FAIR REPRESENTATIONS IN THE DATA DOMAIN

OLIVER TIMOTHY THOMAS

A thesis submitted for the degree of Doctor of Philosophy.

School of Engineering and Informatics

University of Sussex

March 2022

ABSTRACT

Algorithmic fairness is a multi-faceted topic which is of significant consequence to a diverse range of people. The issue that this thesis investigates is a fairness-specific instance of a yet even broader concern — that data can be biased due to spurious correlations. Machine learning models trained on such data learn to exploit these spurious correlations that do not hold in the test distribution. When spurious correlations are found with respect to protected demographic attributes, trained models could be biased towards certain subgroups or populations. A promising approach to counteract biased data is by producing a fair representation as a pre-processing step. The main drawback, however, of existing fair representation learning approaches is that the data often become obscured when projected into an uninterpretable latent space, making intuitive assessment difficult. Noticing that the domain the data resides in is often interpretable, with the structure providing richer information that is easier to understand on a per sample basis, I develop fair representations in the data domain. These convey additional per-sample information that can be easily shared and explained to system designers and stakeholders. This thesis investigates three aspects of fair representations in the data domain. Firstly, I demonstrate a novel application of fair representations to generate counterfactual samples in the data domain. The aim of this application is to promote positive actions to address discrimination in an already existing system; Secondly, I develop a method to produce fair representations in the data domain based on statistical dependence principles; Lastly, I take this approach further, introducing two further methods to achieve fair representations in the data domain based on adversarial learning.

PUBLICATIONS

Some ideas and figures in this thesis have appeared previously in the following publications:

Quadrianto, Novi, Viktoriia Sharmanska and Oliver Thomas (2019). ‘Discovering Fair Representations in the Data Domain’. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, pp. 8227–8236. DOI: `10.1109/cvpr.2019.00842`.

Kehrenberg, Thomas, Myles Bartlett, Oliver Thomas and Novi Quadrianto (2020). ‘Null-Sampling for Interpretable and Fair Representations’. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox and Jan-Michael Frahm. Cham: Springer International Publishing, pp. 565–580. ISBN: 978-3-030-58574-7.

Thomas, Oliver, Miri Zilka, Adrian Weller and Novi Quadrianto (2021). ‘An Algorithmic Framework for Positive Action’. In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO ’21. –, NY, USA: Association for Computing Machinery. ISBN: 9781450385534. DOI: `10.1145/3465416.3483303`.

DECLARATION

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree. Except where indicated specifically in the text, this thesis was composed by myself and the work contained therein is my own.

Hanover, Brighton, March 2022

Oliver Timothy Thomas

ACKNOWLEDGEMENTS

The work contained in this thesis would not have been possible without the support of those close to me. Throughout this process I have felt inspired, deflated, humble, determined, sad, and elated (among others), and I could not have navigated that maze without them.

Thank you to my supervisors — Novi, for pushing me further than I thought possible and supporting me along the way; and David, for wisdom and perspective. Thank you to my co-authors, for listening to my half-formed ideas, fully-formed jokes, and being suitably honest about both. In particular, thank you to my office-mates Myles and Thomas, who make everything fun. I would also like to give a special thank you to the European Research Council (ERC) for partial funding of my work through the BayesianGDPR project (Grant agreement ID: 851538). Lastly, to Clare, who has shared every part of this adventure with me — thank you for everything.

CONTENTS

III CONCLUSION

ACRONYMS

AE	autoencoder
CNN	Convolutional Neural Network
DAG	Directed Acyclic Graph
DI	Disparate Impact
DP	Demographic Parity
DT	Disparate Treatment
EOdds	Equalised Odds
EOpp	Equality of Opportunity
FPR	false positive rate
GAN	Generative Adversarial Network
HSIC	Hilbert-Schmidt Independence Criteria
INN	Invertible Neural Network
KNN	K-Nearest Neighbours
LUPI	Learning Using Privileged Information
ML	machine learning
MLP	multi-layer perceptron
MMD	Maximum Mean Discrepancy
NN	(artificial) neural network
PPR	positive predictive rate
PPV	Positive Predictive Value
RBF	Radial Basis Function
SVHN	Street View House Numbers
SVM	Support Vector Machine

TNR	true negative rate
TPR	true positive rate
VAE	variational autoencoder
cVAE	conditional VAE
WAE	we’re all equal
WYSIWYG	what you see is what you get

NOMENCLATURE

P	Probability
s	Sensitive attribute/protected characteristic/spurious attribute
S	Random variable for the sensitive attribute/protected characteristic/spurious attribute
$\mathcal{S}$	Set of possible values for the sensitive attribute/protected characteristic/spurious attribute
$\boldsymbol{x}$	Input features (without the s attribute)
$\hat{\boldsymbol{x}}$	Reconstructed input
$\mathcal{X}$	Set of possible values for the input features
y	Class label (ground truth)
Y	Random variable for the class label
$\mathcal{Y}$	Set of possible values for the class label
$\hat{y}$	Predicted label
$\hat{Y}$	Random variable for the predicted label
$\boldsymbol{z}$	Encoding of $\boldsymbol{x}$

GLOSSARY

Adult/Census Income	A popular fairness dataset based on census data from the U.S.
CelebA	Face attributes dataset with more than 200K celebrity images.
COMPAS	A popular fairness dataset based on recidivism data in the U.S.
MNIST	Dataset of grayscale handwritten digits.
cMNIST	Dataset of colourful handwritten digits.
demographic group	Set induced by the sensitive attribute.
disparate impact	When policies, practices, rules or other systems that appear neutral result in a disproportionate impact on a protected group.
disparate treatment	Intentional discrimination of a protected group.
fairness definition	An aspirational (often legally inspired) specification of a fair classifier.
fairness metric	A metric which quantifies how well a fairness definition is satisfied.
JPEG	A standard image format.
protected characteristic	See sensitive attribute.
sensitive attribute	An attribute that, usually for legal or ethical reasons, should not be the basis for classification.
spurious attribute	An attribute that is correlated with the prediction target in the training set but not in the deployment setting.
SVG	Scalable Vector Graphics image format.

Part I

PRELIMINARIES

This part covers the introduction, the related work, and a summary of the work presented in Part II.

1 | INTRODUCTION

The increasing capability of machine learning (ML) models to perform well at specific tasks has led to their use in more consequential applications. This increased consequence has in turn led to greater scrutiny, with particular concern about what it means for an algorithmic decision, recommendation, or prediction to be 'fair'. In response, the research community has begun investigating these questions which are grouped together under the term *algorithmic fairness*. This burgeoning field of algorithmic fairness has been the focus of a growing body of research, with a number of definitions being introduced to quantify and measure *un*-fair behaviour, which, as a research community we aim to minimise, or ideally, eradicate. These definitions are often with respect to specific, legally protected characteristics that are observed alongside the features used for training an ML model, but cannot be used during inference. Examples of these protected characteristics may include race, gender, age, or disability status, among others.

Although algorithmic fairness is a multi-faceted problem, this thesis investigates a specific instance of a general concern — that data can contain spurious correlations. These are coincidental correlations between variables that are not related in a direct cause-and-effect manner. Despite these correlations not existing in the broader population, they may appear when non-random subsets of data are observed. Due to difficulties in obtaining representative samples, spurious correlations may be present in the subset of data used for training and validating an ML model, leading to a shortcut being exploited rather than a more complicated underlying true function being learnt. A toy example of such behaviour could involve a 'camels or cows' image classification model. Although in the wider world both camels and cows can be seen in a myriad of different settings, such as cities, beaches, or prairies, in this fictitious example these different settings are not well represented in the obtained data. Here, both the training data and the data withheld to validate model performance predominantly feature camels in a *sand*-based setting, and cows in a *grass*-based setting. There is a risk that instead of training a model to identify the animal present in an image as intended, the model simply learns to associate the image setting with the classification target. A spurious correlation between regions of the input image and a perceived target may be modelled due to underspecification of the task, producing incorrect results should an image of a camel in a pasture, or a cow on a beach be presented. Although the above example is simplistic, this becomes particularly important when the spurious correlation is between the model target and features associated with a protected characteristic. Examples of a model target for a classification task where sensitivity to protected characteristics is paramount may be approval or not, for a hiring, loan or bail decision. Simple rules such as 'invite male candidates to interview for a vacancy', or 'offer higher loans to white applicants' may perform well on the labelled training and validation data, but when deployed they may both perform poorly, and have the potential to cause significant harms to the population. This specific type of spurious correlation, often referred to as *biased data*, is the source of concern in this thesis. Biased data impacts performance and plays a large part in the trust afforded to ML-based systems. The effect of this can have a significant impact, especially in the case of decisions that directly affect a person's livelihood.

A promising approach to counteract biased data is by producing a *fair representation* as a pre-processing step. In fair representation-learning, the aim is to produce a transformation of the data such that it still retains utility for a downstream task, but has been modified so that information about a protected characteristic of concern is either removed, or obfuscated to the point where a downstream model produces 'fairer' decisions by default. The benefits of this approach are that the fairness-promoting aspect is isolated, allowing easier regulation, and allowing the process to be independent of other concerns. However, this approach is not without drawbacks. Completely removing some attributes while retaining utility is non-trivial; the burden of responsibility to check for unfair behaviour can be inadvertently moved away from a downstream system; and the data often becomes obscured when projected into an uninterpretable latent space, making intuitive assessment difficult. Making progress in addressing some of these drawbacks may promote the adoption of fair representations and the benefits that they provide.

1.1 Problem Statement

This thesis investigates fair representations of data and whether they can be used to provide additional insight into a system. Can we retain the benefits of fair representations of data — an isolated and measurable fairness-inducing intervention — while making progress in overcoming the shortcomings? The result should be a transformation of the data that increases the fairness of a post-hoc ML model by default, while retaining the utility of the original input, and still remaining as interpretable as the input data.

This desiderata poses the research question that is tackled in this thesis: 'How can we make fair representations interpretable?'. To approach this, I first develop a method that uses fair representations to interrogate the effect of a specific protected characteristic. This will provide insight into the relationship between the feature of interest, the protected characteristic, and the remaining features. I will demonstrate that this can be used to promote fairer outcomes without necessarily directly manipulating an existing decision system.

Secondly, I will demonstrate that fair representations can exist within the data domain itself. This is not a trivial task. The resulting output of the transformation should reside in the original feature space and retain useful information about features other than the protected characteristic. In addition, the transformation should also obfuscate that particular feature.

Finally, I will improve on this first attempt at producing fair representations in the data domain, introducing models based on alternative approaches to achieve a more robust outcome with improvements to the qualitative results.

1.2 Motivation and Aims

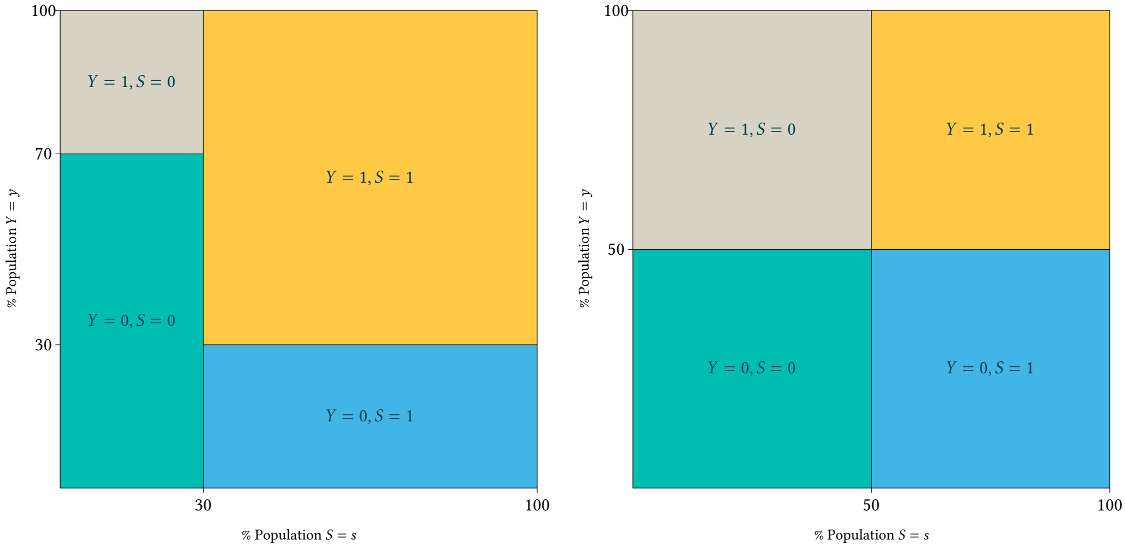
More data cataloguing human behaviour is being produced than ever before. The broad aim of many applications is to use this data to make sensible predictions about future events. These can be to assist the user by preempting their needs and queries, or to make decisions about the effectiveness and cohesion of potential hires. Ideally, to do this, we would aim to have the total information that was available to the user. However, this is not realistic. Instead, we typically have n-pairs of data $(\boldsymbol{x}, y)$, which form a dataset $D = \{(\boldsymbol{x}_0, y_0), \ldots, (\boldsymbol{x}_{n-1}, y_{n-1})\}$. These pairs represent input features $\boldsymbol{x}$ from the set of possible values $\mathcal{X}$, and outcomes y from the set of possible outcomes $\mathcal{Y}$. If the data were total, then we would have all of the information necessary to emulate the true underlying mapping from $\mathcal{X}$ to $\mathcal{Y}$. Instead, we are limited to obtaining, at most, data that can be recorded. As such, the aim is not to reconstruct the ground-truth mapping, but instead produce an approximation. Typically in ML we focus on finding an approximation function $f\colon \boldsymbol{x} \mapsto y$, from the set of possible functions in the hypothesis space $\mathcal{F}\colon \mathcal{X} \mapsto \mathcal{Y}$, that most accurately models this relationship (minimises the Empirical Risk). However, recent works have questioned if this alone is the best criterion for success. Instead, fairness-aware ML algorithms take into account additional information in the form of a protected characteristic s from the set of possible values $\mathcal{S}$, and seek to reduce the hypothesis space to functions that either don't make use of s at all, or allow for a defined margin-of-error[1]. An overview of related work that aims to achieve this is discussed in chapter 2.

One application for ML models is emulating current decision processes. For tasks such as loan approval, decisions have traditionally been made by a number of decision makers employed for the task, each with their own thresholds, preferences, and biases. In such a setting, the promise of automated decision systems is clear. An automated system can process millions of applications incredibly quickly, is available at all times, and crucially, will be consistent in its decision making process. However, there are drawbacks. Any errors or inconsistencies in the logic learned from observing past behaviour have a greater chance of being exposed, and worse, perpetuated. With such a system deployed, it is no longer possible to pass off inconsistent decision making as human error. The challenge to produce a fair system might be difficult, but there is significant opportunity for improvement from any unconstrained system. In such a system, making the outcome 'fairer' in any way can have a significant and practical impact, even if absolute fairness is not achieved.

One criticism that is often levelled at ML models, especially those deployed in human-centred scenarios, is that the decision making process is not clear. In addition to our desire to produce fairer results, it is also important that stakeholders in the system feel confident in any fairness-enhancing interventions introduced. On top of the aim of producing a fairer result, any amendments to the system should also increase the interpretability. An improved fairness intervention solution would not only increase the fairness of the system, but would allow stakeholders to gain some knowledge of what changes are required for this to be met.

Lastly, a concern for generally adopting fair ML is the potential trade-off between model accuracy and how 'fair' the system is. I explore more about different definitions of fairness in chapter 2, however there is a simple case to demonstrate that there may not be a trade-off after all. In figure 1.1 we witness the case where the training dataset is imbalanced in relation to the deployment setting. This can be for a number of reasons, such as using historical data, or only having access to a limited source of data. In the deployment set, however, the data is balanced. If the features available to train a model *are not* sufficiently rich for a function to approximate the mapping of X to Y, but *are* sufficiently rich to map X to S, then the protected characteristic may become a *proxy label* for the target. In this case, the data could be categorised as biased —

[1] The maximum margin-of-error is often legally defined.



(a) Training Population.

(b) Deployment Population.

Figure 1.1: An example of possibly biased data. The training dataset is imbalanced with respect to both outcome y, and s-group, however the deployment setting is balanced with respect to both. The training set does not reflect the deployment set. Depending on the complexity of the task, it may be simpler to use s as an indicator for y. In this setting, enforcing some fairness criteria may improve generalisation performance. This is explored in greater detail in chapter 2.

there exists a spurious correlation between S and Y that is only present in the training set. By providing an additional inductive bias that the outcome should *not* be dependent on S, we may produce a function f that is closer to the ground truth than the training data implies.

The aims of this thesis are to provide an approach that reduces the effect during model training of a specific type of spurious correlations between features related to a protected characteristic and a target label which appear as biased data. I produce this at the *pre-processing* stage in the form of a data transformation. While ideally a protected characteristic would be completely obfuscated, this is an unnecessary aim. Instead, the task of deciphering the protected characteristic need only be more complex than learning to perform the task.

As an additional aim, the resulting transformation should give us some insight into the transformation process itself. Of particular concern is additional problems being introduced by the transformation process. For example, if the protected characteristic is 'gender', then any changes made to hide this feature that in turn affect skin-tone are an indicator that the system designer may also need to consider 'race' as an additional source of potential bias. Similarly, if the system returns a clearly degenerate solution, then it may save months of development time by highlighting this problem earlier.

1.3 Claims and Contributions

In this thesis, I produce three main contributions. Firstly, I demonstrate that fair representations can be used to promote fairer outcomes within an already existing system. I achieve this by drawing a connection between the reconstruction of samples from fair representations and counterfactual examples. This work is catalogued in chapter 4.

Secondly, I demonstrate that fair representations can exist within the data domain, making use of the inherent interpretability that this domain provides. I make a first contribution to this in chapter 5 using a statistical dependence measure to promote a fairer representation under an additive decomposition assumption, allowing the data to be broken down to a 'fair' and 'unfair' component.

Lastly, I improve on this first attempt, assuming a more complex relationship between the 'fair' and 'unfair' components and introduce *null-sampling* in chapter 6 to draw manipulated samples from a designated region of a learned latent space. This opens up alternative techniques to achieve fair representations in the data domain, making use of the properties of both a conditional VAE (cVAE) and an Invertible Neural Network (INN).

1.4 Thesis Outline

This thesis is organised in the following way. Chapter 2 gives an overview of algorithmic fairness, and in particular, publications to date on fair representations of data. Following this, chapter 3 describes each of the three main chapters of this thesis in greater detail, with an emphasis on how they relate to each other. Chapters 4 to 6 contain three peer-reviewed and published works which have been reproduced with minimal changes except where explicitly indicated. The work presented in chapter 4 is currently under review as a journal submission which extends on published conference proceedings. Chapter 5 contains an addendum with experimental results to help motivate chapter 6. Finally, in chapter 7 I present the main conclusions and suggest possible future directions for the presented work.

2 | RELATED WORK

'The world isn't fair, Calvin.'
'I know Dad, but why isn't it ever unfair in my favor?'
— Calvin and Hobbes, 14 April 1986
BILL WATTERSON

This chapter aims to provide a summary of research to date in the area of algorithmic fairness, with a particular focus on fair representations. This is a broad research area which includes the topics of fairness, interpretability, and ultimately, accountability in automated decision-making and decision-support systems. While this is a relatively new area of research, there is a growing body of work with dedicated conferences such as FAccT[1], AIES[2], and EAAMO[3] as well as events at prestigious conferences with a more general scope, such as NeurIPS (Barocas and Hardt, 2017) and ICML (Corbett-Davies and Goel, 2018). The reason for this increase in activity is simple — the problem is complex. Machine learning (ML) models find patterns in, and ultimately reflect the underlying data. This has enabled them to be incredibly successful, performing to a superhuman standard for many tasks (Silver et al., 2017; Vincent, 2017; Brown and Sandholm, 2018; Jumper et al., 2021; Ravuri et al., 2021). Typically, these tend to be objective problems such as predicting the weather (Holmstrom et al., 2016), playing Atari games (Adamski et al., 2018) or distinguishing between plant phenotypes (Singh et al., 2016). Naturally, the success and high performance of machine learning techniques in these areas has led to the desire to apply these same techniques to more subjective applications, such as advertising (Sweeney, 2013), parole hearings (Angwin et al., 2016) and CV screening (Albert, 2019) to name a few. The promise of instant, consistent, and cheap decision making clearly has high impact potential. However, without due care, ML models can exhibit the problem described in Kallus and Zhou (2018) as 'Bias In, Bias Out' — an analogue of the database mantra 'Garbage In, Garbage Out'. This description refers to training a model on biased data — that is, a non-random subset of data that exhibits a spurious correlation

[1] https://facctconference.org/
[2] https://www.aies-conference.com/
[3] https://eaamo.org/

between a subset of features (that do indeed have a correlation with a protected characteristic), and the target, despite the observed correlation not being present in the wider population. An ML model trained on such data may (unwittingly) approximate the spurious correlation between the incorrect features and the target, rather than approximating the intended function. This can have serious unintended repercussions. In principle, this short problem description is appealing, but it represents only part of a larger picture. While data is *a* source of unfairness in ML models, it is not the sole cause. It would be remiss to discuss fairness without highlighting that sociological, economic, and historical factors are a major contributor to unfairness in general. In addition, the role of the system designers in determining criteria for success and monitoring these criteria during the period a model is deployed are also fundamentally important. However, exploring these avenues is outside of the scope of this work. Here, the focus is on the data.

2.1 Outline

The predominant discussion in this chapter, and in this thesis generally, is around *fairness*. Specifically, fairness applied to ML systems — what it is, where it fits in a broader context, and how to promote it. To facilitate this, this chapter is laid out in the following way: Initially, some background for studying this as a research topic in its own right is given in section 2.2. There is then a brief discussion in section 2.3 placing fairness, and in particular research into fair representations, into a wider scope of adjacent problems. Next, in section 2.4 there is a brief discussion of how fairness fits more broadly into the topic of *A.I. Ethics*, of which fairness, transparency, interpretability, and accountability are pillars. In section 2.5, definitions of *fairness* are reviewed from a group and individual perspective, followed by a discussion on when enforcing fairness can be beneficial, and the scenarios in which it can be detrimental to the utility of a model. This is followed by an overview of how fairness constraints are being added to existing models in section 2.6. Lastly, a predominant issue in this area revolves around the problem that what is considered discriminatory is domain-specific, requiring subject matter expertise to identify. For example, the sex of a patient may be an important non-discriminatory feature in a diagnosis system, but would be considered discriminatory by a bank to determine if an applicant should receive a loan. A useful tool for facilitating these discussions around the relationships within data is *Causal Modelling*. A brief summary of work in this is provided in section 2.7.

2.2 Motivation

The area of bias and discrimination is not a new one. Legal scholars have been debating these problems for centuries (Pole, 1978; Kennefick, 2018). As such, there are a number of statutes around the world defining what it means to illegally discriminate. Although there is not consensus, the prevailing opinion is that decisions should not be based on the immutable characteristics of an individual. That is, features that an individual has never had a choice over, and cannot reasonably change should not be used as the basis of a decision — they are not relevant to the task. In the UK, discrimination based on age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation are all covered by the Equality Act 2010, let alone other protections within specific domains. Producing automated decisions, or decision recommendations, *should* then be sensitive to these attributes. 'Ignorantia juris non excusat' — ignorance is not an excuse. A well-meaning practitioner may first think to simply remove the features of concern from a dataset. The rationale is that if an ML model does not have access to a protected characteristic, then it cannot use it to make a decision, thus avoiding *disparate treatment*. Unfortunately, the effect these protected characteristics have is so profound, that even removed, they are often largely recoverable from the remaining data (Pedreshi et al., 2008). As such, the problem about automated bias has been highlighted by researchers for a number of years and institutions are starting to pay attention. Furthermore, deployed applications are having to be justified or withdrawn after investigations have demonstrated that they are falling foul of this very concern. For example, Propublica's *Machine Bias* (Angwin et al., 2016) article sparked debate and raised concerns that needed to be addressed by the community after demonstrating that (at least on the surface) recidivism prediction software produced by Northpointe advised that black people in the U.S. were more likely to re-offend than offenders with a similar profile who were white. Concerningly, this is a pattern that has been repeated in similar areas as reported in Johnson (2022). As a cause for optimism, both the U.K. House of Lords (2018) and the U.S. Whitehouse (2010) now say that this issue should be addressed.

The aim is to approach justifiable concerns head-on. Doing this has a number of benefits. It is changing the questions that we are asking about fairness and biases, the impact they have on our own societies, and prompting researchers to find innovative ways of adapting models to complex real-world problems.

2.3 Context

Before discussing the definitions of fairness, let us consider how fair representations, the predominant focus of this thesis, relate to fairness and in turn, how fairness in ML relates to similar tasks.

