitigate this risk, as a next step, we will create a simple data generating process (i.e. no possible confounding) to test the tradeoffs in DP and EO.

Finally, we observe that the latent variable inference model performs best with respect to both DP and EO. Theoretically, a classifier can perform well on both of these metrics under certain circumstances (e.g. protected and non-protected groups appear in comparable numbers in the test data, and the two groups have comparable success rates). Males only outnumber females ~1.3/1, which is relatively small, so the conditions for a single model performing best on both of these metrics do potentially exist. Therefore, we believe this result is not an error, and that this

is an indication of our latent variable inference model is more fair than simpler models with respect to DP and EO.

4.5.2. Analysis of Counterfactual Fairness with Respect to Gender

To evaluate the fairness of the predictor, we calculate the probability to succeed in college after alternating each sample's sex and compare this counterfactual probability with the original probability.

Specifically, for the latent variable inference model, the fair K attribute was relearned with the previous approach. But since our causal model defined the causal effect of sex on GPA, LSAT and FYA, we studied the linear relationship between sex and the three variables, and use it to predict GPA, LSAT score with counterfactual gender. Then we use these counterfactual values to implement the full model and unawareness model.

Based on the result, there are several observations:

a) As the number of attributes used for prediction decreases and higher expected level of fairness increases , the accuracy of the model decreases. There could be some tradeoff between fairness and accuracy.
b) For both female and male groups, the predicted probability would be higher in a counterfactual situation where each individual is under a different gender identity.
c) For males, the most fair model would be the full model, while for females, the most fair model would be the latent inference model.
d) The full and unaware model where gender is used as an attribute for prediction, the difference between counterfactual probability and original probability is similar between female and male.
e) Using the unaware model, the difference is higher for females, while the difference is higher for male using the latent variable inference model.

According to these observations, different models have different fairness effects on gender. The big difference of fairness between females and males by the latent variable inference model is out of our expectation. We would like to explore possible reasons behind it and ways to explain it. Furthermore, since we don't know how the data is generated, we would need to apply the metric to a toy dataset and evaluate if they are valid enough to reflect fairness.

4.5.3. Counterfactual fairness evaluation on race.

To evaluate race as a protected attribute, we refer to the exploratory analysis done in Figure 4 Figure 4 shows that white students have a higher standardized ZFYA compared to the other races in the Law Schools Admissions Data Set.

The violin plots in figures Figure 6 and Figure 7 show that sex is the only protected attribute in the dataset which should be evaluated in a counterfactual world rather than race which barely has any effect on the ZFYA target variable.

5. Conclusion and next steps

5.1. Brief summary

So far, we have surveyed several papers on counterfactual fairness, specifically through the unawareness model in [4], the latent variable model in [4] and the PIU method in [9]. We implemented and applied two unfair baselines: a full model, a fairness through awareness model, and an advanced latent variable model on a real-world Law School dataset, where our primary task is to make fair predictions with respect to SEX and RACE on student success. Furthermore, to evaluate fairness, we also implemented metrics including demographic parity, equality of opportunity, counterfactual fairness and obtained some expected results on our models. Particularly, while the full model has the best predicting power, it does not perform well in fairness compared to the unawaress model and the latent variable model. However, the latent variable model performs surprisingly poorly on some of the fairness evaluation metrics. Thus, part of our next steps would be to validate the implementation of our metrics and test them on a small toy dataset. In hindsight, using a fully specified toy dataset should have been our starting point in understanding both the fairness methods and evaluation metrics.
Our work up to this point is also leading to the implementation of a more advanced fairness technique: prediction with PIU constraint, which guarantees individual level fairness without restrictive functional assumptions on the data. As our baselines and the latent variable model cannot remove unfairness when multiple unfair pathways exist.

5.2. Next steps

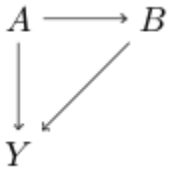

Figure 9: An illustration of a simple casual graph where protected attribute A affect Y in both $A \to Y$ and $A \to B \to Y$ pathways. A is a binary protected attribute. Y is a binary predictor.