2.3.1 *From Fairness to Fair Representations*

As mentioned, this chapter is concerned with fairness. This thesis though is concerned with fair representations. Methods to implement these are elaborated on in section 2.6, but here is a brief introduction to both the definition of fair representations, and notation that will be used throughout this chapter. The main idea behind fair representations is linked to the initial suggestion of removing protected attributes from a dataset, but it goes a little further. Instead of simply removing the features, the aim is to obfuscate the removed features from the remaining data. The core idea is to project the features used for training a model to a new latent embedding space, where the latent embedding is still useful for a task, but makes determining the removed features difficult. Formally, let X be the input space, and Y the label space. The objective is to find hypotheses $g: X \mapsto Z$ and $h: Z \mapsto Y$ such that two conditions hold. First, Empirical Risk should be minimised by the application $h \cdot g$, that is $\min \mathscr{L}(Y, h(g(X)))$. Second, some dependence measure, $f(\cdot,\cdot)$ that can be used to measure the relationship between Z and S should be minimised — they should be independent. Motivation for the core idea behind fair representations is based on the data processing inequality from Information Theory, that 'local processes cannot increase information content' (Beaudry and Renner, 2012). Given a Markov Chain of three random variables $X \to Z \to Y$, then Y is conditionally independent of X. In addition, no post-processing of Z can increase the information that Z contains about X. Expressed in terms of Mutual Information, this can be written as

$$I(X;Z) \geq I(X;Y)$$

If independence between Z and S is achieved, then any mapping from Z to Y must also satisfy independence between Y and S. The benefit of this fairness intervention is that it is a contained step. Furthermore, fair representations present an opportunity for additional applications; some of which are presented and investigated in this thesis.

This concept of expressing data as a constrained representation though, is not unique to the area of fair representations. Indeed, this approach has been applied in other research areas. The arising question then, is what characterises fair representations so that they are worthy of investigation, independently of these related problems?

2.3.2 *The Wider World*

There are of course multiple research areas that overlap to some extent, and fairness is no exception. For many problems, there is a connection to fairness-inducing methods. One similar research problem to fairness is Domain Adaptation, which in turn is similar to Transfer Learning. In Domain Adaptation, the aim is to produce a model that performs well on a different (but related) target data distribution than that which the model was trained on. This is sometimes described as a *distribution shift*. An example of domain adaptation is training on house number signs, such as those in the Street View House Numbers (SVHN) Dataset (Netzer et al., 2011) for deployment reading the handwritten digits dataset MNIST (LeCun et al., 2010) (French et al., 2018; Hoffman et al., 2018; Saito et al., 2018; Wang et al., 2021). A common approach in Domain Adaptation is to project the input to a new latent embedding such that multiple domains project to the same embedding space . An analogy for this type of many-to-one processing could be downsampling an image. Multiple high-resolution images may be visually indistinct in low-resolution. However, the utility of this downsampled image for a task might be severely hampered. The challenge becomes retaining as much information relevant to the task as possible, independent of domain. Additionally, *domain labels* are provided during training, these are a categorical label indicating which domain (dataset) a training sample comes from. Although there is research relaxing this constraint such as in Creager et al. (2021).

Producing embeddings for Domain Adaptation may sound similar to fair representations, and there is certainly a connection. Work originally designed for this problem, such as Ganin et al. (2016), has been re-purposed for fairness in works such as Edwards and Storkey (2016), Beutel et al. (2017), Jaiswal et al. (2018b) and Yang et al. (2020) among others. The differences, however, are subtle but serious. Firstly, there is a difference in the scale of distribution shift. In domain adaptation, the domains are semantically similar, but distinct. A typical problem is training on SVG images of vehicles and evaluating on JPEG images of vehicles in the real world. In fairness problems, the equivalent domains are the values of protected attributes, for example, whether an applicant is male or not. The remaining features, for example, word embeddings from a C.V. in a hiring scenario, are often semantically identical, and the distinctness is arguable. There is also a difference in the effect. In domain adaptation, the aim is to produce *better* results in the deployment setting. The goal is to maximise generalisation from the source to target domains. There is no requirement that performance across domains be equally performant. There is also a difference in the sensitivity to using the domain label at inference time. A valid approach in domain adaptation may be to determine the domain and use a different model based on this inference (Wang et al., 2020), in fairness applications, this would be a problematic design decision.

The difference between Domain Adaptation and fairness when trying to ensure complete independence between a predicted outcome and a protected attribute is perhaps slight. However, independence is only *one* notion of fairness. As will be discussed in section 2.5 there are multiple definitions. When other notions of fairness, such as *Equality of Opportunity*, or *Equal Calibration* are used, the comparison falls flat. Although it is certainly possible to enforce these criteria in a Domain Adaptation setting, because of the difference in aims, these would hinder, rather than help performance.

Due to these differences, fair representations are being designed, implemented and evaluated as distinct from other approaches. However, it is quite common for progress in each field to permeate and provoke improvements across these research areas with differing priorities.

2.4 The Pillars of Ethical ML

Fairness, Accountability, Interpretability and Transparency are all cornerstones of A.I. Ethics. Although this thesis is ostensibly about technical solutions to fairness issues, there is an overlap with the other topics in terms of motivation. The aim is to increase fairness, but to do so in a way that allows for greater accountability by providing a suite of approaches that increase the interpretability of a system, and in turn, making the system more transparent. In this section, all of these topics are briefly expanded upon with two aims. The first is giving a short introduction to the topic and providing an illustration of the research questions within. The second is to provide a wider context to chapters 4 to 6, despite not being the primary focus of these works.

2.4.1 *Fairness*

Fairness, as alluded to, is a reference to outcomes with respect to a protected attribute, and is largely the focus of the contributing chapters of this thesis. The predominant body of literature regarding fairness is based around classification, which is an inherently discriminative[4] task, although this need not be the case — the principles of fair behaviour can apply to any task, including regression (Agarwal et al., 2019; Chzhen et al., 2020), recommendations (Beutel et al., 2019) and resource allocation (Li et al., 2020), among others. The task of binary classification is used throughout this chapter, however, for expediency.

In the U.S., fairness legislation broadly falls into two groups: Disparate Treatment (DT), and Disparate Impact (DI) (Barocas and Selbst, 2016). As such, much of the early work on fairness interventions developed around this language. Understanding the notions that these capture then becomes critical when following the development of fairness enhancing models.

A decision making process is said to suffer from disparate treatment if its decisions are (even partly) based on an applicant's protected attributes. Outcomes that disproportionately affect one group in either a positive or negative way, despite the application of seemingly neutral processes are said to suffer from Disparate Impact.

To avoid DT the simple solution is to ensure that a decision making process does not have access to the protected attributes. However, as Pedreshi et al. (2008) explain, this is not a straight forward process. It is simply not enough to not directly ignore a sensitive attribute. The reason for this is that a sensitive attribute can be effectively 'reconstructed' from the other features. In their paper, which was one of the first to address fairness in a context related to ML they give the example of determining whether to give a loan to an applicant or not. They point out that if we decide not to capture the race of an applicant, but still capture the area code, we could potentially learn the rule 'rarely give credit to applicants in neighbourhood 10451 from NYC'. This may seem harmless, but if a subject matter expert advised that the vast majority of people in NYC area 10451 were black, then the learned rule is equivalent to 'rarely give credit to black-race applicants in neighbourhood 10451 in NYC', which is evidently discriminatory (Pedreshi et al., 2008). To

[4] Discriminative in the Latin sense that we are trying to discriminate between two (or more) classes. This is an important distinction raised in Pedreshi et al. (2008). The membership of the category that we are conscious of not discriminating against is referred to as a *potentially discriminatory*, or *protected* attribute. Their paper argues that this is different to being a *sensitive attribute* giving the example that gender is not often considered sensitive (withheld), but it can be potentially discriminatory. In general, later work has adopted that terms 'sensitive attributes' and 'potentially discriminatory attributes' and 'protected attributes' are used interchangeably, although some works, such as Chiappa (2019) (For more on this paper, see section 2.7) go back to this original view that they are indeed different.

satisfy DT then, a more complicated transformation of the data, rather than a simple masking of some features must be used. The authors distinguish between direct discrimination, which uses a protected attribute directly, and indirect discrimination which uses a non-protected feature (or a combination of features) as a proxy for the protected feature and then use this proxy in their evaluation, which they demonstrate using the 'German Credit' dataset.

Similarly, a simple way to avoid DI is to use protected attributes when making decisions. Then, verifying that the outcomes satisfy the fairness criteria becomes straightforward. However, this of course would constitute DT. Clearly this is a nontrivial task. One of the first papers to investigate a fairness intervention regarding classification was Kamiran and Calders (2009) with their *CND* (Classification with No Discrimination) model. Their first contribution is to define a measure of discrimination in a dataset:

$$Disc \coloneqq P(\hat{y} = 1 | s = 0) - P(\hat{y} = 1 | s = 1)$$

Their solution hinges around the notion that the bias isn't a property of the features of an individual, but within the mapping of input features $\boldsymbol{x}$ to the outcome label y. The remedy proposed is to massage the data prior to training a classifier so that no discrimination is present in the eyes of a Naïve Bayes classifier by simultaneously *promoting* applicants from a protected group and *demoting* applicants not from this group by 'flipping' the outcome label in the training data. The notion that bias exists within the mapping to an outcome label is an interesting one and reflects our understanding of the world. Intuitively, there is no bias in just having an attribute, such as race, the bias only exists in outcomes based on that feature. More recently, however, this assumption has been challenged. It has been observed that due to the inherent feedback loop of decisions regarding people, those decisions that affect a generation have repercussions. If a group are perpetually discriminated against, then over time the sensitive attribute is reflected in other features (Liu et al., 2019), strengthening the case for intervention in the feature space.

Further methods to invoke fair outcomes are discussed in section 2.6, but it's worth noting that research into fairness in ML has developed in several ways. In one set of approaches, researchers investigate different definitions of fairness, particularly cases where existing definitions do not reflect the problem well. This is discussed in section 2.5. Another approach that some researchers take is to try and improve on the bounds of a fairness-enhancing model section 2.6. In addition, there is research that highlights ML reflecting societal inequalities such as Bender et al. (2021), though again, these wider works are outside of the scope of this thesis.

2.4.2 *Accountability*

Although fairness is a measurable quantity (once the definition has been agreed upon), accountability is a less mathematical construct. However, there has been research into what accountability in ML could, and perhaps should look like. This topic provides a compelling justification for fairness methods being adopted. As greater levels of accountability are introduced, it can be reasonably expected that methods that may mitigate any liabilities become of greater interest. Although this topic is largely outside the scope of this thesis, here is a short summary of a selection of works.

In Wachter et al. (2017) open questions are presented — can human-interpretable systems be designed without sacrificing performance? how can transparency and accountability be achieved in inscrutable systems? and how can parallels between emerging systems be identified to set accountability requirements? The first of these questions is motivation throughout this thesis, but particularly for chapters 5 and 6 which attempt to address this issue.

The concern in Citron and Pasquale (2014) is 'arbitrariness by algorithm' and the effect that this may have on society. Although this is slightly beyond the scope of this thesis, the presented concern is an adjacent topic to chapter 4. The authors suggest that individuals assessed by predictive models should be notified that they have been assessed, along with being presented with the opportunity to challenge the assessment. Individuals, or neutral experts should be able to 'open up the black box scoring system'.

A counterargument to the proposal put forward throughout this thesis is Ananny (2018) in which the authors argue that accountable systems cannot simply be created by incorporating transparency. In their eyes, accountability is about addressing power imbalance and transparency is limited in its ability to deal with this. As the models are complex, they argue, transparency is unlikely to be a binary attribute. It is therefore important to not only consider what transparency reveals, but also what is not revealed. This aligns with the findings and limitations in chapters 5 to 6, where transparency is improved for part of the system, but some elements remain opaque.

As a reminder of the importance of remaining vigilant of current developments, the approach to accountability in Diakopoulos (2014) is that 'Journalistic approaches' should be taken to try and interrogate the semantic behaviour of a decision system. On the one hand, the aim is that stakeholders will be more critical of their own products, but also some responsibility also falls on journalists to become fluent in methods to interrogate these systems. This is an element that is advocated for in chapter 7.

In Hwang (2018) they bring in a broader perspective. Computational decision processes are in part determined by the computational power available. As such, regions with the greatest access to computational resources will be the ones to determine the ethics of more complicated models. As we transition into the a period where large 'foundation' models are routinely being used as the basis of other applications, this perspective is becoming more pertinent.

2.4.3 *Transparency & Interpretability*

The last of 'The Pillars of Ethical ML', relate to greater accessibility. The terms interpretability and transparency are inconsistently defined within the literature, but here interpretability is the degree to which a human can understand the cause of a decision. Transparency is the degree to which a human can interpret the training and inference procedures that lead to a decision. Some models are inherently interpretable and transparent, such as Decision Trees, or to a lesser extent linear models, but these are limited in terms of the complexity they can capture. There are also model agnostic interpretability techniques such as local surrogate models (Ribeiro et al., 2016), or game-theory approaches to explanations such as the Shapley Value (Winter, 2002), which are applied post-hoc. Throughout this thesis the aim is not to be completely transparent and interpretable, but instead, to improve on existing methods by making them less opaque.

From Miller (2019), Interpretability is the degree to which a human can understand the cause of a decision. The book *Interpretable Machine Learning* (Molnar, 2019) frames that statement in a slightly different way, describing interpretability as 'the degree to which a human can consistently predict the model's result'. We have already made a distinction between transparency and interpretability, but Molnar (2019) goes further, distinguishing between interpretability and explanation. Even if we are capable of interpreting the results of a model, Miller (2019) argues, unless we receive an explanation of how that model came to make a decision, then we will be unable to reliably reproduce the results. Not only that, but as humans, simply any old explanation will not do, we require a *good* explanation. According to Miller (2019), good explanations are not only truthful and coherent, but are — *Contrastive*: We tend to think in a counterfactual way, i.e., would I have been approved for a loan if I earned more money, and explanations should reflect this. *Selective*: The world is complex and we do not like to receive too much information. As such, we should only give between one and three explanations that cover the majority of cases. *Social*: They should be tailored to the audience. *Causal*: While truth and probability matter, audiences tend to find cause and effect explanations more satisfying.

The concepts above have been introduced despite not being the primary focus of the work in this thesis. They are however relevant to the motivation of the work introduced, particularly in chapters 5 to 6. Achieving fairness without accountability is unlikely. Accountability without interpretability and in turn transparency may be similarly difficult. Incorporating ideas from these adjacent research areas into any potential fairness interventions then could be desirable.

2.5 Definitions of Fairness

As fairness is the primary concern in this thesis, in this section the definitions of this topic are explored. A first disambiguation is that discrimination and fairness are not the same thing. One is the problem, and the other is the remedy. Because of this, typically we describe measures of discrimination and fairness constraints, which are used to combat discrimination. Discrimination measures naturally align into two main groups, characterised by Dwork et al. (2012) as *group fairness* and *individual fairness*. Throughout chapters 4 to 6 of this thesis, the notion of fairness promoted is *group fairness*, although by contrasting this notion of fairness against others, we can obtain a better understanding of its scope and limitations. The term 'group' refers to the measure being applied to the collection of people who form the group. Individual fairness, on the other hand, evaluates fairness at an individual level, rather than as part of a group.

2.5.1 *Group Fairness*

In Barocas et al. (2019) definitions of group fairness are described as belonging to one of three groups, *Independence*, *Separation* or *Sufficiency*. This pattern is adopted in this section. Throughout, the random variables Y, S and $\hat{Y}$ are used to denote the observed outcomes, protected attributes, and predicted outcomes, respectively.

2.5.1.1 *Independence*

The most intuitive notion of fairness is *Independence*. This is the notion that given a prediction ($\hat{Y}$) of a recorded outcome (Y) and a protected sensitive attribute (S), then the prediction should be independent of the sensitive attribute

$$\hat{Y} \perp S$$

In fact, one of the first papers in algorithmic fairness literature, Kamiran and Calders (2009), use this as their discrimination measure. Although written in a different form, they later used the notation latterly adopted throughout the fairness literature in their journal article Kamiran and Calders (2012), which expanded on their previous work (Kamiran and Calders, 2009; Kamiran, 2011). Other works often describe this definition as *statistical parity*

$$\text{Statistical parity} \coloneqq \; P(\hat{Y} = 1|S = s) = P(\hat{Y} = 1|S = \neg s)$$

Statistical Parity (or Demographic Parity (DP) as it is often known) appeals to an intuitive sense of group fairness, namely, that the outcome of the model should be independent of some sensitive attribute(s). For example, the probability that a student is accepted onto a course at a university should be the same regardless of whether of the student is male or female. This aligns directly with the DI discrimination definition and is the fairness measure associated with this type of discrimination.

Independence is a class of fairness notions, however, and although DP is the most prevalent in this class, there are other notions of fairness that fall within this class, including 'Difference in Mean Scores' and 'Difference in average residuals' (Zliobaite, 2015).

$$\text{Difference in mean scores} \coloneqq \; \mathbb{E}[\hat{Y}|S = s] = \mathbb{E}[\hat{Y}|S = \neg s]$$
$$\text{Difference in average residuals} \coloneqq \; \mathbb{E}[\hat{Y} - Y|S = s] = \mathbb{E}[\hat{Y} - Y|S = \neg s]$$

There are situations, however, where this does not work as intended. In these cases, instead of promoting the perceived harmed group based on the quality of the individual, as long as the probability of acceptance is the same, the criterion is met.

To illustrate this, let us consider an example that will be used across all our definitions. Imagine that we are in charge of admissions at a university and we are particularly concerned with complying with a fairness criteria regarding male and female subgroups. At this fictitious university, we can only accept 50% of all applicants. To determine if a potential student is likely to succeed, there is an entrance requirement, which is highly predictive of success. In fact, 80% of people who meet the entry requirement successfully graduate. However, many students apply despite not meeting the requirements. Universally, only 10% of students successfully graduate if they do not meet the entrance requirements. Both male and female subgroups apply to our university in equal numbers, though only 40% of applying males meet the entrance requirements, whilst 60% of applying females meet the entrance requirements. As already stated, we can only accept 50% of applicants to be students. Under Demographic Parity, we would require that 50% of both males and females be accepted, regardless of likely academic performance. Even though only 40% of male applicants meet the qualifying academic requirements, an additional 10% of the unqualified male population would have to be accepted to be 'fair', whilst 10% of qualified females would be rejected. A confusion matrix showing outcomes for this selection rate applied to both groups is shown in tables 2.2a and 2.2b.

To counter this, yet keeping within the frame of *independence*, relaxations of this criterion have also been suggested to include parity up to some threshold, ϵ. This could be expressed in terms of an absolute difference

$$\mid P(\hat{Y} = 1|S = s) - P(\hat{Y} = 1|S = \neg s) \mid \leq \epsilon$$

or via a ratio as suggested in Zafar et al. (2017b) which is seen as being comparable to 80% *rule* mentioned in disparate impact law (Feldman et al., 2015).

$$\frac{\min(P(\hat{Y} = 1|S = s), P(\hat{Y} = 1|S = \neg s)}{\max(P(\hat{Y} = 1|S = s), P(\hat{Y} = 1|S = \neg s)} \geq 1 - \epsilon$$

This rule suggests that as long as the selection rate of the 'harmed' group is within 80% of the 'privileged' group, then it is fair enough. Although critics of this point out that 80% was chosen seemingly arbitrarily.

2.5.1.2 *Separation*

A more complex definition of fairness is *separation*, which is independence given the observed outcome (Y)

$$\hat{Y} \perp S|Y$$

This has been formalised by the metric Equalised Odds (EOdds) (Hardt et al., 2016) which considers all values of Y, and the looser constraint Equality of Opportunity (EOpp) (Hardt et al., 2016), which only constrains independence given the observed outcome is positive.

$$\begin{aligned} \text{Equalised Odds} \coloneqq \; & P(\hat{Y}|Y=0,S=s) = P(\hat{Y}|Y=0,S=\neg s) \\ & \& \\ & P(\hat{Y}|Y=1,S=s) = P(\hat{Y}|Y=1,S=\neg s) \end{aligned} \tag{2.1}$$

$$\text{Equality of Opportunity} \coloneqq \; P(\hat{Y}|Y=1,S=s) = P(\hat{Y}|Y=1,S=\neg s) \tag{2.2}$$

Concisely, EOdds ensures matching both the true positive rate (TPR) and false positive rate (FPR) across the protected groups, whereas EOpp only ensures that the TPR of both (all) protected groups are equal. The benefit of this is that in some cases it may be a truer representation of fairness.

In our university admissions example, EOpp is equivalent to accepting members of both female and male subgroups at different rates, as long as the TPR of both groups is equal. To achieve this, we would be looking to accept 44.5% of males and 55.5% of females, which would give both groups a TPR of 85.4%. If we were enforcing Equalised Odds, we would have to make sure that we were not only matching the TPR, but also the FPR. In our example, the selection rate would be 46.4% for males and 53.6% for females. The effect of this on each group is demonstrated in the confusion matrices in tables 2.2c and 2.2d.

However, could an algorithm satisfy both independence and separation? Theoretically, this is possible in two scenarios, but practically these are unlikely to occur. Let's compare DP and EOpp, the less strict of the two separation-based group fairness measures introduced. DP requires that the positive predictive rate (PPR) per group is equal

$$PPR(s) = P(\hat{Y}=1|S=s) \tag{2.3}$$

and EOpp requires that the TPR is equal across groups

$$TPR(s) = P(\hat{Y}=1|Y=1,S=s) \tag{2.4}$$

For independence and separation to both hold, the PPR and TPR must be equal. In the above, we can see the first of the scenarios where this can occur: If Y and $\hat{Y}$ are independent of each other, which would results in a very poor classifier indeed, then $PPR = TPR$. This could be achieved, for example, by a random number generator, or in the degenerate case where the model consistently predicts only one output class e.g. $P(\hat{Y} = 0) = 1$ or $P(\hat{Y} = 1) = 1$.

Expressing TPR using Bayes Rule, we get

$$\begin{aligned} P(\hat{Y} = 1 | Y = 1, S = s) &= \frac{P(Y = 1 | \hat{Y} = 1, S = s) P(\hat{Y} = 1, S = s)}{P(Y = 1, S = s)} \\ &= \frac{P(Y = 1 | \hat{Y} = 1, S = s) P(\hat{Y} = 1 | S = s)}{P(Y = 1 | \hat{Y} = 1, S = s) P(\hat{Y} = 1 | S = s) + P(Y = 1 | \hat{Y} = 0, S = s)(1 - P(\hat{Y} = 1 | S = s))} \end{aligned} \tag{2.5}$$

The other scenario where both DP and TPR can hold starts with the condition that DP is met: that the prediction $\hat{Y}$ must be independent of S. If in addition, the observed value, Y is independent of the protected attribute S then TPR can also hold. This can be seen in equation (2.5) where if both $\hat{Y}$ and Y are independent of S, then the whole equation becomes independent of S.

2.5.1.3 *Sufficiency*

There is another notion of fairness that is less well utilised in the general fairness literature criteria called *sufficiency*. This is the concept that the true outcome, given the predicted score is independent of s.

$$Y \perp S \,|\, \hat{Y}$$

In our example, we would leave the selection rates alone, giving a selection rate of 40% for the male subgroup and 60% for the female subgroup, as we then treat applicants equally based on our anticipation of their success. Confusion matrices for this are shown in tables 2.2e and 2.2f.

As is the case in our example, sufficiency is typically satisfied by default in modern machine learning pipelines. In the case of a (artificial) neural network (NN) model, the logits are typically inferred as outcome labels during inference by either evaluating $f(\boldsymbol{x}) \geq 0$ in a 1-d output, or $\arg\max(f(\boldsymbol{x}))$ with an n-d output. To violate sufficiency would require s-specific thresholding (or manipulation) of these raw logits, which would be outside of standard practice. This approach is suggested as a manner to implement independence and separation notions of fairness however — see section 2.6.1.

Similarly to the conflict between independence and separation, there is a conflict between independence and sufficiency, and between separation and sufficiency. These conflicts can be described succinctly, so are demonstrated below.

Independence is $\hat{Y} \perp S$ and sufficiency is $Y \perp S|\hat{Y}$. By simply claiming both of these statements to be true, we then have:

$$S \perp \hat{Y} \text{ and } S \perp Y|\hat{Y} \implies S \perp Y$$

This shows that independence and sufficiency can only hold when the observed outcome in the data (Y) is independent of S. In other words, we have a dataset with outcome rates that are equal across protected groups.

The scenarios where separation and sufficiency are not in conflict are easier to view via Bayes Rule. Equality of Opportunity (EOpp) is an implementation of separation and requires $P(\hat{Y} = 1|Y = 1, S = s)$ to be consistent for all values of $s \in \mathcal{S}$. An instance of a sufficiency-based metric is Positive Predictive Value (PPV) and requires $P(Y = 1|\hat{Y} = 1, S = s)$ to be consistent for all values of $s \in \mathcal{S}$. Defining $Br_s = P(Y = 1|S = s)$ as the base rate, PPV can be expanded, giving:

$$\begin{aligned} P(Y = 1|\hat{Y} = 1, S = s) &= \frac{P(\hat{Y} = 1|Y = 1, S = s)Br_s}{P(\hat{Y} = 1|S = s)} \\ &= \frac{P(\hat{Y} = 1|Y = 1, S = s)Br_s}{P(\hat{Y} = 1|Y = 1, S = s)Br_s + P(\hat{Y} = 1|Y = 0, S = s)(1 - Br_s)} \\ &= \frac{TPR_s \cdot Br_s}{TPR_s \cdot Br_s + FPR_s \cdot (1 - Br_s)} \end{aligned} \tag{2.6}$$

For PPV and TPR to be equal across S, then one of three conditions must be met. First, make the right side of equation (2.6) must become independent of S, which can be satisfied if the classifier produces a degenerate result as previously described, resulting in the TPR being either 0 or 1. Or, there must be an impossibly accurate model where an incorrect prediction is never made. In other words, the FPR must be 0, and the TPR must be 1. Alternatively, if the Base Rate is the same across groups implying $Y \perp S$, then when PPV is satisfied, our measure of separation, TPR must also be the same across groups, along with the FPR, by the definition of PPV in equation (2.6). This demonstrates that we cannot have both sufficiency and separation-based notions of fairness unless the data is inherently balanced across protected groups with regard to outcome. This notion of trade-offs and balancing tension is a common one throughout fair machine learning, and certainly one that will be repeated throughout this thesis.

Table 2.1: Confusion Matrices per population subgroup (Male/Female) for the fictitious University Admissions example that runs throughout section 2.5.1. In all examples, the number of both male and Female applicants is 5, 000, giving 10, 000 total applicants for that academic year, of which only 5, 000 can be accepted. *Top Row*: An acceptance rate that promotes Demographic Parity (DP) is used. *Middle Row*: An acceptance rate that promotes Equality of Opportunity (EOpp) is used. *Bottom Row*: An acceptance rate that promotes Positive Predictive Value (PPV) is used.

(a) Outcomes for Male applicants with DP applied.

	Actual	
Predicted	Graduate	Retake
Accepted	1650	850
Rejected	250	2250

(b) Outcomes for Female applicants with DP applied.

	Actual	
Predicted	Graduate	Retake
Accepted	2000	500
Rejected	600	1900

(c) Outcomes for Male applicants with EOpp applied.

	Actual	
Predicted	Graduate	Retake
Accepted	1632	691
Rejected	268	2409

(d) Outcomes for Female applicants with EOpp applied.

	Actual	
Predicted	Graduate	Retake
Accepted	2142	535
Rejected	458	1865

(e) Outcomes for Male applicants with PPV applied.

	Actual	
Predicted	Graduate	Retake
Accepted	1600	400
Rejected	300	2700

(f) Outcomes for Female applicants with PPV applied.

	Actual	
Predicted	Graduate	Retake
Accepted	2400	600
Rejected	200	1800

2.5.2 *Individual Fairness*

All definitions of fairness that have been looked at so far consider statistics applied to subgroups of a dataset — these are referred to as *group fairness* measures. There is another approach, called *individual fairness*. This is the idea that regardless of any group as a whole, similar individuals should be treated similarly. A practical challenge of this approach to fairness though, is that it raises several questions about the nature of the term 'similar'. A simplified version of this idea was implemented by Thanh et al. (2011) in the context of a K-Nearest Neighbours (KNN) classifier, but this approach is generally credited to Dwork et al. (2012) who independently developed and refined the approach.

2.5.2.1 *Similarity Measures*

To determine the similarity between samples, a number of different approaches have been adopted. In Thanh et al. (2011), the authors use a Manhattan distance of z-scores for interval-scaled

Table 2.3: A non-exhaustive example list of different fairness criteria and their categorisation as *Independence*, *Separation*, or *Sufficiency* based.

Fairness Goal	*Definition*	*Example of*
Demographic Parity	$P(\hat{Y}\|S=0) = P(\hat{Y}\|S=1)$	Independence
Equal Opportunity	$P(\hat{Y}=1\|Y=1, S=0) = P(\hat{Y}=1\|Y=1, S=1)$	Separation
Equalised Odds	$P(\hat{Y}=1\|Y=y, S=0) = P(\hat{Y}=1\|Y=y, S=1)\ \forall y \in Y$	Separation
Equal Accuracy	$P(\hat{Y}=Y\|S=0) = P(\hat{Y}=Y\|S=1)$	Independence
Predictive Parity	$P(Y=1\|\hat{Y}=1, S=0) = P(Y=1\|\hat{Y}=1, S=1)$	Sufficiency
Conditional Use Accuracy Equality	$P(Y=y\|\hat{Y}=y, S=0) = P(Y=y\|\hat{Y}=y, S=1)\ \forall y \in Y$	Sufficiency

attributes, and the percentage of mismatching values for nominal attributes to determine the distance between data points. They determine that discrimination has occurred if in its k-nearest neighbours those within the same protected category have been treated differently to the neighbours of a different category. They propose that on finding points where they are confident that some discrimination has occurred, then the class label for that point should be amended. This data should then be used to train subsequent models.

A seminal paper in the field, Dwork et al. (2012) continued with the concept of a distance measure. They characterised individual fairness and proposed that it should be their goal. They suggest that given some *task-specific* similarity metric, δ, then similar samples should have similar outcomes, i.e. $|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq \delta(\boldsymbol{x}, \boldsymbol{x}')\ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ should hold, where $f(\boldsymbol{x})$ produces a continuous score as opposed to a discrete label. The authors acknowledge that obtaining δ is a tricky problem, described as 'one of the most challenging aspects of this framework'. Practically, it may require input from social and legal scholars or domain experts to help formulate this metric.

In Mukherjee et al. (2020) the authors propose learning a similarity score from the data. They propose two methods to achieve this. The first is based on factorising an embedding of the original data into a fair and *un*-fair representation, and then calculating the similarity based on the Mahalanobis distance between the fair embeddings. In their second approach, they require human feedback to determine 'comparable' pairs of samples. A logistic regression model is then trained to predict if two given samples are similar.

2.5.3 *Which Fairness Measure to use?*

The above is a useful framework for viewing fairness constraints and helps us to categorise various definitions of fairness, such as those in table 2.3, but that should not diminish the work that seeks to make novel strides within each of the areas. For example, Foulds et al. (2020)

expand the independence notion of fairness. Their inspiration comes from third-wave feminism and intersectional privacy, which they expand beyond binary sensitive groups and measure an unfairness value at each intersect. For example, consider we have a dataset with three sensitive attributes, sex, race, and religion. Most prior approaches consider these to be one feature *sexRaceReligion*. This paper measures the difference in outcome with respect to independence between each combination of sensitive attributes so that sex, race, and religion are all viewed as separate, measurable points of potential discrimination — the authors are concerned with whether discrimination occurs in any, some, or if only with all attributes present.

A question that may be reasonably asked is which fairness definition (or family of definitions) should one be using? This is a complex question to answer. Some, such as Heidari et al. (2018) and Yeom and Tschantz (2018) make the assumption that the choice of which to apply should come from the designer's moral perspective, arguing that this is a task outside of the expertise of computer scientists and instead should be debated by philosophers.

However, there is a view that the best approach is to model the problem and investigate the effects. Recently, works have started analysing the delayed impact that fair interventions in machine learning have on society (Liu et al., 2019). This work looks to model the impact that different notions of fairness have on the groups involved, recognising that there is more than the initial 'accepted for a loan' or 'rejected for a loan' dichotomy, but that this has an impact in terms of credit scores for the individual applying. This is a bold approach that tries to measure the effect that automated decisions may have one generation into the future.

The problem of which fairness criteria to apply was looked into in Hinnefeld et al. (2018). They look at a dataset[5] which they use to create four datasets, with combinations of Sample Bias, No Sample Bias, Label Bias and No Label Bias. They consider a binary race attribute (Black or White), where White race is $s = 0$ and Black race is $s = 1$. Label Bias is where there are different label thresholds based on race, and Sample Bias is where one group (in this case White race) has people selected at a higher rate. This paper demonstrates that no single fairness metric is able to pick up all discrimination and that all fairness metrics require 'a healthy dose of human judgement'.

[5] The dataset in the paper is unable to be released or referenced, according to the authors, due to 'contractual limitations'.

2.5.4 *Effect on utility*

It has already been demonstrated in section 2.5, that when enforcing fairness objectives, one degenerative approach to satisfy all fairness criteria is that the predictive value $\hat{Y}$ loses all relationship to the target variable Y. Clearly, this is not a practical solution, but it highlights that the trade-off between utility (accuracy) and fairness may need to exist. The question is, although a lack of utility can be introduced, is there necessarily a trade-off? Is it not possible to increase both utility *and* satisfy notions of fairness? In the following section, this is explored using hypothetical datasets from figure 2.1.

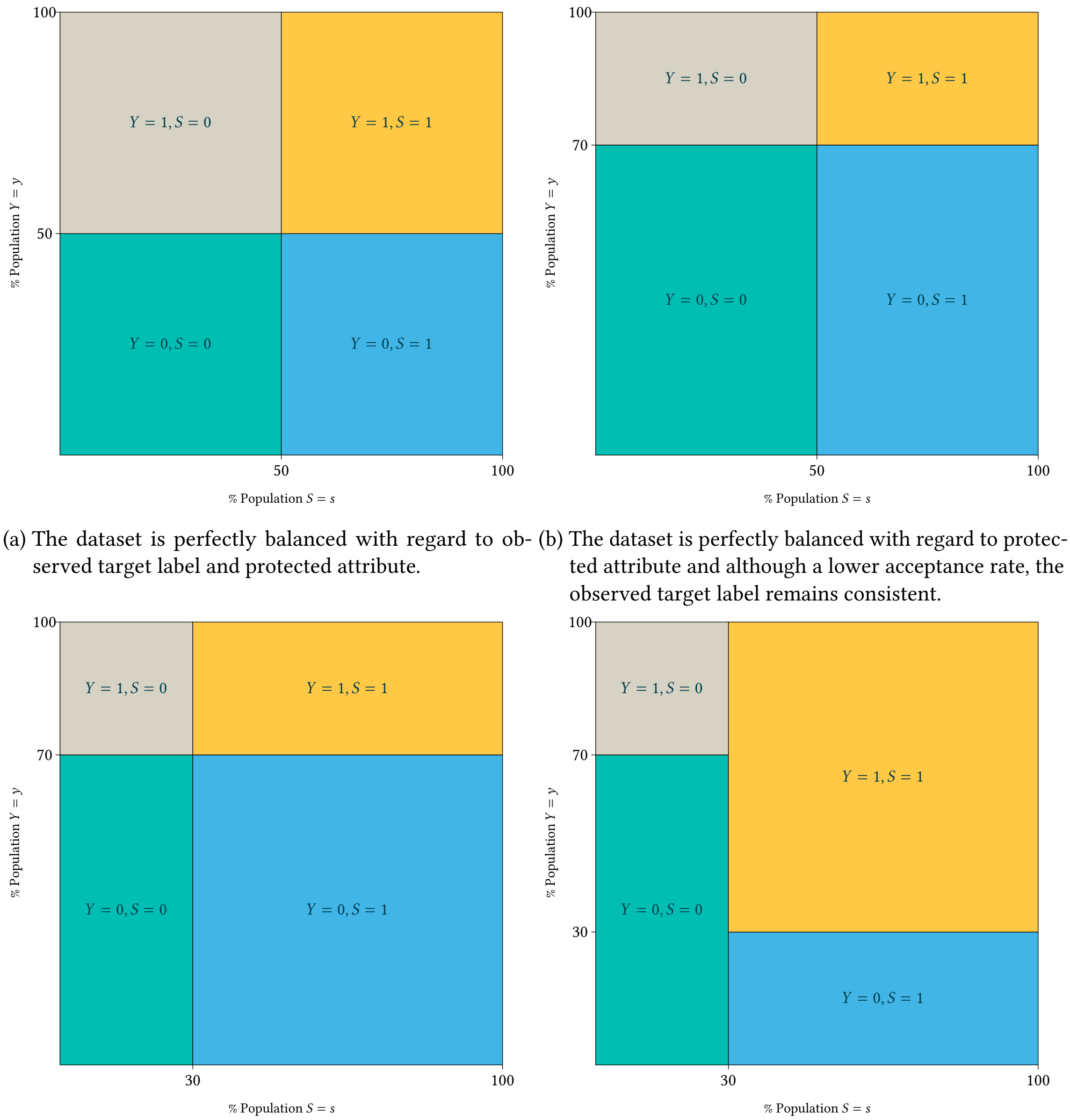
There are a number of scenarios to take into account, with various combinations of balanced and imbalanced training and deployment settings, and the type of fairness that is enforceable. To make the discussion simpler, only Demographic Parity (DP) and Equality of Opportunity (EOpp) are considered as fairness measures as they are most predominantly featured thorughout this thesis.

2.5.4.1 *Deployment Setting Reflected in the Training Data*

Let us first consider when the training data is representative of the deployment data. In the simplest case where the training and deployment setting are both 'balanced' (figure 2.1a), that is $0.25 = P(Y = y, S = s)\ \forall s \in \mathcal{S},\ y \in \mathcal{Y}$ in the binary case for both S and Y, there is no fairness-utility trade-off. A classification model exists that can minimise empirical risk and also achieve DP and EOpp. Furthermore, in this situation, a 'perfect'-classifier would achieve DP, EOpp and perfect accuracy. The constraint of being balanced in all quadrants can be relaxed to the case where the data is represented by figure 2.1b and the above statement about the existence of an ideal classifier remains true.

In the case where the data is not exactly balanced regarding S, but the outcomes are balanced, as in figure 2.1c, a model exists that still displays the positive characteristics of the above, but there is a risk that less importance will be put into correctly modelling the minority outcome ($Y = 1,\ S = 0$ in this case). This in turn potentially affects the generalisation capabilities of the model.

Lastly, there is a dataset that is imbalanced with respect to both the observed outcome and protected attributes figure 2.1d. With this kind of dataset, it is no longer possible to satisfy both DP and EOpp with a 'perfect' classifier. A model that maximises accuracy cannot be fair with



(a) The dataset is perfectly balanced with regard to observed target label and protected attribute.

(b) The dataset is perfectly balanced with regard to protected attribute and although a lower acceptance rate, the observed target label remains consistent.

(c) The dataset is imbalanced with respect to the protected attribute, but the observed outcomes across these groups is balanced.

(d) The dataset is imbalanced with respect to the protected attribute and the observed outcomes across these groups is imbalanced.

Figure 2.1: Four examples of the distribution of data in a dataset. In figures 2.1a and 2.1b the datasets are balanced with respect to protected attribute. In figures 2.1a and 2.1c, the datasets are balanced with respect to outcome. In figure 2.1d the dataset is imbalanced with respect to both protected attribute and observed outcome. The combinations of these datasets is used to discuss various settings where fairness-promoting methods can help, or harm generalisation of the model to a deployment setting.

regard to DP as the labels themselves don't satisfy DP. Depending on the definition of fairness being encouraged, in this scenario, a fairness-utility trade-off is introduced. This is because if DP is enforced, the model will have to incorrectly classify some samples to meet this requirement. However, if EOpp is enforced, then both the fairness criteria and utility can be mutually improved.

2.5.4.2 *Misleading Training Data*

Alternatively, there is the case where the training data does not represent the deployment setting. This can be due to a number of reasons, including a procedural screening during data collection as discussed in Kallus and Zhou (2018). This is one of a number of situations where a fairness-utility trade-off can be introduced, but there are also scenarios where fairness interventions can improve the generalisation of the model.

There are two main scenarios. First, where the data is balanced with respect to outcomes in the training set, but not in the deployment data, and secondly, where the reverse is true. When the training data conforms to data such as that in figures 2.1a to 2.1c, but the deployment setting is closer to figure 2.1d then a model achieving several definitions of fairness can be produced on the training dataset, but when deployed, the model may achieve a fair result in terms of DP, but would fare poorly in terms of utility. The effect on EOpp is not clear to determine, equal TPR (recall) may be achieved by virtue of the model performing poorly. Similarly, unequal TPR may be observed as equal performance is not guaranteed.

In the case where the training data is imbalanced (as in figure 2.1d) and the deployment setting is balanced, then during training it will not be possible to achieve both DP and high utility, though EOpp will be able to be met during training with an high-accuracy model. However, conversely, if DP is enforced though reflecting prior knowledge of the deployment setting, then encouraging DP will actually help the generalisation of the model, making the performance in terms of utility better.

Similar analysis was conducted empirically in Wick et al. (2019) and more theoretically in Maity et al. (2021). Clearly, understanding the context around any fairness-enhancing intervention becomes paramount, as no intervention is universally applicable. Furthermore, a 'blind' application of an intervention without consideration can lead to unintended results.

2.5.4.3 *Trade-offs between Fairness and Interpretability*

As demonstrated above, the relationship between utility and fairness is not simple. The work in this thesis, in addition to fairness, also introduces a form of interpretability into the decision process. Similarly to the espoused trade-off between utility and fairness, there is generally expected to be a trade-off between fairness and interpretability (Agarwal, 2021), but the intersect is still in the early stages of characterisation (P et al., 2021). In chapters 5 to 6 methods for introducing interpretability are introduced with limited trade-offs with respect to both utility and fairness. However, this method of incorporating additional interpretability is within the context of retaining freedom of model selection, which may not be available under some stricter definitions of interpretation.

2.6 Implementation of Fairness Enhancing Methods

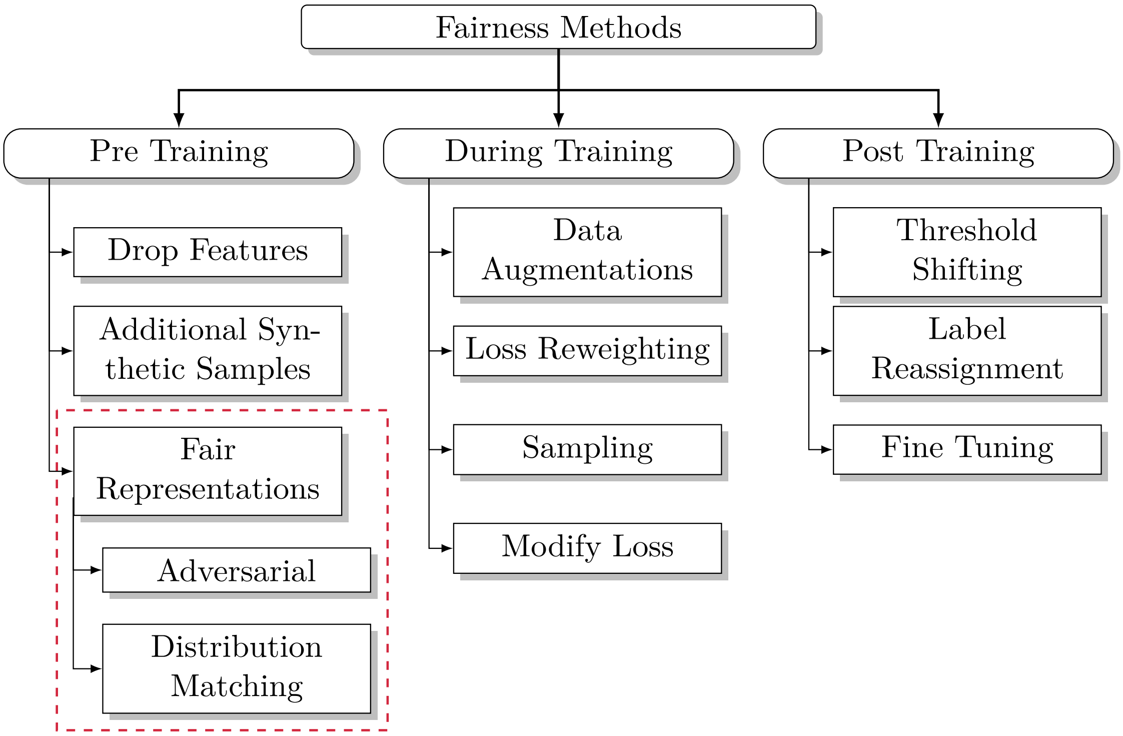


Figure 2.2: A taxonomy of fairness intervention techniques. The dashed box surrounds the topics that are contributed to in chapters 4 to 6 of this thesis.

After diagnosing a problem, the natural next step is to consider methods to remedy it. In the case of detecting discrimination in an automated system, the method to remedy this is adding fairness-enhancing interventions to the system. As with all areas, the line between the various points where fairness constraints can be injected is at times blurred. However, in general, we can think of fairness interventions occurring after, during, or before the training of a model with a taxonomy of approaches shown in figure 2.2. In this section a non-exhaustive selection of methods will be discussed, giving a brief summation of a variety of fairness-enhancing approaches, many of which are used as baseline models in chapters 4 to 6. The section is divided into three broad categories of fairness intervention methods *Post*-Training, *During*-Training - which covers minimising fairness constraints directly and *Pre*-Training - which includes feature selection and feature adjustment, adversarially removing sensitive attributes, and learning fair representations.

2.6.1 *Post Processing*

One of the first approaches was by Calders and Verwer (2010) who 'flip' the predicted outcomes for some samples close to the decision boundary so that the notion of fairness being aimed for is satisfied. To achieve this, they use a stochastic classification model and massage the probabilities of the outcome of each sample. This is later extended by Lohia et al. (2019) to address direct bias in the form of Disparate Treatment[6] by aiming to reduce the disparity in the predicted outcomes of a model directly conditioned on the protected attribute.

In Hardt et al. (2016) the thresholding value for the soft outputs of a classifier is amended per group. Further analysis of Hardt et al. (2016) is conducted in Awasthi et al. (2020) who show that even when only a subset of the protected attributes are known at inference time, this method still produces an optimal level of fairness. A criticism of this approach, however, is given in Woodworth et al. (2017) who highlight a scenario with biased data where this approach would fail.

A fundamentally different approach to that taken above is the fine tuning of an existing model's weights to make the model produce fairer outputs. This approach is presented in Savani et al. (2020). In this work, three approaches are suggested. Random perturbations of the weights; a layer-wise Bayesian Optimisation procedure; and an adversarial fine-tuning approach. In all approaches, the best performing model on a withheld validation set is selected. Results are demonstrated on tabular datasets that are common in the literature: the COMPAS recidivism dataset (Flores et al., 2016; Larson et al., 2016), the UCI Adult Income dataset (Dheeru and Karra Taniskidou, 2017), and Bank Marketing dataset (Moro et al., 2014) as well as the image-based CelebA dataset of Celebrity images (Liu et al., 2015).

[6] In this paper, the authors refer to Disparate Treatment as 'a type of individual fairness'. This is not a definition in common use.

2.6.2 *During Training*

A less intrusive approach, rather than modifying the model outputs directly, is to reduce the hypothesis space of valid models so that the outputs are fairer. Some methods to achieve this are given below.

One of the more direct ways to enforce fairness is to modify the loss term directly. An example of this is Zafar et al. (2017b) where the covariance between the protected attributes of a user and the signed distance from the remaining features to the decision boundary (the predicted score $d(\boldsymbol{x})$) are used as an approximation of DP, as shown in equation (2.7). The authors then minimise this according to one of two strategies. In the first, they optimise for accuracy under the fairness constraint that the fairness measure should be within some region of tolerance. In the second, they optimise for fairness under the constraint that a

$$\begin{aligned} Cov(s, d(\boldsymbol{x})) &= \mathbb{E}[(s-\bar{s})d(\boldsymbol{x})] - \mathbb{E}[(s-\bar{s})]\bar{d}(\boldsymbol{x}) \\ &\approx \frac{1}{N}\sum_{i=1}^{N}(s_i-\bar{s}) \cdot d(\boldsymbol{x}_i) \end{aligned} \tag{2.7}$$

A further example of modifying the loss function directly is to follow Quadrianto and Sharmanska (2017) who noticed that enforcing fairness constraints is an application of Distribution Matching — the distribution of outcomes across groups should be identical (or up to a small tolerance). They use a modified Support Vector Machine (SVM) from the Learning Using Privileged Information (LUPI) paradigm to ensure the sensitive feature is not used during inference, but is available during training. They then pose a question about how much fairness to apply. In their experiments, a fairness-utility trade-off is present and the authors suggest a human should be responsible for selecting how much fairness (within a legal limit) to apply. This concept of bringing accountability into automated decision making is an important, although often overlooked addition.