As mentioned in our brief summary, a natural next step is to generate a fully specified toy dataset with one binary protected attribute that affects the predictor in multiple unfair pathways. A naive construction is shown in figure 9. With a known data generating process and distribution $P(A),\ P(B),\ P(Y)$ that induces the casual graph in figure 9, it will be easier to compute all necessary conditional probabilities and joint probability distribution. In this way, we will be able to more easily validate our implementation for both fairness techniques and metrics. In addition,

this is a minimally sufficient causal model that allows us to apply advanced fairness techniques and compare results with our baselines without extra complexities.

Furthermore, we have reflected on our result up to this point that warrants next steps. As we get some unexpected result, I think it would be a good idea to actually re-read the paper again and make sure that we build the model in the correct way, also it might be helpful to make sure that we understand the relationship among data correctly. For example, does it make sense to define a parent-child relationship between two attributes, it's always a good idea to draw a relationship map among attributes. Lastly, if we actually get an unexpected result, we can try to work on another dataset to test if our implementation is correct, since fairness cannot be easily defined in a standard way.

As we find baseline models, latent variable models and counterfactual models, there's still other models mentioned in the paper, we can take a review and apply it to our dataset. Also, if we work on a more complicated dataset other than law school dataset, can we achieve similar results? How will the performance of the new model compare with our old one?

As mentioned in section 4, since we don't know how the data is generated, we would like to apply the metrics to a toy dataset and evaluate if they are valid enough to reflect fairness of each model. Once the validity of metrics is confirmed, we would like to investigate deeper into the models and causes of discrepancies between gender groups in terms of fairness probability.

6. Contributions

Index	Task	Owner
1.1.	Introduction	mk4427
1.2.	Research framework, data pipeline design	ovj2101
2	Literature survey	Full team
3	Data exploration	mk4427
4.1.	Implementation: full, unaware, latent variable models	yw3576, jg4281
4.2.	Implementation: individually-fair path-specific approach	xg2353
4.3.1.	Metrics: demographic parity	ovj2101
4.3.2.	Metrics: equality of opportunity	ovj2101
4.3.3.	Metrics: counterfactual fairness	yw3576, mk4427
5.1.	Conclusions	xg2353
5.2.	Next steps	xg2353, yw3576, jg4281, ovj2101
7	References	xg2353, mk4427
--	Proofread, edit, submit	ovj2101, xg2353

7. References

[1] Sanjeev Verma, Rohit Sharma, Subhamay Deb, Debojit Maitra, Artificial intelligence in marketing: Systematic review and future research direction, International Journal of Information Management Data Insights,Volume 1, Issue 1, 2021, 100002, ISSN 2667-0968, https://doi.org/10.1016/j.jjimei.2020.100002

[2] Ana Beduschi, International migration management in the age of artificial intelligence, *Migration Studies*, 2020;, mnaa003, https://doi.org/10.1093/migration/mnaa003

[3] Ulenaers, Jasper. "The Impact of Artificial Intelligence on the Right to a Fair Trial: Towards a Robot Judge?" *Asian Journal of Law and Economics*, vol. 11, no. 2, 2020, pp. 20200008. https://doi.org/10.1515/ajle-2020-0008

[4] Matt J Kusner et al. “Counterfactual fairness”. In: arXiv preprint arXiv:1703.06856 (2017).

[5] S. Chiappa and T. Gillam. Path-specific counterfactual fairness. arXiv:1802.08139, 2018.

[6] Nabi, R., Shpitser, I.: Fair inference on outcomes. CoRR abs/1705.10378 (2017).

[7] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Scholkopf. Avoiding discrimination through causal reasoning. arXiv preprint arXiv:1706.02744, 2017.

[8] Zhang, J., & Bareinboim, E. (2018). Fairness in Decision-Making — The Causal Explanation Formula. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/11564

[9] Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima. Learning individually fair classifier with path-specific causal-effect constraint. arXiv preprint arXiv:2002.06746, 2020.

[10] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In AAAI, pages 1931–1940, 2018. https: //github.com/raziehna/fair-inference-on-outcomes.