A subsection of modifying the loss term is incorporating adversarial training to minimise the empirical risk of predicting the target, but maximising the empirical risk when predicting the protected attribute. This approach overlaps in terms of implementation with section 2.6.3.2. There are parallels between what we are trying to achieve with fairness constraints and the work that is being progressed in domain adaptation. One of the major breakthroughs in this work was adding a 'gradient reversal layer' introduced by Ganin et al. (2016) and . This has been applied in many fields including fairness. The gradient reversal layer is applied at the conceptual input to a discriminator network (adversary) and allows the *min-max* game to become a direct minimisation, as gradient descent is applied to the discriminator network, but gradient *ascent* with regard to the discriminator loss is applied to the network prior to this point. This framework was then used in Edwards and Storkey (2016) for making a representation that censored a sensitive attribute. This was built on by Beutel et al. (2017) who applied the technique explicitly to fairness, demonstrating that this method is particularly useful even with very small amounts of data. Other papers have tried to build on this work, such as Wadsworth et al. (2018) who, instead of using a representation as the input to the adversary, use the soft output from the predictor.

Other approaches during training include loss reweighting (Kamiran and Calders, 2012). During training the proportion of samples in each group related to a combination of S and Y is calculated. Then, for each group, a weight is assigned (the same weight for all samples in the group) — assigning an instance weight determined by the number of samples in each Y/S group combination as in equation (2.8), where W is a function providing the group weights and N is the number of samples. During training of a classifier, this group weight associated with each sample is multiplied by the loss for each sample.

$$W(s, y) = \frac{N_{X|S=s} \cdot N_{X|Y=y}}{N_X \cdot N_{X|S=s,Y=y}} \tag{2.8}$$

Although this does not specifically minimise any fairness criteria, it assists in making the model 'pay attention' to under-represented samples. A parameterised variation of this approach is presented in Yan et al. (2022) where they propose finding instance weights via meta-learning rather than precomputing the weights such that a specified fairness criteria is satisfied.

Another approach is to view the fairness constraint as a regularisation term, as in Chuang and Mroueh (2021). Here the authors use the data augmentation technique Mixup (Zhang et al., 2018b) to interpolate between the training samples and a sample from the alternative protected group in the case of DP (they also extend this to a procedure for EOdds). They measure the gradient of the model on the interpolated features and try to minimise the inner product of the Jacobian on interpolated samples and the difference between the two original samples. The intuition provided for this approach is that the gradient of the model should remain consistent throughout different interpolation values if the model is indeed not sensitive to the protected attribute.

Lastly, sampling techniques can be used to promote fairness in a model. Approaches such as Roh et al. (2021a) effectively 'upsample' from under-represented groups during mini-batch selection during training of a NN model. They achieve this by adaptively optimising the selection rates for each group to promote the training of a model that satisfies the selected notion of fairness. This idea was developed further in the follow-up paper by the same authors in Roh et al. (2021b). Here they additionally consider a scenario where only a subset of the training data contain records of the protected attribute value.

2.6.3 *Pre-Training*

The final broad category of fairness interventions is prior to training a classification model. This encompasses a number of interventions, including producing additional samples and learning a new representation of the data. The element that connects these approaches is that they all modify the underlying dataset and is the intervention method that is predominantly developed in this thesis.

2.6.3.1 *Fair Representations*

Work in this field was pioneered by Zemel et al. (2013). In their work, they argue that fairness can be achieved through representation learning. The authors suggest that the population in X should be mapped to one of K-prototypes that reside in the same space as the original data. This is presented as a discriminative clustering model, where each of the prototypes is a centroid. The 'job' of the model is to position all K-prototypes is positions that satisfy the three competing desiderata. The first is that a linear classification model should be equally as predictive when provided with a prototype as the model would be if provided with the original data. Secondly, the prototype that a sample is mapped to should be as close as possible to the original sample. Lastly, individuals from each protected group should be equally likely to be assigned to a prototype. Clearly, these goals are in conflict.

Another approach is to explicitly change the input features, such as in Feldman et al. (2015). This paper compares the probability distributions of individual features across protected groups and seeks to rectify this. Enforcing $P(\bar{\boldsymbol{x}}|s = 0) = P(\bar{\boldsymbol{x}}|s = 1)$ where $\bar{\boldsymbol{x}}$ is a modified version of the original feature $\boldsymbol{x}$. In this approach, each individual feature is ranked within sensitive subgroups, then the feature distributions are shifted to retain the same rank while having an identical distribution for each subgroup such that $P(s|\bar{\boldsymbol{x}}) = 0.5$.

An end-to-end approach to learning fair representations was presented in Louizos et al. (2016). In this paper the authors present two methods using the framework of a variational autoencoder (VAE) to produce a fair embedding Z. The first is an 'unsupervised' model where a VAE model produces an embedding $\boldsymbol{z}$ conditioned on $\boldsymbol{x}$ and S, and aims to reconstruct the input $\hat{\boldsymbol{x}}$ by conditioning on $\boldsymbol{z}$ and s. During training, the embedding Z is encouraged to be independent of S by encouraging the embedding to be close to a prior that is independent of S. In practice, they additionally add a penalty based on distribution matching (they use Maximum Mean Discrepancy (MMD)) to encourage the distributional embedding of each group to be identical. The second approach is a 'semi-supervised' method. Here the authors take an additional step by further factorising the embedding based on the outcome, producing a second embedding $\tilde{\boldsymbol{z}}$ conditioned on $\boldsymbol{z}$ and y, and the reconstruction of the first embedding $\boldsymbol{z}$ conditioned on $\tilde{\boldsymbol{z}}$ and s. Then, in the case that the target label Y is not observed, a predicted value of y conditioned on $\boldsymbol{z}$ can be used.

2.6.3.2 *Adversarial Learning*

One benefit of learning fair representations is that there is the potential for transfer learning — producing a fair representation that can then be used for a multitude of (data appropriate) similar tasks. The idea of transfer learning in the context of fair representations was explored further in Madras et al. (2018a), using a similar framework to that of Beutel et al. (2017), which in turn is based on Edwards and Storkey (2016) and Ganin et al. (2016), both discussed in section 2.6.2. They demonstrate that fair representations can indeed be used to predict other features and give a more robust set of experiments than presented in Zemel et al. (2013) (which mentioned the potential for transfer learning as a motivation for learning a fair representation). They give motivation for this by defining two roles. There is a data collector role who obtains the data and sells it, there is also a data vendor role who purchases the data and uses it to create models. They argue that the data vendor may not care about fairness, and as such the responsibility falls on the data collector to amend the data to a new, fair representation. This provides a difficulty for the data collector as they do not know what the vendor intends to do with the data. This is the motivation for learning a fair, transferable representation. The model is constructed as an autoencoder, where a bottleneck layer is used as input to an adversarial classification network with a gradient reversal layer between the bottleneck layer and the adversarial network. In addition, there is a task classification network from the same bottleneck layer. During training, all three losses (reconstruction, adversarial, and task) are minimised. The learned representation for downstream classification tasks is then the output of the bottleneck layer that was used during training.

Another adversarial approach is based on Generative Adversarial Networks (GANs). In the GAN framework, a generator model produces a new sample conditioned on random data and a discriminator model determines between the generated sample and a genuine sample from the data. The training of the dual models in tandem takes place as a *min-max* game. On the one hand, we have the generator trying to produce a sample which is rich enough to accurately model the data distribution. On the other hand, we have a discriminator that is similarly trying to determine the data distribution so that it is able to determine which samples are drawn from it. The proposed use of GANs in fairness literature is to produce new, diverse samples on which a classification model can be trained.

In Xu et al. (2018), the authors use a typical GAN set-up, but have an additional discriminator to not just determine if the data is real or not, but to also query whether the data generated is fair, in this case, regarding demographic parity, by predicting the value of s for the generated sample. A second paper, Sattigeri et al. (2019) use a similar approach, but in addition to generating plausible fair samples, they also produce task classification labels. Both new samples and labels are encouraged by a discriminator to either satisfy DP or EOpp. The theme of building a more diverse dataset was built on by Sharmanska et al. (2020). They use a conditional GAN model to increase the representation of under-represented groups in the training dataset by generating new data conditioned on existing samples within the dataset, and the .

Finally, a wholly alternative adversarial NN approach is provided in Zhang et al. (2018a). In this approach the network has two classification heads. One predicts the task, the other the protected attribute using the logits from the classification network as an input. During training, the gradients for the task update are projected onto the gradients for the adversarial loss. Some guarantees are provided that the resulting gradient update should improve task performance, but at the same time, actively harm the adversarial model, resulting in a debiased classifier.

2.7 Causal Inference

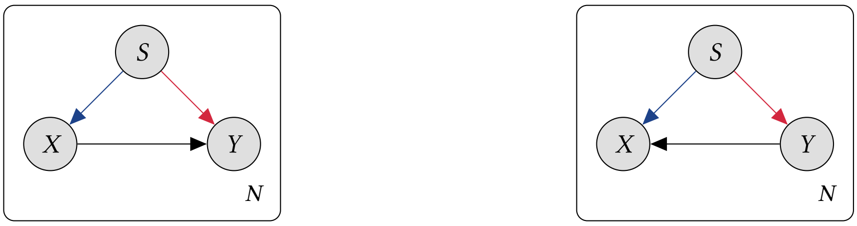
The final section in this chapter is about Causal Modelling, which is influential later in this thesis, notably in chapter 4 where a particularly narrow form of this area is emulated. Causality is a framework to describe cause and effect, rather than solely correlations, modelling these relationships in a Directed Acyclic Graph (DAG) structure. With the acknowledgement that fairness is difficult to solve arithmetically, methods to incorporate subject matter expertise are being explored. Causal models are appealing in this regard because they allow for an explicit causal relationship to be accounted for rather than relying on correlation.

The authors of Dedeo (2014) argue that without understanding the causal relationship between attributes, then it becomes particularly difficult to differentiate between innocent relationships and those which at first glance may appear innocent, but when the socio-economic background of those attributes is understood, they might infer a less innocent relationship.

Stemming from the work of Pearl (2009), causal models are an attempt to model cause and effect. An example would be atmospheric pressure and the position of the needle on a barometer reading. We know that the two are linked and our data about this will demonstrate a high correlation between observations, but the correlation does not imply causation. While we know that changing the pressure will effect the barometer, moving the needle on the barometer will not effect the pressure in the room. The benefit of viewing the world in this way is that we can transparently interpret why decisions have been made. Clearly, the relationships between features are complex, but we can utilise experts from the domain we are trying to apply our model to. This is a nice feature given that fairness itself is domain specific. While this may seem simple on the surface, it is highly complicated to correctly model the world. For example, not all features are captured. There may be an unobserved feature that confounds two features, so while they may look as if they are connected in some way, they are actually both reflective of the unseen confounder. An example of this is height and level of education. On the surface we could draw a correlation that the taller (on average) the population is, the higher the level of education. This can be observed by visiting any primary or secondary school, but we are missing a confounder — age. Furthermore, there can be multiple confounders that affect different sets of features. While not insurmountable, this is nonetheless a very labour intensive approach. In many ways, if this approach is fully realised, it is the gold standard for ethical models.

2.7.1 *Modelling the Data*

A useful first step when approaching any ML problem is to spend time understanding the data. This is a nontrivial stage that at times can be overlooked. Prior to modelling the relationships



(a) Causal relationship from X to Y, with possible affects of S on both.

(b) Causal relationship from Y to X, with possible affects of S on both.

Figure 2.3: Possible paths for a protected characteristic to influence data. In both models the presence of S acts a possible descendent of both the input features X and the target variable Y.

between features, it can be useful to populate a datasheet, as proposed in Gebru et al. (2021) to understand some of the scope and limitations of any particular dataset.

In terms of algorithmic fairness, the causal relationship of particular concern is between the protected attribute S and the remaining features. Given the nature of protected attributes — immutable properties of an individual — these are almost always descendants, as opposed to children, of either the input features X, or the target outcome Y, or indeed both. Depending on the relationship between the variables, different types of discriminatory practices may enter the dataset. These potential relationships from S to X and from S to Y are shown in blue and red respectively in figure 2.3. In figure 2.3a the target label is determined by the features X. This is a relationship that is closer to that of decision systems such as loan applications, hiring decisions, or recidivism prediction — outcomes are based on observations in the data. The alternative relationship from Y to X is shown in figure 2.3b. This model is closer to the relationship in image classification tasks — detecting the presence of an item or characteristic in an image.

2.7.1.1 *S and X*

The first potential relationship of concern is between the protected attribute and the other input features as modelled by the blue arrow in figure 2.3. This characterises that the features of an individual are the source of bias in the dataset. This relationship can be forged through a number of biases, such as institutionalised bias and measurement bias (Tolan, 2019). If this type of bias is present, it may be reasonable to adjust the method of encouraging fairer outcomes. For example, fair representations may be a sensible choice as they manipulate the features of a dataset used for training an auxiliary model.

2.7.1.2 *S and Y*

An additional source of concern is that the outcomes themselves are causal dependent of the protected attribute. This may occur if direct discrimination is present in the data (Tolan, 2019). If this type of bias is present, then broader questions about the use of this data for training a machine learning model to emulate this process should be asked. In the case that this process must be emulated, then more direct interventions, such as the *during*, or *post* training models may be more appropriate.

2.7.2 *Counterfactual Fairness*

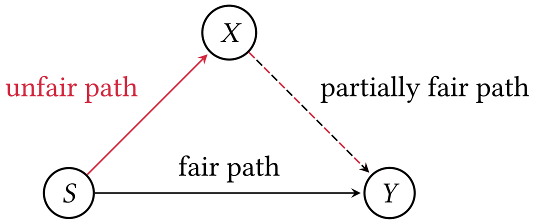


Figure 2.4: An example of Path-Specific causal modelling. The path from S to X (in red) is considered *unfair*. The relationship between X and Y is then

Fairness interventions using causal modelling were first presented in concurrent works Kilbertus et al. (2017) and Kusner et al. (2017).

$$P(\hat{y}|x, s) = P(\hat{y}|x, \text{do}(s = \tilde{s}))\ \forall\ x \in X,\ \tilde{s} \in S \tag{2.9}$$

In Kilbertus et al. (2017) the authors introduce a counterfactual notion of fairness, that is based on interventions described in equation (2.9), which they use to determine if a model displays *proxy discrimination*. They define two types of variables related to fairness:

1. *Proxy Variables*: These are causal descendants of the protected attribute that are used in a predictive model.
2. *Resolving Variable*: A proxy variable, where the causal effect of he protected attribute is considered valid to use in a predictive model.

In addition to proxy discrimination, the authors discuss *Unresolved Discrimination* — when there is a path from S to a variable, V that is not blocked by a resolving variable, and when V itself is not a resolving variable. The authors assume an additive, linear relationship between variables, allowing for the effect along specific paths to be corrected for using equation (2.10).

$$V = v(X - \mathbb{E}[X|\mathrm{do}(P = p')]) \tag{2.10}$$

During the construction of a predictive model, the authors propose either restricting the variables available for training a model, or explicitly correcting for the influences along specific paths.

In Kusner et al. (2017) the authors propose defining *Counterfactual Fairness*, which is evaluated at an individual level. A classification model displays Counterfactual Fairness if equation (2.11) is satisfied.

$$P(\hat{Y}_{S\leftarrow s}(U) = y|X = x, S = s) = P(\hat{Y}_{S\leftarrow s'}(U) = y|X = x, S = s) \;\; \forall \; x \in X, \;\; s' \in S \tag{2.11}$$

In this definition, U represents the set of ‘background’ variables that are not captured by the graphical model. The implication of the value of s on the left hand side of the conditioning is that the prediction for the original sample, and a counterfactual sample, intervened on with regard to s, should be identical. This is achieved through a three-step process:

1. *Abduction*: Compute a the prior distribution over U for a given prior.
2. *Action*: Perform a do-operation on the protected attribute, replacing relevant equations of the Structural Equation Model with intervened values.
3. *Prediction*: Compute the distribution of the remain variables using the posterior distribution and intervened values from steps (1) and (2).

The crucial difference in the approaches of Kilbertus et al. (2017) and Kusner et al. (2017) is that the former is intended as a framework to reason about fairness and producing a set of variables on which a classifier can be trained, whereas the latter is a distinct method that requires counterfactual modelling at an individual instance level.

The idea of resolved variables in Kilbertus et al. (2017) was expanded on in Chiappa (2019). They note that not all the effects of a sensitive attribute on the outcome are potentially discriminatory. They give an example of the Berkley admissions data that was suggested to be discriminatory against women. They note that women were applying in greater proportions to classes with low acceptance rates, thus the influence of gender on the class applied for is not discriminatory and should be taken into account to learn a highly predictive model. This is similar to the idea first mentioned in Pedreshi et al. (2008) that there is a difference between sensitive attributes and potentially discriminatory attributes. In Chiappa (2019) they use the power of a causal model to isolate this effect along specific pathways noting ‘approaches based on statistical relations among observations are in danger of not discerning correlation from causation, and are unable to distinguish the different ways in which the sensitive attribute might influence the decision’. This paper views unfairness as the presence of an unfair causal effect of S on $\hat{Y}$. This idea is not new. It is specifically mentioned in Kusner et al. (2017) that ‘a decision is unfair toward an individual if it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute were different’. This assumes that the entire effect of S on $\hat{Y}$ is problematic. The path-specific approach uses the same definition, but modifies the ending to be ‘... counterfactual world in which the sensitive attribute *along the unfair pathways* were different’. They achieve this by measuring the effect of s along unfair pathways and disregarding it. In the simple case in figure 2.4 where the direct effect of s on y is fair, but the effect of s via m is unfair, our goal would be to remove the effect of s along unfair pathways. In this case (and in the case of Kusner et al. (2017) the goal is to achieve fairness in a counterfactual world as described above — this can be seen as a form of individual fairness.

3 | Summary of Contributions

The following is a summary of the main contributions in this thesis. All presented approaches aim to address the problem of biases, with respect to a protected attribute *S*, which are captured in a dataset that will be used to train an automated decision support system. The proposed solutions all improve upon the fairness of such a system via transformations of the data. Furthermore, all of the proposed solutions provide additional passive information that provides insight into the changes required to meet the various statistical definitions of fairness.

3.1 Replaying Biases

The initial approach taken in Thomas et al. (2021) (expanded in chapter 4) is to build on the concept of fair representations. Fair representations are a preprocessing transformation step that, when applied to the data, promote fairer outcomes in a system that is unrestricted by fairness constraints. This is achieved by removing as much information about the protected attribute, *S*, as possible from the representation. The main idea is that if a fair representation is sufficiently devoid of information about *S*, then this can be viewed as a disentangling procedure, separating *S* from the remaining features. To encourage as much information as possible that is *not* about the protected attribute to be retained, this approach is trained in an unsupervised manner, with the original sample being constructed from a combination of the fair representation and the (known) protected attribute value. The part of this approach that 'reconstructs' the original data from the fair representation and the protected attribute can be seen as *replaying* the effect of *S* on the remaining features, thus replaying a suite of bias-inducing processes. In the text, a connection is drawn between the unsupervised aspect of fair representation learning as a disentangling procedure; and a limited form of causal modelling where a counterfactual sample can be drawn with respect to a specific, predefined treatment variable. Then, this reconstruction procedure is repeated with alternative plausible values for the protected attribute. This produces samples that are similar except for the replayed effect of the protected attribute. A decision support system can then be interrogated, allowing for the question ‘would two individuals, who are similar in all ways except for those influenced by a protected attribute, receive a similar outcome?’. A practical mechanism is then introduced to promote more equal outcomes over time by identifying individuals who do not receive similar outcomes in the scenario described.

3.2 Fair Representations in the Data Domain

In the previous work, biases are replayed to generate cross-domain samples, for example, an image of a man could be translated into an image of a woman via a representation that is disentangled from S. The approach taken in Quadrianto et al. (2019) (chapter 5) is fundamentally different. A classification model is designed with an explicit data transformation stage, where the transformed data is constrained to both exist in the same feature space as the input (e.g. a space of RGB values for images), and also to be independent of protected attribute. This explicit data transformation stage to a latent feature embedding is used in previous works, but crucially, here we constrain the fair representation to be the same data domain as the input. The approach here is to translate an input into a new domain that not only *isn't present* in the data, but also *can* be represented within the data domain. This is challenging as there are no examples to aid in training. Instead, the approach taken is to try and satisfy four objectives. The first is that the representation should contain information relevant to the classification of the sample. Second, the representation should be similar to the input. Third, the representation should be statistically *independent* of the protected attribute according to the novel application of Hilbert-Schmidt Independence Criteria (HSIC) which measures statistical independence between two random variables. Lastly, the residual of the representation when taken away from the original input should be statistically *dependent* on the protected attribute.

The benefit of constraining the fair representation to exist in the data domain is that the transformations made by the model become inherently transparent. This transparency, in turn, presents several opportunities for greater accountability and monitoring of the model in a deployment setting.

This chapter has an addendum that seeks to bridge the gap between chapter 5 and chapter 6. This is done by demonstrating an unsupervised version of the training objective with some practical alterations.

3.3 Fair Representations in the Data Domain with Controllable Replayed Biases

The final work of this thesis is presented in Kehrenberg et al. (2020b) (chapter 6). The motivation for this approach is to improve on the work of the previous chapter. There, the assumption is that all samples can be decomposed into a 'fair' and 'unfair' component that, when simply added together, forms the original input. This work loosens this restriction and allows the original input to be an arbitrarily complex function of the two components. Two methods to achieve this are presented. The first proposed method is a conditional VAE (cVAE) which uses adversarial training to constrain the latent embedding of a variational autoencoder (VAE) framework. The second proposed method 'partitions' the latent space in a lossless manner using an Invertible Neural Network (INN). The procedure for the cVAE is given below, followed by the training procedure for the INN.

When training a VAE, the aim is to model the data distribution by conditioning on a learned posterior distribution $P(Z|X)$. During training of a VAE, in this work the posterior distribution is encouraged to be independent of S using an adversarial training approach. During reconstruction, the model that samples from the data distribution is additionally conditioned on a One-Hot encoding of the protected characteristic to 'replace' the information removed by the adversary. During inference, this encoding is given with all 0-values, so that the reconstruction is not given any value for the protected characteristic — a 'null-sample' is drawn.

Typically, during training of the INN, a multivariate Normal-, or an array of independent Normal-distributions with 0-mean and unit standard deviation receive a series of information-preserving transformations such that the likelihood of modelling the data distribution is maximised. However, because the network is invertible, we can train the network in the reverse direction and try to fully capture the distribution of samples from the dataset by transforming the data to match the prior distribution. Adversarial training is used to encourage a small, predefined, independent selection of the variable to contain all the information about the protected attribute.

To then generate a transformation of the data that remain in the data domain but belong to no particular protected group, a process coined *Null-Sampling* is used. A data sample is fed into the INN, resulting in an embedding. The predefined region of the embedding associated with the protected attribute is then set to 0, the mean value of the prior distribution. The amended embedding is then fed through the INN but in reverse, resulting in the exact image, but with the protected attribute manipulated to be the mean of the training dataset.

3.4 List of Publications and Author Contributions

This thesis is based on three publications, corresponding to chapters 4 to 6. The following is a detailed listing of all the individual author contributions.

3.4.1 *Publication 1*

A shorter version was published as a conference paper:

Thomas, Oliver, Miri Zilka, Adrian Weller and Novi Quadrianto (2021). 'An Algorithmic Framework for Positive Action'. In: *Equity and Access in Algorithms, Mechanisms, and Optimization.* EAAMO '21. –, NY, USA: Association for Computing Machinery. ISBN: 9781450385534. DOI: 10.1145/3465416.3483303.

Contributions:

- I conceived the idea of capturing multiple biased effects in a single model and then developed the strategy to isolate these behaviours. I wrote the entire codebase and ran all experiments. I led the discussions and wrote a significant portion of the text.
- M. Zilka was a discussion partner.
- A. Weller acted as an editor of the conference proceedings.
- N. Quadrianto was a discussion partner and helped to strengthen the technical contributions.

3.4.2 *Publication 2*

Quadrianto, Novi, Viktoriia Sharmanska and Oliver Thomas (2019). 'Discovering Fair Representations in the Data Domain'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Computer Vision Foundation / IEEE, pp. 8227–8236. DOI: 10.1109/cvpr.2019.00842.

Contributions:

- The initial idea came from a brainstorming session with my coauthors. I then developed the idea further, writing code, implementing baselines, running experiments, and presenting analysis. All authors contributed to the text.
- V. Sharmanska helped with writing the code, led the image-based experiments, and participated in discussions.
- N. Quadrianto helped with writing the code and led the initial discussions.

3.4.3 *Publication 3*

Kehrenberg, Thomas, Myles Bartlett, Oliver Thomas and Novi Quadrianto (2020). 'Null-Sampling for Interpretable and Fair Representations'. In: *Computer Vision – ECCV 2020.* Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox and Jan-Michael Frahm. Cham: Springer International Publishing, pp. 565–580. ISBN: 978-3-030-58574-7.

CONTRIBUTIONS:

- I developed the problem statement of producing fair representation in the data domain in an unsupervised manner and proposed that this could combat distribution shift. I designed the coloured MNIST experiments, wrote some code, ran some experiments, and structured and wrote a significant portion of the text.
- T. Kehrenberg suggested using an INN for my proposed problem, formalised a reference set to tackle severe distribution shift, wrote a large proportion of the code, and wrote a significant portion of the text.
- M. Bartlett wrote a large proportion of the code, ran most of the experiments, and structured and wrote a significant portion of the text.
- N. Quadrianto gave feedback on progress and suggested directions to explore.

Part II

PUBLICATIONS

This part comprises two peer-reviewed publications and one work-in-progress journal extension of a third peer-reviewed publication. They are reproduced here with minimal changes except where clearly marked.

4 Paper 1: An Algorithmic Framework for Positive Action

Authors: Oliver Thomas[1], Miri Zilka[1], Adrian Weller[2,3] and Novi Quadrianto[1,4]

Affiliations:

[1] Predictive Analytics Lab (PAL), University of Sussex, Brighton, UK

[2] University of Cambridge, Cambridge, UK

[3] The Alan Turing Institute, London, UK

[4] BCAM – Basque Center for Applied Mathematics, Bilbao, Spain

Conference: *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2021

DOI: `10.1145/3465416.3483303`

Extended version submitted to: Machine Learning (Journal) *Under Review*, 2022

4.1 Abstract

Positive actions are additional voluntary steps that can be taken to address an imbalance of opportunities for individuals belonging to groups of the population that share protected attributes such as race, disability or gender. They are defined within anti-discrimination legislation as a legal approach to address under-representation in the workplace for members of these groups and create more equal outcomes over time. Within this theme, we propose a novel algorithmic fairness framework to identify candidates to receive positive action outcomes. The aim is to advance equal representation while respecting anti-discrimination legislation and equal treatment rights. We use a counterfactual fairness approach to distinguish between two behaviours within a decision system. The first is direct discrimination — outcomes that are directly influenced by group membership — which we aim to identify and remove. The second, which we refer to as structural discrimination, is where outcomes depend on the causal consequences of a protected attribute. During inference, our aim is to determine which candidates have likely suffered from structural discrimination and to recommend them for positive action steps.

4.2 Introduction

Allocating limited resources, such as jobs or university placements, among individuals requires assessing their suitability for the role as part of the candidate selection process. At the same time, machine learning (ML) systems are becoming more capable. It is natural then that such systems are increasingly being used to inform, support, or even make decisions directly within consequential domains such as candidate selection, which affects millions of lives (Barocas and Selbst, 2016). To ensure that these systems remain trusted, it is important that the selection process is fair and, in addition, that positive outcomes are fairly distributed within the population. Therefore, it is necessary to consider how the notions of fair process and fair outcomes translate into algorithmic decision support frameworks (Wachter et al., 2020; Xiang, 2021).

In the E.U., U.S., and U.K., among others, anti-discrimination legislation dictates that a fair selection process requires *equal treatment*. To achieve this, protected attributes, for example, gender and race, are not to be considered within the decision-making process without a good reason (Dwork et al., 2012; Bent, 2019). In addition, we have the concept of a *fair outcome*, that outcomes should be assigned consistently throughout the population based on the merits of the applicant. Simply ignoring the protected attributes, however, guarantees neither a fair process nor fair outcome (Pedreshi et al., 2008; Harned and Wallach, 2019; Wachter et al., 2020; Simons et al., 2021; Xiang, 2021), except in trivial cases.

Decision support algorithms are typically trained using data that document previous decisions. An example of this is to use a candidate's CV to anticipate how likely they are to be successful when interviewed based on the outcomes of previous candidates. Without due diligence, the resulting algorithm may learn to disproportionately predict positive outcomes in favour of applicants who most closely resemble the existing workforce. This may lead to a lack of opportunities for groups that have historically been under-represented in the workforce, in contrast to those belonging to a majority group[1](Kamiran and Calders, 2012; Kallus and Zhou, 2018). Statistical disparities in the training data can arise from two mechanisms: (i) unequal treatment; or (ii) equal treatment when the status quo in the environment itself is not neutral. The former occurs when the data contain discriminatory past decisions. The latter, when historically under-represented groups struggle to compete with the majority under a standard *equal treatment selection process* — a selection

[1] The group enjoying an advantage is not always the majority. We aim for clarity of exposition when referring to the over-represented group as the majority.

process that is 'blind' to the applicant's protected attributes. Often, the statistical disparity in the training data and, as a result, in the model's prediction, is a combination of both.

When evaluating whether an outcome is fair, measuring *Demographic Parity* (DP) — the difference in the probability of positive outcomes between subgroups — as the definition of fairness is appealing. It is an intuitive concept, but it is often impractical. Except in restricted scenarios, enforcing it hinders the accuracy of the model's predictions. Additionally, this approach often does not align with anti-discrimination legislation. A common alternative is to instead promote an algorithmic fairness constraint *Equal Opportunity* — that True Positive rates should be equal across subgroups — as it better aligns with the notion of equal treatment, however, this often means maintaining a disparity in the positive[2] outcome rates (Hardt et al., 2016; Wachter et al., 2020).

Anti-discrimination legislation acknowledges the need to bridge the gap between equal treatment and equal representation. The Equality Act 2010 (UK) defines *positive action* as 'lawful measures taken to encourage and train people from under-represented groups to help them overcome disadvantages in competing with other applicants'[3]. Examples of positive action include, but are not limited to: additional training opportunities and mentoring programmes available to an under-represented group, targeted advertising, outreach, networking, and bursaries. For example, Target Oxbridge is a free, UK based programme that 'aims to help black African and Caribbean students and students of mixed race with black African and Caribbean heritage increase their chances of getting into the Universities of Oxford or Cambridge' (Rare Recruitment, 2021). Policies designed to meet the specific needs of under-represented groups may also be considered as positive action. The European Research Council introduced an automatic extension of eligibility only for women with children when applying for grants[4] and the UK Department for Education has made it a requirement to improve the outcomes of disadvantaged students at universities by providing positive action in the form of additional tutoring, additional summer schools, and targeted recruitment for academic vacancies (Donelan, 2021). The action taken must be 'proportionate' to both the extent and the longevity of the under-representation and the barriers experienced by the under-represented group.

[2] Throughout this paper, we use the terms 'accepted', 'successful', and 'positive' interchangeably when referring to outcomes.

[3] European legislation defines positive action similarly. In the US, similar measures can be employed under affirmative action; however, the definitions do not completely overlap.

[4] These measures are included in the European Research Council's Gender Equality Plan for 2021-2027.

We argue that incorporating the notion of positive action within decision support algorithms respects anti-discrimination legislation while promoting equal representation and equal treatment rights. In this work, we propose a novel algorithmic fairness framework to identify *positive action candidates*. These are individuals who would be rejected under a standard equal treatment selection process due to an earlier disadvantage experienced because of their under-represented group membership. A natural question may be why a novel approach is needed; why not identify the top-rejected candidate from an under-represented group as a positive action candidate? We compare our approach with this baseline in section 4.4.2.1 (figure 4.4).

Our goal is not to produce a fair system but rather to produce an accurate system that also provides additional insight into the deployment setting so that equal outcomes may be achieved over time. Instead of just making every input invariant to some sensitive attribute, we want to make predictions that are as accurate as possible but also ask counterfactual questions about the outcomes that may have occurred in a counterfactual world where the individual has other values for their protected attributes, e.g., what if a white applicant were Latino. In this way, we are not compromising the efficacy of the prediction model, but we are also able to provide additional information about who, to our best guess, appears to have been overlooked, perhaps rationally, depending on the application, due to some protected attributes. These individuals are the ones we want to identify.

4.3 Background

We begin this section with an overview of applicable fairness definitions, before an overview of counterfactual modelling and the connections to our approach. There is then a discussion of related research areas and work.

4.3.1 *Definitions*

In this work we discuss subgroups with respect to *protected attributes* — characteristics that, by law, must not be the basis for discrimination. These include, but are not limited to, race, gender, age, religion, and disability. We define a protected subgroup as an under-represented group separated from the majority by the perception of one or more protected attributes. For example, women in the engineering profession are under-represented when compared to their representation within the population. In the context of a decision support system, we may observe a statistical disparity — a disproportionate positive outcome (e.g., hiring or admission) rate — in favour of an existing majority, compared to a protected subgroup. This can be a result of the model being trained on previous discriminatory decisions, but it can also be the result of a genuine statistical difference in the input features (e.g., grades or qualifications) as observed in the data. In this work, we define bias as a mechanism by which a statistical disparity between a protected subgroup and the majority is created or exacerbated. Bias within the decision-making process will affect the decision outcome. Bias that occurred earlier may affect the features on which a decision is made. To expand the discussion of bias, it may be useful to refer to the framework presented by Friedler et al. (2021), which defines three spaces — the *construct space*, *observed space*, and *decision space* — and uses the mappings between them to formalise several definitions of bias.

The *construct space* represents the '*ground truth*' — an unobservable space that correctly captures differences between individuals with respect to a task; the *observed space* represents the measurable features for consideration, and the *decision space* represents the outcome (Friedler et al., 2021). For example, intelligence resides in the construct space, measured IQ resides in the observed space, and acceptance or rejection from the International Mensa Club resides in the decision space.

The observed space is an estimate of the construct space. The decision space, in turn, is an estimate of the construct space based on the observed space. In any application, we are required to make assumptions regarding the mapping between spaces. Friedler et al. (2021) refer to these assumptions as *worldviews*, highlighting two common worldviews, we're all equal (WAE) and what you see is what you get (WYSIWYG), which are often in tension with each other. WAE assumes that any disparity between subgroups in the observed space is due to structural bias — an incorrect mapping between the construct and the observed space. WYSIWYG on the other hand, allows for a disparity between protected subgroups, assuming the observed discrepancies are a true reflection of disparities in the construct space. In this work, we adopt a 'hybrid' worldview (section 4.4.1.1) that allows a version of both worldviews to coexist by further delineating the construct space.

To better understand the potential for statistical disparities in the data, we discuss the specific types of bias that we attempt to address in this work. *Sample selection bias* originates from training on a non-representative sample of the population (Tolan, 2019). *Label bias* occurs when the dataset contains past discriminatory decisions (Wick et al., 2019; Jiang and Nachum, 2020). Mitigation efforts that independently consider selection bias (Kamiran and Calders, 2012; Agarwal et al., 2018), or label bias (Calders and Verwer, 2010; Jiang and Nachum, 2020; Kehrenberg et al., 2020a) are available. In addition, biases can also be introduced from outside the environment that we control, for example, outside the training population, measurements, and learning algorithms. Recognising that these biases may not occur in isolation, our proposed framework aims to acknowledge and mitigate a broad range of biases rather than focussing on addressing a solitary issue. This includes bias that cannot normally be mitigated by an automated rejection / acceptance model while respecting anti-discrimination legislation and the right to equal treatment.

4.3.2 *Counterfactual Modelling.*

To identify positive action candidates, the subset of rejected candidates from an under-represented group who have the potential to succeed, we take a counterfactual approach. A counterfactual outcome is a hypothetical outcome for a scenario that is identical in all respects except for a specific, well-defined change and its causal consequences (Hume, 2000; Miller, 2019). In the context of this work, we focus on counterfactual scenarios with respect to a change in a protected attribute and distinguish between two types of counterfactual questions:

Question 1: Would the outcome change if *only* the protected attribute were different?

Question 2: Would the outcome change if the protected attributed *and its causal consequences* were different?

For example, if a female applicant is not invited for a job interview, we can ask the following two questions: If your CV was identical but the application *appeared* to be from a male applicant, would she be invited to interview?[5] If she had been *born* male, experienced life as a male, and then applied for the same job, would she have been invited for an interview? The second counterfactual question is critical to our approach, as it is used to identify candidates for positive action. We use the first counterfactual question to detect and mitigate label bias.

Our proposal is two-fold. First, identify those who did not succeed but likely would have done so in a counterfactual world where they did not suffer from direct discrimination. Second, identify those who did not succeed but likely would have in a counterfactual world where they had the advantages associated with being a member of the over-represented group.

Counterfactual modelling with protected attributes is a subject that provokes a discussion of the argument that there is ‘no causation without manipulation’(Holland, 1986). This refers to

[5] An experiment by Bertrand and Mullainathan (2004) looked at exactly this question.

the position of some practitioners that only mutable features should be considered as treatment variables and their effects evaluated within a causation setting. This position requires us to refrain from evaluating immutable features, such as protected attributes, as there is no method to evaluate this experimentally. Opponents of this view, including Pearl (2009) argue that causation does not require manipulation of features, but rather requires modelling the relationship between interactions. An example supporting this viewpoint is that we can reason about acts of God, such as earthquakes or volcanic eruptions, without being able to manipulate them directly. Instead, by understanding the relationship between these events and how they interact with their environment, we can make reasonable assumptions about cause and effect.

To model causal interactions, a common tool is a *Structural Causal Model* (SCM), a graphical model whose vertices represent features and whose edges represent the *causal* pathway between them. For example, the first type of counterfactual question would manifest in an SCM as a direct pathway between the protected attribute and the outcome if such a relationship were present. The SCM is a *high-level map* of the data generation procedure. However, a complete structural model is challenging to obtain; they are application-specific and require specification by domain experts, who often have conflicting views. Producing a counterfactual sample with an SCM with respect to a specific *treatment variable* involves intervention and 'playing out' the effect. For more information, see Pearl's Do-Calculus. In practise, if we do not have access to an SCM, we could follow a statistical matching approach and find two as close as possible individuals (differing by the protected attribute) within the data, which assumes little interaction between the variable of interest and the rest of the features. Alternatively, as we are only interested in intervening on a small subset of features, we could create a plausible counterfactual sample using an adversarially trained generative model. Current work looks at a similar problem of unpaired domain adaptation. We often see this in terms of image-to-image translation where methods such as Cyclegan (Zhu et al., 2017) or Stargan (Choi et al., 2020) have excelled. In our case, we treat the protected attribute as a domain, e.g. the domain of male applicants and the domain of female applicants, and try to make a mapping between the domains. The intention is that the mapping learns the complex relationship between the domains, creating a 'likely' counterfactual. Although the SCM would remain hidden, a complicated relationship between the variable of interest and the remaining features can be emulated. In this paper, we compare both approaches and demonstrate the effectiveness of the latter.

4.3.3 *Related Works*

As far as we know, this is the first work to address positive action in the context of a decision support system. However, previous works have looked at related problems. We briefly describe the relevant literature to place the problem of determining positive action candidates in context.

Deferral: Learning to defer is an extension of the 'learning with a reject option' (Hendrickx et al., 2021), or 'selective prediction' (Geifman and El-Yaniv, 2017) paradigm. In both cases, the challenge is to identify which unseen candidates the model is uncertain about. Once identified, in the learning to defer framework, these candidates are directed to a human decision maker at some cost (Madras et al., 2018b; Mozannar and Sontag, 2020). Uncertainty generally fits into two broad categories, *epistiemic* and *aleatoric*, depending on whether the uncertainty is related to the model (epistemic) or the data (aleatoric). Epistemic uncertainty is reducible, and a model can become more certain with respect to this type of uncertainty by training with a greater number of diverse samples. This is sometimes referred to as deferring due to sample novelty, distance, or being an outlier in relation to the training data. Aleatoric uncertainty, on the other hand, is irreducible and represents inherent differences present in the data. Samples deferred due to this are sometimes referred to as being ambiguous. Using this language, deferral can be categorised as the identification of samples that display either epistemic or aleatoric uncertainty with respect to the training data. Our approach, on the other hand, can be categorised as counterfactual-based aleatoric uncertainty modelling.

Deferral poses interesting questions about the practical quantification of uncertainty and could be a potential extension to our framework. However, deferment differs from identifying positive action candidates. The system we are evaluating may be confident in its assessment that a candidate who would be suitable for positive action should be rejected. Deferment models may be able to capture some elements of uncertainty with respect to **question 1**, as they evaluate uncertainty about the target of the decision. We also evaluate **question 2**. We are proposing a similar approach, but the uncertainty is due to a re-enactment of the data generation process. In this sense, the system may be confident of the outcomes, but it is only when we view the outcomes across a range of values for a specific treatment variable that we see the discrepancy. We are using counterfactual modelling to measure aleatoric uncertainty and to make recommendations about how to promote fairer outcomes in this uncertain context. Furthermore, we highlight people who may be at risk of receiving a biased decision based on past data. In this case, deferring to a human who has potentially reinforced past behaviours themselves may not be the ideal outcome. Instead, we suggest taking alternative steps to help the individual.

Actionable Recourse: Another related field is that of recourse. Works in this area, such as Joshi (2019), Ustun et al. (2019) and Karimi et al. (2021) similarly ask counterfactual questions constructed as interrogations of an existing decision system. In these works, the aim is to determine how the world would have had to be different for an alternative outcome to occur. They aim to explain what changes would need to have been made for a rejected candidate to be accepted. This can be challenging if the model recommends that an immutable feature be adjusted. Due to this, the framework *actionable* recourse has developed. In this approach, only features that are mutable are considered valid pathways to produce an alternative result, with further extensions to this framework only considering feasible feature adjustments. Instead, our framework asks, in some ways, simpler and more direct questions of a decision system: If a candidate were perceived to have an alternative protected attribute value, would the outcome be different? And would the outcome change if the protected attribute and its causal consequences were different? We then use this information to determine individuals who have been positively and negatively impacted.

Auditing Systems: This is a broad and multifaceted area, but, in general, auditing aims to evaluate either a dataset (Saleiro, 2018), or a system (Kearns et al., 2018) for potential bias. Examples of auditing systems similar to ours include Black et al. (2020). In their work, the authors take an alternative counterfactual approach based on finding the nearest sample in the data with a different protected attribute and comparing outcomes broadly across population subgroups. Our works differ in motivation, as while the authors use their auditing method to look at which *groups* are most affected, we evaluate which *individuals* are likely to be affected.

4.4 Approach

We propose a new algorithmic fairness framework to advance equal representation while respecting current anti-discrimination legislation and the right to equal treatment. We identify *positive action candidates*: roughly speaking, these are individuals who would be rejected under a standard equal treatment selection process because of an earlier disadvantage experienced due to their under-represented group membership. More precisely, we use generative adversarial modelling to construct counterfactual samples per applicant which are used to determine assignment to one of three outcome labels:

1. Successful applicants; and applicants from under-represented groups who were unsuccessful, but appear to have suffered from *direct discrimination*. That is, those who would have been successful if they had a different set of protected attributes, without considering causal consequences (i.e. **question 1**) — these are accepted;
2. Unsuccessful applicants from under-represented groups who are not in (1), and for whom there exists a set of protected attributes which would have caused them to be successful (considering causal consequences, i.e. **question 2**) — these are flagged as positive action candidates; and
3. Everyone else — these are rejected.

Note that for applicants from the majority group, our approach only alters the outcome from an unconstrained classification model where the protected attribute has a direct treatment effect on the outcome. For all others, the outcome remains unchanged as either accepted or rejected. Additionally, our approach only attempts to identify unsuccessful candidates with strong potential to succeed.

4.4.1 *Positive Action Framework*

The framework we employ follows a rule-based solution to process counterfactually uncertain outcomes. These are outcomes evaluated across a range of counterfactual samples generated from the same individual, differing only with regard to a protected attribute. Predefined rules (table 4.1) account for uncertain outcomes that arise during inference and promote fairer outcomes over time.

A reasonable question may be why only define a subset of rejected candidates as being suitable for positive action as opposed to offering positive action outcomes to all rejected candidates? We welcome this approach, but we assume that there is some cost associated with positive action that makes such an approach prohibitively expensive. Our method identifies those who would have likely succeeded in a counterfactual world. A possible future direction of our work is to incorporate ranking of candidates to meet some budget constraint.

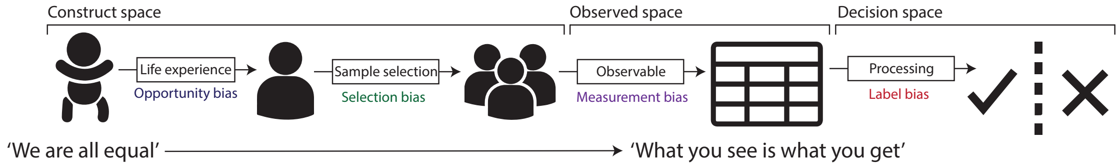


Figure 4.1: A 'hybrid' worldview showing biases potentially introduced at each step of a timeline leading up to a decision. Aptitude is characterised by the infant at the beginning of the timeline and is assumed to be independent of all protected attributes, aligning with the WAE worldview. By the point of observation however, the construct space might have altered. Our 'hybrid' worldview allows for disparity between subgroups, aligning with the WYSIWYG worldview. Opportunity bias, selection bias, measurement bias and label bias can introduce or further aggravate the disparity between the protected subgroup and the majority.

4.4.1.1 *Fairness Worldview*

Where does positive action fit within the technical definitions of fairness? If we consider WAE and WYSIWYG, the worldviews discussed in section 4.3.1, then our approach to remedy unfairness does not fully align with either of these worldviews. In WAE, the assumption is that the observed space can be 'corrected' for structural bias to emulate the equal construct space. In WYSIWYG, the assumption is that decisions based on the observed feature space reflect the construct. In our 'hybrid' worldview, the decisions based on observed features *do* match the construct space as in WYSIWYG, but we also allow for structural biases to exist and be corrected for as in WAE. The hybrid worldview adopted in this work is illustrated in figure 4.1 and as a graphical model in figure 4.2. The crux of our approach is that we expand the *construct space* to include the element of *time*, with the observed space representing measurements of the construct space at points along this additional axis. Consider a set of measurable features X within the observed space $\mathscr{X}$, where $\mathscr{X}$ represents the space of all potential feature values. Each individual sample $\boldsymbol{x} \in X$ is an approximation of its non-measurable construct counterpart $\tilde{\boldsymbol{x}} \in \tilde{X}$, giving the decomposition $X \approx \tilde{X} = \alpha \cdot \tilde{X}_{apt} + \beta \cdot \Delta\tilde{X}$ where $\tilde{X}_{apt} \perp S$ and $\Delta\tilde{X} \not\perp S$. Here, S is the protected attribute to which we are sensitive, and α and β are non-negative values that sum to 1. In other words, we assume that an individual's suitability for the task, at the time of measurement, is a combination of their aptitude ($\tilde{X}_{apt}$), a natural-born ability, and their experiences over time ($\Delta\tilde{X}$)[6]. We assume that the aptitude component, $\tilde{X}_{apt}$, is independent of any protected attribute and therefore complies with the worldview WAE[7]. The component 'life-experience' $\Delta\tilde{X}$ changes aptitude positively or

[6] We make no claims regarding the strength of 'nature' vs. 'nurture'. The framework holds for all potential ratios, including those where $\alpha = 0$ or $\beta = 0$.

[7] We are excluding tasks where success may be strongly correlated with physical attributes — for example, playing professional basketball and height.

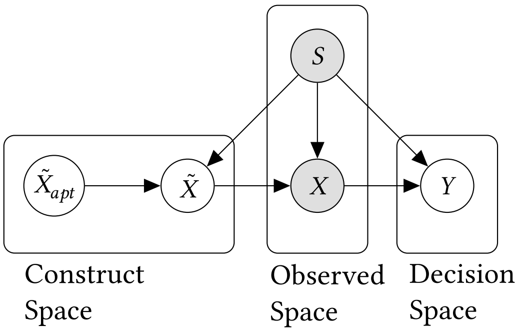


Figure 4.2: The effect of a protected attribute S on descendants of $\tilde{X}_{apt}$ throughout a data-generation procedure. $\tilde{X}$ within the construct space, X within the observed space and Y within the decision space.

negatively and may not be independent of S. A graphical model of our worldview is shown in figure 4.2.

4.4.1.2 *Underlying mechanisms and bias*

We consider a setting in which we observe a statistical disparity between population subgroups separated by the value of $S \in \mathcal{S}$ that occurs in both the observed space and the decision space. The disparity within the decision space may be worse than the disparity within the observed space. One mechanism that can cause this aggravation is label bias, a direct impact of the protected attribute S on the outcome $Y \in \mathcal{Y}$ due to previous discriminatory decisions within the training dataset. To achieve equal treatment, the effects of label bias must be eliminated. The disparity within the observed space can be caused by several mechanisms or their combination: selection bias occurs when the training set contains a non-representative sample of the population; measurement bias occurs when the mapping from the construct space to the observed space is not as faithful for certain groups or individuals[8]. Furthermore, part of the disparity within the observed space can be a true reflection of a disparity within the construct space itself, at the time of measurement. We assume that the distribution of aptitude $\tilde{X}_{apt}$ in the construct space is the same across subgroups, though this need not remain the case with the application of $\Delta\tilde{X}$. Although variation in opportunities between individuals is normal, when the imbalance of opportunity affects a protected group more than the majority, it will result in a disparity between the subgroups within the construct space itself. Addressing this imbalance of opportunity is a principal component of positive action and our framework.

[8] We note that this is not an extensive discussion of bias and there are other underlying mechanisms that can lead to a statistical disparity between an under-represented group and the majority.

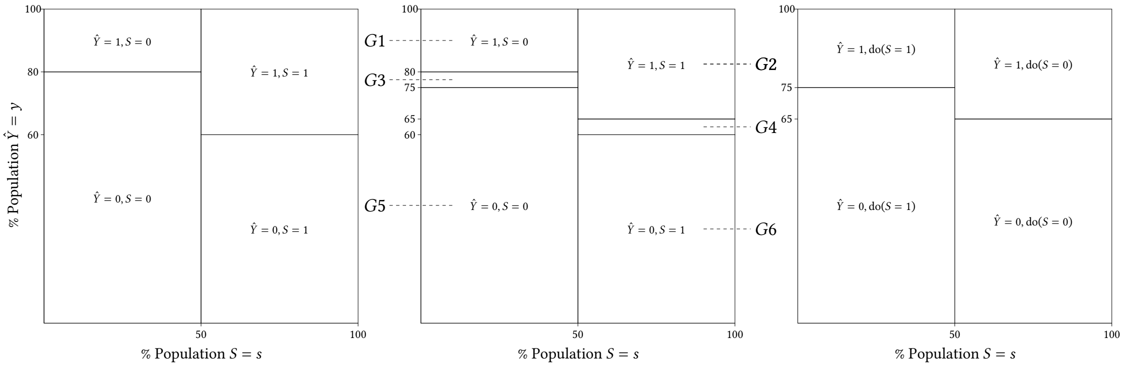


Figure 4.3: The accepted ($\hat{y} = 1$) and rejected ($\hat{y} = 0$) ratios difference between a protected subgroup ($s = 0$) and the majority ($s = 1$). *Left*: under a standard equal treatment selection rule reflecting the original dataset. *Right*: in a counterfactual version of the dataset where $s = s'$. *Middle*: 'Overlapping' the two outcomes. The population captured by groups G_1, G_2, G_5 and G_6 have consistent outcome across both worldviews. Groups G_3 and G_4 represent individuals that will receive different outcomes comparing outcomes in the original dataset to counterfactual versions of the same individuals.

Our work is an application of counterfactual fairness, which can be viewed as an instance of individual fairness. Counterfactual fairness requires that outcomes in the real world align with those in a counterfactual world where the individual had belonged to a different group. Individual fairness requires that similar individuals be treated similarly. To achieve this, however, we use approaches from fair representation literature rather than intervening directly on a causal model. Fair representations are data transformations that preserve the original data, but disentangle the effect of a protected attribute. This is necessary to try and produce a counterfactual model with regard to a specific, predetermined variable. This would normally be restrictive; however, in our case, this is in line with our goals. We do not aim to produce a structural causal model for all variables, we are only concerned with understanding the relationship of one particular variable, the protected characteristic, and its causal effects. We can then marginalise the outcomes with respect to S.

4.4.2 *Positive Action Candidates*

To quantify the effects of structural biases, we divide the data into six subgroups, as shown in figure 4.3, representing areas of agreement and disagreement in the outcome, conditioned on s. This procedure can be done for any pair of model outcomes, although here we demonstrate with example outcomes from an unconstrained classification model. One set of evaluations is on the original data, and the other is the model but applied to a counterfactual dataset. The counterfactual dataset is produced by intervening on the protected attribute S, as previously described, e.g., $P(X|\mathrm{do}(s))$. We conceptually overlay the original outcomes (figure 4.3, left) on the counterfactual outcomes (figure 4.3, right). When overlaid, the data can be separated into six subgroups, as shown in figure 4.3, middle. The subgroups G_1 and G_2 receive a positive outcome in both cases. Subgroups G_5 and G_6, receive a negative outcome in both cases. However, the subgroups G_3 and G_4, represent *a different outcome under counterfactual-based aleatoric uncertainty*. The subgroup G_3 represents the subgroup that would have received a negative outcome based on the observed data, but would have received a positive outcome in a counterfactual world. We propose that these candidates are those that have been impacted negatively by structural biases. This subgroup may be interpreted as individuals who would be rejected under a standard equal treatment selection process because, we hypothesise, of an earlier disadvantage experienced due to their under-represented group membership. Although we cannot accept these applicants while aligning with anti-discrimination legislation, we can highlight them as candidates for positive action — targeted support to help them succeed under a future equal treatment selection process. The subgroup G_4 represents the conceptually opposite subgroup. These are those that would have received a positive outcome based on the observed data but receive a negative outcome in the counterfactual world. We propose that members of this group should remain accepted under a policy of no-detriment. We apply this policy so that candidates are not punished for receiving the benefits of structural biases. Over time, the objective of our positive action approach is to reduce the proportion of the population that fall into the G_3 and G_4 groups.

4.4.2.1 *Candidate Selection*

We still need to identify which applicants we want to highlight for positive action. The reader might now consider a straightforward approach of selecting the top rejected candidates from the under-represented group. The drawback of this approach is that it is only applicable when there is a clear way to rank candidates. We illustrate these two issues with the following motivating example.

Consider a minority who traditionally send their children to schools that teach English to a good level but teach Maths only to a basic level. This minority is under-represented within STEM subjects. To keep this example simple, we consider the application to consist of grades in only two subjects, Maths and English, with equal weight. Blindly taking the best rejected applicants will not spot the applicants who did exceptionally well in Maths, considering the poor education they

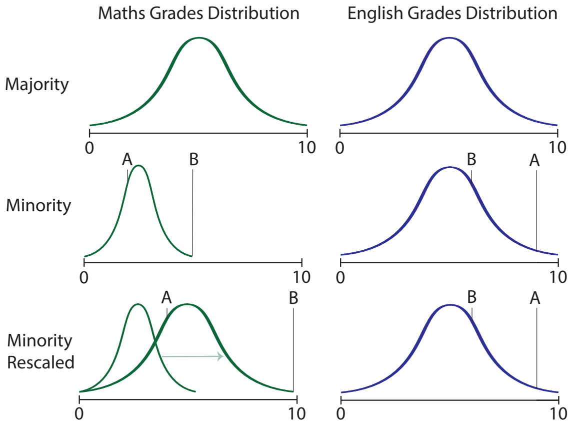


Figure 4.4: How our approach to choosing positive action candidates compares to choosing the top rejected candidates from the under-represented group. With an equal weight selection rule, Applicants A and B have the same overall score. Re-scaling the minority's Maths grade distribution to match the majority's distribution highlights applicant B as the better positive action candidate.

received in this subject. In our approach, the minority's Maths grade distribution gets recalibrated to match the majority's distribution, while the distribution of the English grades is left unaffected because there is no disparity with the majority's distribution. This means that a minority applicant who is good at Maths, relative to their minority subgroup, will be preferred compared to one who is relatively good at English. Figure 4.4 illustrates how two applicants would be ranked under our approach compared to the baseline of choosing the top rejected candidates. For the majority, the distribution ranges between 0–10 for both English and Maths. For the minority, the English distribution ranges between 1–10, but the Maths distribution only ranges between 1–5. The grades of applicant A are 2 and 9 in Maths and English, respectively. Applicant B's grades are 5 and 6 in Maths and English, respectively. With an equal weight selection rule, both have an overall score of 11. When we rescale the Maths grade distributions of the minority to match the majority's distribution, applicant B is highlighted as the better positive action candidate with an overall score of 16 compared to 13 for applicant A. This recalibration is only put into effect when populating the positive action candidates' group. When applicants are considered for acceptance, the features are taken as they are. In the case of this example, we may not be able to accept applicant B, but they are flagged as a positive action candidate — a Maths foundation course, for example, is likely to allow them to successfully compete in a subsequent selection process.

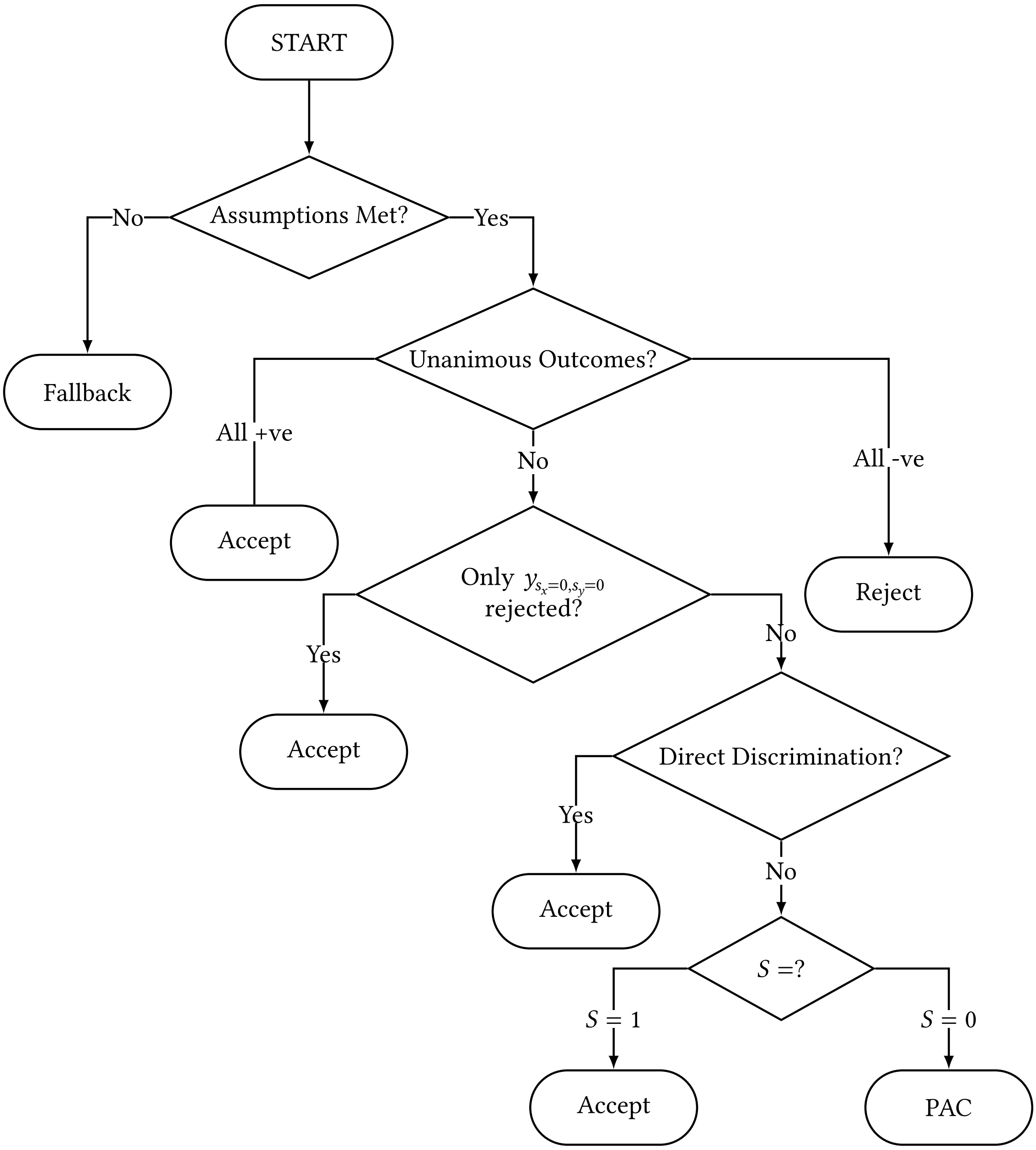


Figure 4.5: A flowchart depiction of the Outcome Comparator, which is also represented in Table 4.1. The input to the flowchart is the ensemble of predicted outcomes, one for each counterfactual scenario, per individual. The result is a recommended outcome: Accept, Reject, or designate as a 'Positive Action Candidate' (PAC). The first decision point is to confirm whether our modelling assumptions, that members of the minority group are disadvantaged, do indeed hold. In the case that they do not, we revert to an unconstrained classifier model. If our modelling assumptions hold, then we follow a process to determine the recommended outcome. Where there is unanimous consensus in the outcomes, the recommendation is followed. Where the decisions are not unanimous, we first determine if the majority of counterfactual outcomes recommend acceptance, if so, then an Accept outcome is recommended. In the case that there is neither a unanimous decision, nor a majority accept outcome, we observe the predictions of the $S_Y = 1$ outcomes – if these all positive, but the $S_Y = 0$ outcomes are not, then we determine that direct bias has been emulated, and the candidate is accepted. If this is not the case, the remaining situation is where Structural Bias has been emulated; the outcome is determined by characteristics of the individual that are related to a protected attribute. In this case, we recommend an Accept outcome for candidates from the majority group, and identify the candidates from the minority group as a PAC.

Table 4.1: Selection rules for mapping the groups represented in figure 4.3 and figure 4.6c to a decision. As $s = 0$ represents a disadvantaged group, we identify those in group 3 as suitable for *positive action*. Combinations not listed are identified, and the outcome reverts to the outcome of an unconstrained model. A flowchart depiction of the decision process captured in this table is given in figure 4.5. An expanded version of this table with all possible outcome combinations is available in Appendix Table 4.10.

Selection Rule	s	$y_{s_x=0,s_y=0}$	$y_{s_x=0,s_y=1}$	$y_{s_x=1,s_y=0}$	$y_{s_x=1,s_y=1}$	Subgroup	y
1	0 or 1	1	1	1	1	G_1 or G_2	1
2	0 or 1	0	1	1	1	G_1 or G_2	1
3	1	0	0	1	1	G_4	1
4	1	0	0	0	1	G_4	1
5	1	0	1	0	1	G_4	1
6	0	0	0	1	1	G_3	2
7	0	0	0	0	1	G_3	2
8	0	0	1	0	1	G_1	1
9	0 or 1	0	0	0	0	G_5 or G_6	0

4.5 Implementation

We describe how to implement our framework as a two-step procedure. The first step generates counterfactual samples, and the second determines if the data allow for equal treatment of candidates.

To identify which candidates may benefit the most from positive action, we use a two-step approach following the scheme in figure 4.6. The aim of this procedure is to analyse, with respect to a protected attribute, two elements. The first element is counterfactual decisions (**Q1**), that is, if the value of S directly observed by a classification system were to be altered, would the outcome change? The second element is outcome with regard to counterfactual samples (**Q2**) — if a counterfactual sample for a given sample were generated, would the decision outcome change? The first accounts for decisions that are potentially discriminatory, as the outcome directly relates to the protected attributes. The latter accounts for differences in the features produced by a change in the protected attributes and their causal consequences. We take this two-step approach as we can then discern which outcomes are affected by direct bias and should be categorised as **outcome label 1**, and those that suffer from structural bias and should be categorised as **outcome label 2** — positive action candidates.

4.5.1 *Generating Counterfactual Samples*

The first step in our two-step procedure is to produce counterfactual samples. For this step, we perform a data generation procedure following a common approach from the literature on fair representation: make a representation of the data $\boldsymbol{z}_x \in \mathcal{Z}_X$, where $\mathcal{Z}_X$ is an intermediate latent space, that is, as best possible, invariant to S. This represents our aptitude construct space $\tilde{X}_{apt}$. The main idea is that, during reconstruction, the protected characteristic is supplied to the decoder in addition to the invariant representation, so that the information required for reconstruction is available. Once trained, we then manipulate the value of the supplied variable S and observe the effect on the reconstruction. First, we train a Generator, an adversarial autoencoder model comprising of an encoder $g\colon (\boldsymbol{x}, s) \mapsto \boldsymbol{z}_x$ and a decoder $k\colon (\boldsymbol{z}_x, s_x) \mapsto \hat{\boldsymbol{x}}_{s_x}$. The encoder maps the observed data point $\boldsymbol{x}$ from the dataset X into a latent representation $\boldsymbol{z}_x$, such that it is independent of the protected attribute, $s \in \mathcal{S}$, where $\mathcal{S}$ is the set of possible values of the protected attribute, for example, $\mathcal{S} = \{0, 1\}$ if the protected attribute is binary. From this latent value $\boldsymbol{z}_x$, our objective is to replicate structural biases associated with each possible s-value when reconstructing the input $\boldsymbol{x}$ as $\hat{\boldsymbol{x}}$ by conditioning the decoder on the s-value in addition to the latent representation. To denote that the value of s can be used to manipulate the reconstruction of $\boldsymbol{x}$, we use s_x. However, in practise, this conditional model proves to be unstable during training. To obtain more consistent results, we follow the approach of Madras et al. (2019) and train s-specific decoders, $k_s\colon \boldsymbol{z}_x \mapsto \hat{\boldsymbol{x}}_s$. In the case of a binary protected attribute, two decoder heads can be used, resulting in two reconstructions created, $\hat{\boldsymbol{x}}_{s_x=0}$ and $\hat{\boldsymbol{x}}_{s_x=1}$. To ensure that the final classification is based on observed samples, we replace the appropriate reconstruction with the original data, so if the actual s-value was 0, the samples returned by this first step would be $\boldsymbol{x}_{s_x=0}$ and $\hat{\boldsymbol{x}}_{s_x=1}$.

4.5.2 *Detecting Discrimination*

In the second of our two steps, we train a classifier model with approximately the same architecture as above. The inputs are the features available to the model, along with an additional input dimension for the protected attribute, which corresponds to the value of s_x. The target labels $y \in Y$ are obtained from the dataset. The model used for the classifier is similar to the generator model in that it comprises an encoder $u\colon (\boldsymbol{x}_{s_x}, s_x) \mapsto \boldsymbol{z}_y$ mapping the input to an invariant latent space $\boldsymbol{z}_y \in \mathcal{Z}_Y$, and a series of s-specific classifier-heads. Then, each classifier-head $v_s\colon\ \boldsymbol{z}_y \mapsto y_{s_y}$ is trained to produce a classification for a given s group.

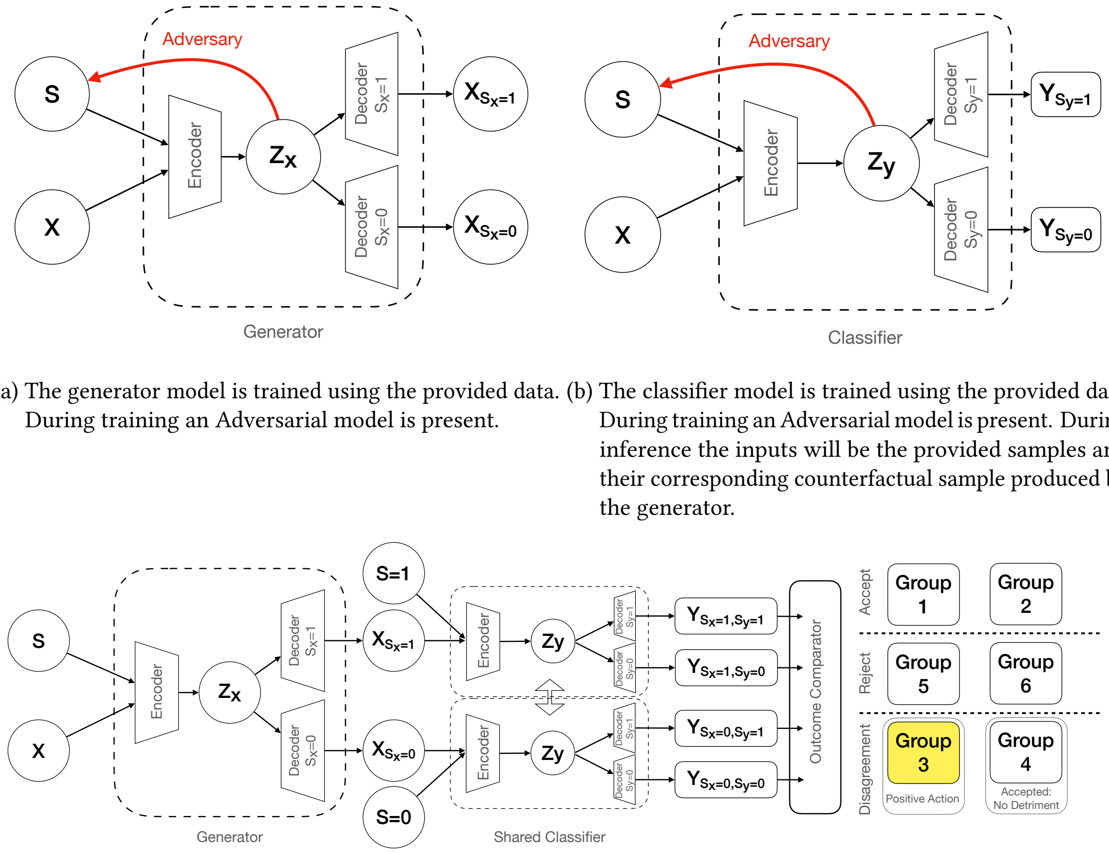
For a given sample, $\boldsymbol{x}$, the output of the classifier model is a set of outcomes: $y_{s_x=0,s_y=0}$, $y_{s_x=0,s_y=1}$, $y_{s_x=1,s_y=0}$, and $y_{s_x=1,s_x=1}$. These outcomes capture counterfactual-based aleatoric uncertainty between outcomes differing by s_x and direct discrimination with outcomes differing by s_y. We use a set of selection rules, referred to as the Outcome Comparator in figure 4.6c to sort the set of original samples X into one of six subgroups G_{1-6}. The full selection rules are presented in table 4.1, but we give some intuition: Groups 1 & 2 ($G_{1,2}$) consist of candidates whose outcomes were either unanimously accepted across all counterfactual inputs (selection rule 1), or differed due to S_y, the concatenated *perceived* protected attribute, changing (selection rules 2 & 8). Unanimous *negative* outcomes for all counterfactual inputs are assigned to groups $G_{5,6}$ (selection rule 9). Lastly, applicants who receive a disagreement amongst the outcomes, i.e., their outcome depends on the value of S_x, are assigned to groups $G_{3,4}$ (selection rules 3-7). The members of group G_4 are accepted as they would by an unconstrained classifier. This is because the selection rules we have adopted follow a no-detriment policy — we only aim to identify candidates that have the potential to succeed, as opposed to punishing candidates for benefiting from a system that they don't control. The members of group G_3 are highlighted as *positive action candidates.*

In the case that the selection rules do not capture the pattern of predicted outcomes, a *fallback* option exists. In this case, the outcome corresponding to the 'true' prediction is used. As an example, if $s = 0$, and none of the selection rules apply, the outcome $Y_{s_x=0,s_y=0}$ is used.

4.5.3 *Model*

Our model is implemented as two successive neural networks (a generator and classifier) representing distinct phases as mentioned above[9]. The goal of each model is to produce counterfactual representations with respect to the protected attribute S. First, we aim to train an autoencoder-based generator model capable of producing a counterfactual *reconstruction* in $\mathcal{X}$; and then we aim to train an autoencoder-based classifier model capable of producing a counterfactual *decision* in $\mathcal{Y}$.

[9] Our code is available at `https://github.com/wearepal/positive-action-framework`



(a) The generator model is trained using the provided data. During training an Adversarial model is present.

(b) The classifier model is trained using the provided data. During training an Adversarial model is present. During inference the inputs will be the provided samples and their corresponding counterfactual sample produced by the generator.

(c) During inference a given sample produces two counterfactual versions, corresponding to all values of *S*. At this point the counterfactual sample that corresponds to the *s*-value observed in the data is discarded and replaced with the original sample. Both the original sample and the corresponding counterfactual are processed by the classifier model, producing 4 outputs. The selection rules in table 4.1 are applied by the 'Outcome Comparator' resulting in the original sample being allocated to one of 6 groups. Candidates that are determined to belong to group 3 are selected for Positive Action.

Figure 4.6: Diagram illustrating our method. *Top*: The individual components are shown during training. The Generator and Classifier models consist of a similar architecture. During this period, an adversary is present in both models. *Bottom*: The composite Positive Action Framework during inference on new data. The four corresponding predicted outcomes then determine the group classification according to one of three final outcomes: *accept*, *reject*, or *disagreement* which has two outcomes associated. Candidates from under-represented groups that were rejected, but would have received a positive outcome in a counterfactual world are *flagged for positive action*. Candidates from majority s that were flagged for acceptance, but would *not* have received a positive outcome in a counterfactual world remain accepted under a 'no-detriment' policy.

The underlying autoencoder model for both generator and classifier has a similar architecture to Madras et al. (2018a), but with multiple decoders, similar to Shalit et al. (2017), Madras et al. (2019) and Park et al. (2021), and comprises:

1. Encoder functions $g_x\colon (\mathcal{X}, \mathcal{S}) \mapsto \mathcal{Z}_X$ and $g_y\colon (\mathcal{X}, \mathcal{S}) \mapsto \mathcal{Z}_Y$ map the input $\boldsymbol{x}$ to a more malleable representation $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$ respectively.

2. Adversary functions $h_x\colon \mathcal{Z}_x \to \mathcal{S}$ and $h_y\colon \mathcal{Z}_y \to \mathcal{S}$ to encourage the representations in the latent space to *not* be predictive of *s*.

3. An ensemble of $\mathcal{S}$-specific decoders. The task in the generator model is to produce a reconstruction $\hat{\boldsymbol{x}}_s$ from $\boldsymbol{z}_x$ and is defined as a function $k : \mathcal{Z}_x \mapsto \mathcal{X}_s \ \forall s \in \mathcal{S}$. Where $\mathcal{X}_s$ is an array of reconstructions, each corresponding to a possible s-value. The classifier task is to produce a prediction score $\hat{y}_s$ from $\boldsymbol{z}_y$ and is defined as a function $m : \mathcal{Z}_y \mapsto \mathcal{Y} \ \forall s \in \mathcal{S}$. Similarly to the generator, $\mathcal{Y}_s$ is an array of prediction score, each corresponding to a possible s-value. During training, $\mathcal{X}_s$ and $\mathcal{Y}_s$ are indexed by the real s value so that only the $\mathcal{S}$-head that corresponds to the true protected attribute is used for training.

The purpose of the generator is to produce a *likely* counterfactual $\hat{X}$ with respect to S. To do this, we produce a latent embedding Z_x, which removes as much information about S as possible. Then we have one decoder-head per possible S-label, allowing the effect of s to be reintroduced[10]. We train this model by optimising the objective function in equation (4.2), where ℓ_{recon} is an appropriate loss between the reconstructions and the features. A hyper-parameter λ is incorporated to allow for a trade-off between the two competing losses[11]. In our experiments, we use the mean of a combined reconstruction loss, with cross-entropy used per categorical feature group and L1-loss used for the remaining non-categorical features. The adversarial loss ℓ_{adv} is realised as cross-entropy between the predicted and the target S coupled with a supplementary non-parametric measure, Maximum Mean Discrepancy (Gretton et al., 2007), with a linear kernel between the embeddings per group (i.e. $\text{MMD}(Z_{s=0}, Z_{s=1})$) giving equation (4.1). We add the additional MMD term, as the core of our method relies on the independence between both Z_x and Z_y, and S. Furthermore, adversarial training can be notoriously unstable, and we found that additional use of MMD stabilises adversary performance.

$$\ell_{\text{adv}}(Z, \hat{S}, S) = \ell_{\text{BCE}}(\hat{S}, S) + MMD(Z_{s=0}, Z_{s=1}) \tag{4.1}$$

$$\mathcal{L}_{\text{AE}} = \min_{\theta,\pi} \max_{\phi} \mathbb{E}_{x,s\sim D}[\ell_{\text{recon}}(k^{\pi}_{S=s}(g^{\theta}_{x}(\boldsymbol{x}, s); \boldsymbol{x}) - \lambda_1 \ell_{\text{adv}}(g^{\theta}_{x}(\boldsymbol{x}, s), h^{\phi}(g^{\theta}_{x}(\boldsymbol{x}, s)), s)] \tag{4.2}$$

The classification model is identical in architecture to the generator, and consists of a shared network with, in a similar fashion to the autoencoder, S-specific classifier-heads. This is done to capture any potential direct discrimination that the model determines to exist based on past data.

[10] This could be performed with a conditional decoder that additionally accepts the protected attribute as input, but in practise, we found our approach to work more consistently.

[11] In our experiments, we use $\lambda_{\{1,2\}} = 1.0$

For the classification model, the task is to produce an ensemble of predictions of the class label y_{s_y} from $\boldsymbol{x}$ and is defined as $g_y : (\mathcal{X}, \mathcal{S}) \mapsto \mathcal{Z}_Y$, $h_y : \mathcal{Z}_y \to \mathcal{S}$, and $m_s : (\mathcal{X}) \to \mathcal{Y}_s \;\forall\, s \in \mathcal{S}$. As with the generator, only the $\mathcal{S}$-head that corresponds to the true protected label is used for training. The objective is shown in the following equation:

$$\mathcal{L}_{\text{Clf}} = \min_{\omega,\xi} \max_{\psi} \mathbb{E}_{x,s\sim D}[\ell_{\text{recon}}(m^{\omega}_{S=s}(g^{\xi}_{y}(\boldsymbol{x}, s)); y) - \lambda_2 \ell_{\text{adv}}(g^{\xi}_{y}(\boldsymbol{x}, s), h^{\psi}(g^{\xi}_{y}(\boldsymbol{x}, s)), s)] \tag{4.3}$$

At inference time, the generator model produces one reconstruction per S-label, per sample, and likewise for the classification model, one outcome per S-label, per reconstruction is produced. In the case of a binary S label, this produces two reconstructions per sample and two decisions per reconstruction, resulting in 4 outcomes per sample.

4.6 Experiments

To determine the effect of our approach in selecting positive action candidates, we evaluate our counterfactual-based approach against a number of baselines. First, we use an unconstrained Logistic Regression (`LR`) model. Then we compare with the following established fairness methods: Instance Reweighting (`RW`) (Kamiran and Calders, 2012); Prejudice Regularisation (`Reg`) (Kamishima et al., 2012); FairLearn (`FL`) (Agarwal et al., 2018); and Disparate Impact Removal (`DIR`) (Zafar et al., 2017b). Lastly, as a baseline, we include a Demographic Parity Oracle (`Oracle`) that 'flips' as few outcomes as possible to enforce *exact* Demographic Parity with maximum possible accuracy. We use this to demonstrate the optimal fair outcome possible for all datasets.

Next, we compare against a naïve implementation of our Positive Action Framework. In this strategy, we train two models: The first is an unconstrained equal treatment model (`LR`); the second is a Demographic Parity enforcing model. As such, we refer to this approach as `LR-DP`. We then apply a simple selection rule: candidates from the under-represented group that were rejected by the unconstrained model but accepted by the demographic parity model are identified as positive action candidates. As there are a number of methods to promote Demographic Parity, we evaluate using a selection of fairness enhancing methods as the second model — these are the `FL`, `RW`, `Reg` and `DIR` methods from above. Furthermore, as a further study, we repeat the procedure, but use the Equal Opportunity enforcing model of Hardt et al. (2016) instead of the unconstrained model. We refer to this approach as `EQ-DP`. The reason for this is to do with our motivation: we want to promote demographic parity in the long-term, but cannot enforce this immediately due to a potential violation of anti-discriminatory legislation. However, Equality of Opportunity does not violate this concern. We use Equality of Opportunity in the first model because it is a valid fairness-promoting method, but we also produce Demographic Parity results in the second model, because that is our long-term goal. Finally, we evaluate our proposed Positive Action Framework, which we refer to as `PAF`. As an ablation study of our approach we also evaluate two variations of the generator model: `PAF-NN` in which the adversarial method to obtain a counterfactual model is replaced by a nearest neighbour model – during inference, the nearest sample based on cosine-similarity from the opposing S-group in the training data is selected as the counterfactual sample; and `PAF-CG` in which the adversarial model is replaced with the established cross-domain translation model CycleGAN (Zhu et al., 2017).

4.6.1 *Datasets*

We first use synthetic data to demonstrate how our approach can be applied to a candidate filtering task within a biased setting. We consider applicants to a university course in a fictitious world inhabited by *blue* and *green* people. We take the *colour* of a person as the protected attribute $s \in \{\text{Green}, \text{Blue}\}$. This university course is for a traditionally *blue* profession, making the setting potentially biased in terms of both direct and structural biases. Because the Blue group is advantaged, we assign this group $s = 1$, and the Green group $s = 0$. The advantage of using a synthetic dataset is that we have access to the data generation procedure, so we can faithfully determine if positive action candidates have indeed been identified. We then demonstrate our approach on the following real-world datasets: *UCI Adult Income*, *Brazilian Admissions*, *UCI Communities and Crime*, and *NYPD Stop Question Frisk*. We use these to highlight the practical benefits and some potential challenges of deploying our Positive Action Framework in a real-world setting.

4.6.1.1 *Synthetic Data Generation*

We define a data generation procedure for a dataset containing opportunity, measurement, and label biases as shown in figure 4.7 with binary S labels, with 2 imperfect observers of 3 features, making a feature-space $\mathcal{X}$ comprising 6 features. We then generate two outcome scores: 1) An

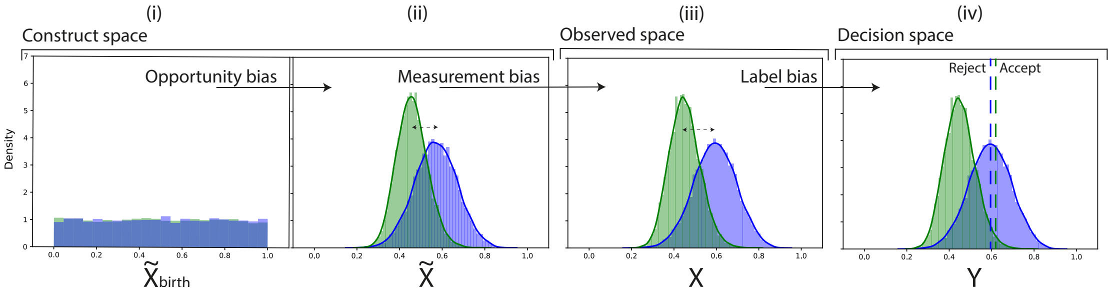


Figure 4.7: Changes in the engineered synthetic data. Starting from a uniform distribution, we visualise how the additive effect of bias can result in a significant disproportion of success between groups differing by a protected attribute. The opportunity bias and measurement bias are modelled as a shift between the distributions. The label bias is modelled by having different acceptance thresholds for the different groups (vertical dashed lines in the right figure).

'acceptance score' based on a linear combination of the observed features with a label bias introduced by setting different acceptance thresholds depending on the value of S (figure 4.7(iv)). 2) A 'graduation grade' based on a linear combination of the features in $\tilde{X}$, bypassing the effect of the introduced measurement bias and label bias.

Full details of the data generation process are available in section 4.9.1.

4.6.1.2 *UCI Adult Income Data*

The UCI Adult Income Dataset (Dheeru and Karra Taniskidou, 2017) is often used to assess algorithmic fairness systems. This dataset comprises $45,222$ samples from the 1994 U.S. census with 14 features including occupation, maximum attained education level, nationality, and relationship status. Of these 14 features, we reserve the binary `salary` feature as the target label, with `>$50K` being the positive outcome and the binary feature `gender` (Male/Female) as the protected attribute with Male as the advantaged group ($s = 1$). Of the remaining 12 features, we reduce the categorical feature `Nationality` to a binary case of American/not American. In addition, there are 6 further features that consist of categorical features. We one hot encode these, resulting in samples with 62 features in total.

4.6.1.3 *Brazilian University Admissions Data*

The UFRGS Entrance Exam and GPA Data, known as the Brazilian University Admissions Dataset (Castro da Silva, 2019) consists of 9 entrance exam scores for students applying to the Federal University of Rio Grande do Sul in Brazil. In addition to these exam scores is the binary label `Gender` (Male/Female) with Male as the advantaged group ($s = 1$), along with the mean of the students GPAs during the first three semesters. This dataset comprises $43,303$ samples with 11 features. Of these 11 features, we reserve the GPA feature as the target label, which we binarize with >3.0 on a 4.0 scale being the positive outcome and the gender feature as the protected attribute resulting in 9 input features total.

4.6.1.4 *UCI Communities and Crime Data*

The UCI Communities and Crime dataset (Redmond and Baveja, 2002) is a composite dataset based on the 1990 U.S. Census, 1995 U.S. F.B.I. Uniform Crime Report and the 1990 U.S. Law Enforcement Management and Administrative Statistics Survey, which is hosted on the UCI Machine Learning Repository. The dataset is created following the procedure of Kamiran and Calders (2012) and consists of 98 features. They dictate that the feature PctBlack ($s = 1$ if $> 6\%$) is used as a protected attribute and that the target label is HighCrimeRate. Of the remaining 96 features, only the feature State is a categorical feature with 46 unique values. We one hot encode this resulting in 136 total input features for $1,993$ samples.

4.6.1.5 *NYPD Stop, Question and Frisk Data*

The New York Police Department release data about their Stop, Question and Frisk programme. The data used is from 2016 and comprises $12,347$ samples with 67 features. We reserve the binary feature sex (Male/not Male) as the protected attribute with Male as the advantaged group ($s = 1$) and the feature hasWeapon as the binary target. Of the remaining 65 features, 59 contain categorical values and become one hot encoded. 6 features are non-categorical. This results in each sample comprising 145 input features.

4.6.2 *Evaluation*

When evaluating our approach[12], we want to determine if we can solve our two questions. **Q1**: would the outcome change if *only* the protected attribute was different (direct discrimination)? **Q2**: would the outcome change if the protected attributed *and its causal consequences* were different (structural discrimination)? However, because we are working with counterfactual outcomes we have to consider the well-known problem that the ground-truth remains unknown (Butcher et al., 2021), which results in a third question: **Q3**: how confident are we in the model's outcomes?

[12] All experiments are repeated with 10 random initialisations of model weights, and dataset splits.

Lastly, we would like to know **Q4**: does our Positive Action Framework remain accurate? And, **Q5**: does it identify positive action candidates in a manner than promotes Demographic Parity?

To determine the answers to the first two questions, we use the group allocations described in section 4.5.2. Simply, for each dataset, we monitor the array of outcomes and report the percentage of outcomes that appear to demonstrate direct, and structural discrimination. Where the outcomes differ by S_y, e.g. $\hat{Y}_{S_x=s,S_y=0} \neq \hat{Y}_{S_x=s,S_y=1}$, we characterise this behaviour as direct discrimination — the outcomes of the classification model are not consistent for the original sample when the value of s perceived by the classifier model changes. Where the outcomes differ by S_x, e.g. $\hat{Y}_{S_x=0,S_y=s} \neq \hat{Y}_{S_x=1,S_y=s}$, we characterise this behaviour as structural discrimination — the outcomes of the classification model change between the original and counterfactual samples. The results of this are shown in table 4.3. Although the true levels of discrimination that we want to observe are not available for all datasets, for the synthetic data, we can produce exact outcomes. We record these results as `CF`.

To determine confidence in the model, we propose two techniques. The first technique is to train an auxiliary `LR` model during evaluation to predict S from each of the inputs and the two embeddings, Z_x and Z_y. Results for this are shown in table 4.2. If the accuracy of this model is equal to the probability of the majority group in the dataset, then we can be confident that the input is somewhat invariant to S. The second technique is to analyse the proportion of groups that are allocated using synthetic data. We show this in figure 4.8.

To determine whether our proposed model retains accuracy and promotes Demographic Parity through positive action candidates, we define the following metrics which are applicable to all datasets: *Positive Predictive Rate* (`Acceptance`) and *Accuracy*. In addition, we also define metrics that are only suitable for the synthetic dataset: *True Capture Rate* (`TCR`), and *False Identification Difference* (`FID`). In all these metrics, we report the results when the predicted outcome is positive ($\hat{Y} = \{1\}$) and when we also treat the positive action outcome $\hat{Y} = 2$ as positive ($\hat{Y} = \{1, 2\}$). The metrics applied to all models are:

Positive Predictive Rate (PPR). This corresponds to the probability of a positive outcome. When this is equalised across groups, demographic parity is satisfied. We expect a disparity when the positive action outcome ($\hat{Y} = 2$) isn't considered a positive outcome, and for this disparity to be greatly reduced when it is.

$$\mathrm{PPR}(y, s) = P(\hat{Y} = y | S = s)$$

Accuracy: We evaluate the utility of the model for predicting the target outcome. As we have an additional outcome of $\hat{Y} = 2$ to depict positive action candidates, we parameterise the definition of True Positive (TP) to accept the values that are considered positive. The count of True Negative (TN) samples remains standard.

$$\text{Acc}(y) = \frac{\text{TP}(y) + \text{TN}}{N}$$

In addition, for the synthetic data, we created an unbiased outcome (G), based on the outcome that would have occurred if structural and direct biases were not included in the data generation procedure. We evaluate how well the models are able to identify these outcomes using the following metrics which we explain using the language of the synthetic data (graduation G, acceptance Y):

True Capture Rate (TCR). This measures the sensitivity of the model with respect to acceptance $\hat{Y} = [\{1\}, \{1, 2\}]$ conditioned on the ability to graduate ($G = 1$):

$$\text{TCR}(y, s) = P(\hat{Y} = y | S = s, G = 1)$$

False Identification Difference (FID) measures the difference in negative graduation outcomes conditioned on acceptance. In other words, once a candidate is accepted, does their chance of graduating depend on the protected attribute?

$$\text{FID}(y) = |P(G = 0 | S = 1, \hat{Y} = y) - P(G = 0 | S = 0, \hat{Y} = y)| \tag{4.4}$$

4.7 Results & Discussion

In our experiments, we aim to answer five questions, the first two of which (**Q1** and **Q2**), are addressed in table 4.3. We can see that, in comparison to the rate of direct discrimination in the data, depicted by the model `CF`, this behaviour is under-detected by our `PAF` model, and its variations. Assuming that our model continues to under-detect this type of discrimination, we can see that drawing conclusions about the level of direct discrimination in the Adult and SQF datasets is not possible as these display relatively low levels. However, the Admissions, and particularly Crime datasets, where the procedure for mapping from input features to observed outcome is more complex, appear to display a greater level of this type of discrimination. Analysing the rate of structural discrimination, we can see that our `PAF` model approximates the `CF` rate quite well, while both ablation models fail. On the assumption that our model can detect the presence (or not) of this type discrimination, the Adult and Crime datasets which contain features about individuals and neighbourhoods, respectively, appear to have high-levels of structural discrimination, whereas the SQF dataset in particular, which mostly contains features describing the frisking procedure, has a much lower rate. We propose that this type of analysis is an important facilitator of discussion to determine potential challenges and mitigation strategies for different types of tasks.

With **Q3**, we aim to determine confidence in the model's outputs. In figure 4.8 we can view the allocation of groups based on the array of results produced for the dataset. In comparison again to the 'ground-truth' counterfactual data, our `PAF` model allocates to groups at approximately the same rate. Encouragingly, the groups G_3 and G_4, which are associated with outcomes that would change based on structural biases seem to be well-captured. Unfortunately, the nearest neighbour, and inter-domain translation model, CycleGAN fail in these groups, over-allocating to G_2 and G_5, groups where there is a unanimous decision. In addition, to approach this question, we report figures in table 4.2 about whether invariance has been achieved across datasets. We propose that we can be more confident in the outcome in datasets where invariance is achieved. In the case where invariance isn't achieved, it doesn't necessarily invalidate the claims — a test of predictive ability doesn't mean that the information is used during the task, but they should be treated more cautiously. For the synthetic dataset, we can see that the generator embedding Z_x which is associated with our ability to detect structural biases, contains little information about the protected attribute. The classifier embedding, associated with direct discrimination, however, still has a lot of information removed, but to a lesser extent. Applying this to the real-world datasets, the results for the Admissions and SQF datasets show that a good level of information was removed about S for both embeddings. The Adult dataset shows that we should be more confident in the detection of direct discrimination than in our ability to detect structural discrimination for this dataset. Lastly, we should be the least confident in drawing conclusions for the Crime dataset.

Lastly, we look at **Q4** and **Q5** — does our suggested approach retain accuracy *and* promote fairer outcomes? Throughout all datasets, we want to see that when the positive action outcomes are included, the difference in PPR is reduced. In addition, accuracy of our should remain high, although we fully expect this to reduce if we include positive action candidates as receiving a positive outcome, as these are candidates that would historically receive a negative outcome. First, the results of the synthetic dataset are shown in table 4.4. We can see that the dataset is challenging.

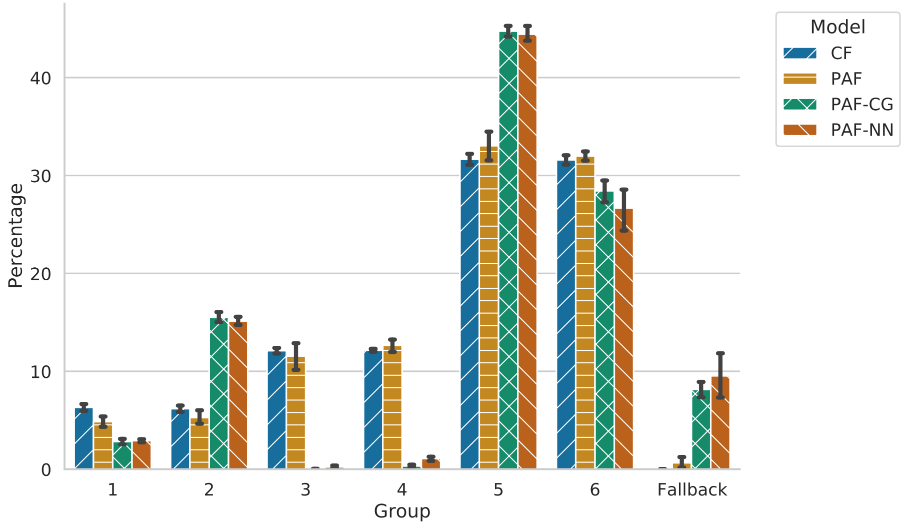


Figure 4.8: Percentage of outcomes assigned to each group described in table 4.1 for the Synthetic dataset. The aim is to match the 'true' counterfactual outcomes (CF) as well as possible. Our proposed fair-representation based PAF model outperforms variations of the model based on CycleGan (PAF-CG) and a Nearest Neighbour (PAF-NN) approach, both of which fail to identify groups G_3 and G_4, the groups of interest in this work.

Of the baseline models, only the Oracle and DIR models manage to encourage demographic parity, although this is at the expense of accuracy. The two general naïve approaches show opposite results. The LR-DP models largely place too much emphasis on accuracy — the demographic parity approaches rely on a mechanism to reduce the acceptance rate for the privileged group, which this approach does not facilitate, so only a small number of candidates are selected for positive action. The EQ-DP models have the alternative issue that the results are somewhat *fairer*, but accuracy suffers considerably. This pattern is borne out in the real-world datasets shown in tables 4.6 to 4.9.

For the synthetic dataset we also evaluate how well the model determines candidates that would have gone on to be successful if discriminatory behaviour had not been present during the data creation. To do this, we look at the rate at which candidates are selected, given that they had the potential to succeed which is measured by TCR in table 4.5. In comparison to every other approach, our PAF model is the *only* approach that captures these candidates. Similarly, we also measure the difference in false identification, that is, what is the difference in rates between subgroups of candidates that are accepted, but unlikely to succeed? Results are shown in the table 4.5. Again, our positive action approach is the only method that actively reduces this measure of disparity.

Table 4.2: The accuracy of a Logistic Regression model at predicting the protected attribute, S from various layers of the trained PAF model alongside the probability that the protected attribute is the favoured group. When the accuracy is close to the probability, the model is unable to correctly classify the input.

Metric	Synthetic	Admissions	Adult	Crime	SQF
Accuracy $S\|X$	0.828 ± 0.006	0.685 ± 0.005	0.847 ± 0.002	0.86 ± 0.016	0.929 ± 0.006
Accuracy $S\|Z_x$	0.564 ± 0.073	0.546 ± 0.034	0.749 ± 0.039	0.998 ± 0.003	0.925 ± 0.006
Accuracy $S\|Z_y$	0.63 ± 0.017	0.539 ± 0.009	0.673 ± 0.005	0.67 ± 0.032	0.925 ± 0.005
P(S=1)	0.503 ± 0.007	0.481 ± 0.004	0.675 ± 0.006	0.492 ± 0.023	0.925 ± 0.005

Table 4.3: Percentage of samples with outcome arrays associated with *direct discrimination* and *structural discrimination* for a number of datasets. On the case of the synthetic data, we have access to *true* counterfactual outcomes, which are reported as the model CF.

				Dataset		
Bias	Model	Synthetic	Admissions	Adult	Crime	SQF
Direct	CF	10.137 ± 0.425	—	—	—	—
	PAF	4.133 ± 1.098	6.068 ± 2.6	3.726 ± 0.609	7.035 ± 2.694	2.329 ± 0.289
	PAF-NN	3.92 ± 0.887	5.768 ± 3.697	3.873 ± 0.784	9.347 ± 5.758	2.45 ± 0.398
	PAF-CG	4.12 ± 1.091	5.301 ± 1.68	4.157 ± 0.704	6.884 ± 1.606	2.438 ± 0.4
Structural	CF	20.95 ± 0.568	—	—	—	—
	PAF	23.663 ± 3.269	15.214 ± 4.044	24.844 ± 3.149	30.427 ± 4.854	7.663 ± 0.354
	PAF-CG	4.963 ± 1.335	7.497 ± 3.65	10.646 ± 3.119	22.085 ± 5.645	7.882 ± 2.501
	PAF-NN	4.963 ± 1.335	10.485 ± 1.165	14.205 ± 0.725	18.97 ± 1.359	13.151 ± 1.792

4.7.1 *Limitations and Intended Use*

When we are considering an algorithmic decision and support system deployed in a real-world setting, we can distinguish between different mechanisms that may lead to a disparity in positive outcome rates between population subgroups: bias we can successfully intervene on, by mitigating, or even completely removing, label bias from the training data and the learnt model; and, a disparity we can detect, but cannot directly intervene on without employing positive discrimination, which is opposed to anti-discrimination legislation.

In this work, we assume that we are required to enforce the mapping between the observed and the decision space to be independent of the protected attribute, that is, we assume that it is a requirement to mitigate direct discrimination (selection rules 2 & 8, Table 4.1). This is the only bias that is mitigated at the accept / reject level. Inclusion of the positive action candidate outcome and the G_3 subgroup enables us to audit and mitigate, in the form of recommending candidates for positive action, any additional effects that may cause structural disparities, e.g., selection bias and imbalance of opportunities.

Table 4.4: Synthetic Dataset results. Positive Predictive Rate should be equalised between outcomes conditioned on S. In our positive action approach we expect that there will be disparity in these values when model only traditional positive outcomes ($\hat{Y} = 1$) are considered, but that this disparity should be reduced when positive action outcomes ($\hat{Y} = 2$) are also included. For the accuracy of the model, we aim for this value to be high when $\hat{Y} = 1$ and expect a reduction when $Y = 2$ is evaluated as a positive outcome.

		Positive Predictive Rate			Accuracy	
	Model	PPR($Y = \{1\}, S = 0$)	PPR($Y = \{1, 2\}, S = 0$)	PPR($Y = \{1\}, S = 1$)	Acc($Y = \{1\}$)	Acc($Y = \{1, 2\}$)
Baseline	`Oracle`	22.697 ± 13.989	-	22.698 ± 13.986	86.138 ± 0.47	-
	`FL`	8.728 ± 0.525	-	27.844 ± 0.916	93.903 ± 0.414	-
	`Reg`	3.187 ± 0.342	-	31.658 ± 0.775	97.225 ± 0.302	-
	`LR`	6.041 ± 0.516	-	34.091 ± 0.815	99.6 ± 0.181	-
	`RW`	6.512 ± 0.455	-	31.605 ± 0.724	98.065 ± 0.153	-
	`DIR`	11.513 ± 0.559	-	11.313 ± 0.506	83.177 ± 0.406	-
Ours	`PAF`	10.022 ± 1.61	33.262 ± 4.416	35.784 ± 1.133	96.987 ± 0.869	85.43 ± 2.107
Ours (Naïve)	`(EQ-DP) (FL)`	6.046 ± 0.509	9.78 ± 0.629	5.369 ± 0.256	85.403 ± 0.504	83.555 ± 0.562
	`(EQ-DP) (RW)`	6.046 ± 0.509	6.653 ± 0.427	5.369 ± 0.256	85.402 ± 0.504	85.125 ± 0.455
	`(EQ-DP) (Reg)`	6.046 ± 0.509	6.046 ± 0.509	5.369 ± 0.256	85.402 ± 0.504	85.402 ± 0.504
	`(EQ-DP) (DIR)`	6.046 ± 0.509	13.113 ± 0.604	5.369 ± 0.256	85.402 ± 0.504	81.91 ± 0.38
	`(LR-DP) (FL)`	6.041 ± 0.516	9.775 ± 0.631	34.091 ± 0.815	99.6 ± 0.181	97.752 ± 0.185
	`(LR-DP) (RW)`	6.041 ± 0.516	6.648 ± 0.431	34.091 ± 0.815	99.6 ± 0.181	99.323 ± 0.194
	`(LR-DP) (Reg)`	6.041 ± 0.516	6.041 ± 0.516	34.091 ± 0.815	99.6 ± 0.181	99.6 ± 0.181
	`(LR-DP) (DIR)`	6.041 ± 0.516	13.108 ± 0.613	34.091 ± 0.815	99.6 ± 0.181	96.108 ± 0.332
Ours (Ablation)	`PAF-CG`	7.782 ± 0.675	7.822 ± 0.656	35.88 ± 1.059	98.115 ± 0.418	98.095 ± 0.411
	`PAF-NN`	7.719 ± 0.875	8.179 ± 0.737	35.789 ± 1.147	98.132 ± 0.628	97.903 ± 0.605

We choose to adopt a no-detriment, or positive-corrective approach. This means that no individual, even if they allegedly benefit from past biased decisions, will be made worse off by the positive action approach. In practice, selection rules can be adapted to suit the context and objectives at hand.

4.8 Conclusion

We present a novel algorithmic fairness framework that builds on the notion of positive action to advance equal representation while respecting anti-discrimination legislation and the right to equal treatment. We aim to identify high-potential under-represented applicants, even if they cannot yet successfully compete in an equal treatment selection process against applicants from the majority group. As we are unable to accept them directly, they are highlighted as promising candidates for positive action measures.

Positive action initiatives can already be found in practice and can include outreach activities, targeted training and adaptive policies. Specific positive action measures will be case and context dependent and should be determined by domain experts. Our aim is to demonstrate that machine

Table 4.5: Synthetic Dataset results for metrics associated with an unobserved, unbiased outcome. True Capture Rate measures the rate at which candidates with the potential to succeed based on $\tilde{X}$ are correctly identified for acceptance. False Identification Difference measures the disparity in likely failure rates based on $\tilde{X}$. The PAF model is the only one that is successful in these measures.

		True Capture Rate			False Identification Difference	
	Model	TCR($\hat{Y} = \{1\}, S = 0$)	TCR($\hat{Y} = \{1, 2\}, S = 0$)	TCR($\hat{Y} = \{1\}, S = 1$)	FID($Y = \{1\}$)	FID($Y = \{1, 2\}$)
Baseline	Oracle	76.578 ± 6.199	-	64.442 ± 40.108	20.255 ± 23.068	-
	FL	59.465 ± 2.316	-	79.897 ± 2.929	21.531 ± 2.161	-
	Reg	46.572 ± 3.992	-	94.618 ± 1.951	61.88 ± 3.567	-
	LR	71.364 ± 2.706	-	96.714 ± 1.771	48.065 ± 3.447	-
	RW	70.469 ± 3.04	-	95.491 ± 1.718	41.903 ± 3.75	-
	DIR	51.971 ± 3.375	-	51.617 ± 4.777	2.692 ± 2.401	-
Ours	PAF	82.778 ± 6.222	93.479 ± 6.374	96.935 ± 1.75	30.42 ± 3.443	2.431 ± 2.133
Ours (Naïve)	(EQ-DP) (FL)	71.364 ± 2.706	71.364 ± 2.706	16.72 ± 2.83	46.687 ± 4.972	22.942 ± 4.152
	(EQ-DP) (RW)	71.364 ± 2.706	72.216 ± 2.619	16.72 ± 2.83	46.687 ± 4.972	41.677 ± 5.506
	(EQ-DP) (Reg)	71.364 ± 2.706	71.364 ± 2.706	16.72 ± 2.83	46.687 ± 4.972	46.687 ± 4.972
	(EQ-DP) (DIR)	71.364 ± 2.706	73.582 ± 3.124	16.72 ± 2.83	46.687 ± 4.972	14.178 ± 4.207
	(LR-DP) (FL)	71.364 ± 2.706	71.364 ± 2.706	96.714 ± 1.771	48.065 ± 3.447	24.279 ± 2.149
	(LR-DP) (RW)	71.364 ± 2.706	72.216 ± 2.619	96.714 ± 1.771	48.065 ± 3.447	43.039 ± 3.889
	(LR-DP) (Reg)	71.364 ± 2.706	71.364 ± 2.706	96.714 ± 1.771	48.065 ± 3.447	48.065 ± 3.447
	(LR-DP) (DIR)	71.364 ± 2.706	73.582 ± 3.124	96.714 ± 1.771	48.065 ± 3.447	15.508 ± 2.749
Ours (Ablation)	PAF-CG	78.4 ± 3.615	78.4 ± 3.615	96.873 ± 1.513	39.744 ± 4.761	39.461 ± 4.766
	PAF-NN	78.519 ± 4.546	79.124 ± 5.104	96.827 ± 1.789	40.269 ± 4.306	37.453 ± 3.605

Table 4.6: Brazilian Admissions Dataset results. For details of metrics see table 4.4.

		Positive Predictive Rate			Accuracy	
	Model	PPR($Y = \{1\}, S = 0$)	PPR($Y = \{1, 2\}, S = 0$)	PPR($Y = \{1\}, S = 1$)	Acc($Y = \{1\}$)	Acc($Y = \{1, 2\}$)
Baseline	Oracle	37.93 ± 0.523	-	37.931 ± 0.526	91.137 ± 0.387	-
	FL	37.369 ± 0.506	-	47.51 ± 1.014	64.036 ± 0.676	-
	Reg	25.515 ± 0.508	-	64.898 ± 1.154	65.799 ± 0.45	-
	LR	37.046 ± 0.957	-	46.451 ± 3.168	64.035 ± 0.637	-
	RW	40.325 ± 0.327	-	42.707 ± 0.742	63.595 ± 0.623	-
	DIR	40.807 ± 0.403	-	41.831 ± 0.827	63.433 ± 0.578	-
Ours	PAF	41.102 ± 3.862	49.68 ± 6.441	44.712 ± 3.967	64.279 ± 0.806	62.56 ± 1.209
Ours (Naïve)	(EQ-DP) (FL)	37.046 ± 0.957	37.665 ± 0.835	36.967 ± 1.044	56.855 ± 0.489	56.781 ± 0.504
	(EQ-DP) (RW)	37.046 ± 0.957	41.988 ± 0.449	36.967 ± 1.044	56.855 ± 0.489	56.177 ± 0.475
	(EQ-DP) (Reg)	37.046 ± 0.957	37.233 ± 0.97	36.967 ± 1.044	56.855 ± 0.489	56.894 ± 0.479
	(EQ-DP) (DIR)	37.046 ± 0.957	42.755 ± 0.635	36.967 ± 1.044	56.855 ± 0.489	56.028 ± 0.421
	(LR-DP) (FL)	37.046 ± 0.957	37.665 ± 0.835	46.451 ± 3.168	64.035 ± 0.637	63.961 ± 0.72
	(LR-DP) (RW)	37.046 ± 0.957	41.988 ± 0.449	46.451 ± 3.168	64.035 ± 0.637	63.357 ± 0.683
	(LR-DP) (Reg)	37.046 ± 0.957	37.233 ± 0.97	46.451 ± 3.168	64.035 ± 0.637	64.074 ± 0.629
	(LR-DP) (DIR)	37.046 ± 0.957	42.755 ± 0.635	46.451 ± 3.168	64.035 ± 0.637	63.208 ± 0.65
Ours (Ablation)	PAF-CG	40.837 ± 2.133	42.751 ± 2.523	43.934 ± 3.506	64.364 ± 0.587	64.027 ± 0.615
	PAF-NN	41.324 ± 2.991	45.837 ± 2.776	44.52 ± 2.799	64.311 ± 0.672	63.517 ± 0.684

learning has the potential to help identify applicants who would benefit from this additional support.

We consider the different mechanisms that can lead to an observed disparity in the rate of positive outcomes between a protected subgroup and the majority. We highlight that, at least in part, this disparity can be due to disadvantages affecting applicants belonging to a protected subgroup, hindering their ability to compete with other applicants.

We demonstrated that an adversarially trained model can determine positive action outcomes and justified this model selection with a range of alternative implementations and ablations.

Table 4.7: UCI Adult Income Dataset results. For details of metrics see table 4.4.

		Positive Predictive Rate			Accuracy	
	Model	PPR($Y = \{1\}, S = 0$)	PPR($Y = \{1, 2\}, S = 0$)	PPR($Y = \{1\}, S = 1$)	Acc($Y = \{1\}$)	Acc($Y = \{1, 2\}$)
Baseline	`Oracle`	31.305 ± 0.466	-	31.303 ± 0.463	93.513 ± 0.25	-
	`FL`	9.19 ± 0.458	-	24.088 ± 0.535	84.335 ± 0.397	-
	`Reg`	7.277 ± 0.529	-	26.368 ± 0.609	84.445 ± 0.35	-
	`LR`	8.182 ± 0.4	-	26.142 ± 0.645	84.708 ± 0.424	-
	`RW`	12.379 ± 0.484	-	21.783 ± 0.495	83.86 ± 0.334	-
	`DIR`	12.641 ± 3.163	-	21.2 ± 3.045	83.674 ± 0.789	-
Ours	`PAF`	20.535 ± 1.555	38.036 ± 5.587	42.671 ± 2.152	80.326 ± 1.1	75.273 ± 1.839
Ours (Naïve)	`(EQ-DP) (FL)`	16.717 ± 0.362	17.977 ± 0.317	18.776 ± 0.766	80.009 ± 0.424	79.92 ± 0.441
	`(EQ-DP) (RW)`	16.717 ± 0.362	20.754 ± 0.333	18.776 ± 0.766	80.009 ± 0.424	79.654 ± 0.421
	`(EQ-DP) (Reg)`	16.717 ± 0.362	17.176 ± 0.402	18.776 ± 0.766	80.009 ± 0.424	80.008 ± 0.438
	`(EQ-DP) (DIR)`	16.713 ± 0.365	20.838 ± 2.789	18.773 ± 0.767	79.992 ± 0.422	79.499 ± 0.605
	`(LR-DP) (FL)`	8.182 ± 0.4	9.591 ± 0.384	26.142 ± 0.645	84.708 ± 0.424	84.624 ± 0.435
	`(LR-DP) (RW)`	8.182 ± 0.4	12.616 ± 0.451	26.142 ± 0.645	84.708 ± 0.424	84.332 ± 0.394
	`(LR-DP) (Reg)`	8.182 ± 0.4	8.689 ± 0.447	26.142 ± 0.645	84.708 ± 0.424	84.718 ± 0.433
	`(LR-DP) (DIR)`	8.17 ± 0.383	12.983 ± 3.181	26.167 ± 0.622	84.724 ± 0.408	84.156 ± 0.469
Ours (Ablation)	`PAF-CG`	19.109 ± 1.225	19.967 ± 1.242	42.701 ± 2.016	80.732 ± 0.853	80.584 ± 0.857
	`PAF-NN`	18.866 ± 0.776	21.888 ± 1.068	42.582 ± 1.859	80.751 ± 0.793	80.021 ± 0.868

Table 4.8: UCI Communities and Crime Dataset results. For details of metrics see table 4.4.

		Positive Predictive Rate			Accuracy	
	Model	PPR($Y = \{1\}, S = 0$)	PPR($Y = \{1, 2\}, S = 0$)	PPR($Y = \{1\}, S = 1$)	Acc($Y = \{1\}$)	Acc($Y = \{1, 2\}$)
Baseline	`Oracle`	30.237 ± 18.208	-	30.219 ± 18.141	81.734 ± 2.087	-
	`FL`	15.554 ± 3.084	-	34.153 ± 3.857	81.457 ± 2.441	-
	`Reg`	8.139 ± 2.03	-	51.496 ± 3.171	83.894 ± 1.789	-
	`LR`	11.637 ± 2.277	-	50.341 ± 3.627	84.899 ± 1.97	-
	`RW`	15.857 ± 2.967	-	42.405 ± 2.758	83.065 ± 1.963	-
	`DIR`	18.385 ± 4.447	-	34.347 ± 7.179	76.96 ± 4.345	-
Ours	`PAF`	28.63 ± 3.921	54.972 ± 5.069	57.524 ± 3.15	79.422 ± 2.249	68.065 ± 2.151
Ours (Naïve)	`(EQ-DP) (FL)`	11.637 ± 2.277	15.954 ± 3.119	10.897 ± 1.161	71.457 ± 2.222	70.905 ± 1.933
	`(EQ-DP) (RW)`	11.637 ± 2.277	16.008 ± 3.029	10.897 ± 1.161	71.457 ± 2.222	70.829 ± 2.002
	`(EQ-DP) (Reg)`	11.637 ± 2.277	11.836 ± 2.37	10.897 ± 1.161	71.457 ± 2.222	71.558 ± 2.111
	`(EQ-DP) (DIR)`	11.637 ± 2.277	20.563 ± 4.561	10.897 ± 1.161	71.457 ± 2.222	68.894 ± 2.743
	`(LR-DP) (FL)`	11.637 ± 2.277	15.954 ± 3.119	50.341 ± 3.627	84.899 ± 1.97	84.347 ± 1.959
	`(LR-DP) (RW)`	11.637 ± 2.277	16.008 ± 3.029	50.341 ± 3.627	84.899 ± 1.97	84.271 ± 2.051
	`(LR-DP) (Reg)`	11.637 ± 2.277	11.836 ± 2.37	50.341 ± 3.627	84.899 ± 1.97	85.0 ± 1.84
	`(LR-DP) (DIR)`	11.637 ± 2.277	20.563 ± 4.561	50.341 ± 3.627	84.899 ± 1.97	82.337 ± 2.654
Ours (Ablation)	`PAF-CG`	25.49 ± 3.017	28.145 ± 3.934	57.117 ± 2.728	79.749 ± 2.251	78.769 ± 2.366
	`PAF-NN`	23.897 ± 4.638	28.934 ± 4.574	54.601 ± 4.79	80.628 ± 2.615	78.643 ± 2.838

Although this is a first attempt to include positive action in a decision-support setting, our counterfactual implementation achieves our goal: it maintains predictive utility while minimising the rejection of candidates with high potential from the disadvantaged group. There are surely improvements that can be made, but the results are encouraging and warrant further investigation. We hope this work will form part of a larger, constructive discussion around the role of machine learning in promoting the use and effectiveness of positive action measures.

Table 4.9: NYPD Stop, Question and Frisk Dataset results. For details of metrics see table 4.4.

		Positive Predictive Rate			Accuracy	
	Model	PPR($Y = \{1\}, S = 0$)	PPR($Y = \{1, 2\}, S = 0$)	PPR($Y = \{1\}, S = 1$)	Acc($Y = \{1\}$)	Acc($Y = \{1, 2\}$)
Baseline	`Oracle`	10.468 ± 0.714	-	10.5 ± 0.718	99.797 ± 0.113	-
	`FL`	4.291 ± 1.241	-	5.729 ± 0.421	91.956 ± 0.572	-
	`Reg`	4.527 ± 1.4	-	5.984 ± 0.448	91.847 ± 0.486	-
	`LR`	3.539 ± 1.382	-	5.527 ± 0.466	92.001 ± 0.584	-
	`RW`	3.732 ± 1.229	-	5.725 ± 0.445	92.029 ± 0.553	-
	`DIR`	3.571 ± 1.17	-	5.869 ± 0.419	91.981 ± 0.507	-
Ours	`PAF`	10.291 ± 2.734	12.707 ± 2.664	13.407 ± 2.795	89.291 ± 1.484	89.154 ± 1.417
Ours (Naïve)	`(EQ-DP) (FL)`	0.532 ± 0.039	3.931 ± 1.265	0.0439 ± 0.0003	89.413 ± 0.701	89.445 ± 0.786
	`(EQ-DP) (RW)`	0.536 ± 0.04	3.732 ± 1.229	0.0438 ± 0.0003	89.737 ± 0.720	89.818 ± 0.755
	`(EQ-DP) (Reg)`	0.536 ± 0.04	4.575 ± 1.351	0.0438 ± 0.0003	89.737 ± 0.720	89.700 ± 0.718
	`(EQ-DP) (DIR)`	0.536 ± 0.04	3.621 ± 1.111	0.0438 ± 0.0003	89.737 ± 0.720	89.810 ± 0.752
	`(LR-DP) (FL)`	3.539 ± 1.382	4.457 ± 1.296	5.527 ± 0.466	92.001 ± 0.584	91.952 ± 0.620
	`(LR-DP) (RW)`	3.539 ± 1.382	3.896 ± 1.275	5.527 ± 0.466	92.001 ± 0.584	91.989 ± 0.604
	`(LR-DP) (Reg)`	3.539 ± 1.382	5.747 ± 1.057	5.527 ± 0.466	92.001 ± 0.584	91.867 ± 0.559
	`(LR-DP) (DIR)`	3.539 ± 1.382	3.835 ± 1.254	5.527 ± 0.466	92.001 ± 0.584	91.985 ± 0.601
Ours (Ablation)	`PAF-CG`	9.659 ± 2.198	12.686 ± 5.226	13.094 ± 2.479	89.514 ± 1.249	89.324 ± 1.254
	`PAF-NN`	10.465 ± 3.017	14.775 ± 4.347	13.777 ± 2.915	89.072 ± 1.534	88.862 ± 1.651

4.9 Appendix

4.9.1 *Data Generation*

We define a data generation procedure for a dataset with binary S-labels and a binary outcome, with 2 imperfect observers of 3 features, creating a feature space $\mathcal{X}$ comprising 6 features. We first draw samples for S from a Bernoulli distribution and model the underlying construct as a Uniform distribution (figure 4.7(i)) — this is where the WAE worldview is applied, as $\tilde{X}_{apt}$ is independent of S:

$$S \sim \mathcal{B}(0.5) \quad \text{and} \quad \tilde{X}_{apt} \sim \mathcal{U}(0, 1)$$

To represent unequal treatment prior to observation, we map from the uniform distribution to an S-conditioned distribution for each feature using an inverse-CDF (percent point) function, $\Delta : \tilde{\mathcal{X}}_{apt}, \mathcal{S} \mapsto \tilde{\mathcal{X}}$. This mapping is captured by $\tilde{X} = \Delta(\tilde{X}_{apt}, S)$ (figure 4.7(ii)).

The features $\tilde{X}$ are still in the construct space, representing the potential to successfully graduate from the university course at the point of applying. Faithful measurement of this data will more closely align with the WYSIWYG worldview. The mapping from construct to observation $\text{Obs} : \tilde{\mathcal{X}}, \mathcal{S} \mapsto \mathcal{X}$ is made of two noisy observations for each feature. A measurement bias further aggravates the disparity between the blue and green distributions (figure 4.7(iii)). We then generate two outcome scores: 1) An ‘acceptance score’ based on a linear combination of the observed features with a label bias introduced by setting different acceptance thresholds depending on the value of S (figure 4.7(iv)). 2) A 'graduation grade' based on a linear combination of the features in $\tilde{X}$, bypassing the effect of the introduced measurement bias and label bias.

We then take the inverse-CDF (product point function) of a distribution at point $\tilde{\boldsymbol{x}}_b$ (CDF^{-1}(distribution, point)) for three unobserved features.

$$\tilde{\boldsymbol{x}}_0 \sim \begin{cases} CDF^{-1}(\mathscr{N}(0.65, 0.15), \tilde{\boldsymbol{x}}_b), & \text{if } s = 1 \\ CDF^{-1}(\mathscr{J}_U(-2, 3, 0.35, 0.2), \tilde{\boldsymbol{x}}_b), & \text{otherwise} \end{cases}$$

$$\tilde{\boldsymbol{x}}_1 \sim \mathscr{N}(0.4 + (2s - 1), 0.2)$$

$$\tilde{\boldsymbol{x}}_2 \sim \begin{cases} CDF^{-1}(\mathscr{L}(0.5, 0.075), \tilde{\boldsymbol{x}}_b), & \text{if } s = 1 \\ CDF^{-1}(\mathscr{T}(100, 0.4, 0.15), \tilde{\boldsymbol{x}}_b), & \text{otherwise} \end{cases}$$

Where $\mathscr{N}$ is a Normal distribution, $\mathscr{J}_U$ is Johnsons-SU distribution, $\mathscr{L}$ is a Laplace distribution and $\mathscr{T}$ is a Student-T distribution.

We then have two imperfect observers of each feature. Both observers add noise from a Normal distribution, but with different mean and standard deviation values.

$$\begin{array}{llll} \tilde{\boldsymbol{x}}_0 \text{ Observer 1}: & \tilde{\boldsymbol{x}}_0 + \mathscr{N}(0.03, 0.02) & \tilde{\boldsymbol{x}}_0 \text{ Observer 2}: & \tilde{\boldsymbol{x}}_0 + \mathscr{N}(0.01, 0.04) \\ \tilde{\boldsymbol{x}}_1 \text{ Observer 1}: & \tilde{\boldsymbol{x}}_1 + \mathscr{N}(0, 0.02) & \tilde{\boldsymbol{x}}_1 \text{ Observer 2}: & \tilde{\boldsymbol{x}}_1 + \mathscr{N}(0, 0.05) \\ \tilde{\boldsymbol{x}}_2 \text{ Observer 1}: & \tilde{\boldsymbol{x}}_2 + \mathscr{N}(0.03, 0.01) & \tilde{\boldsymbol{x}}_2 \text{ Observer 2}: & \tilde{\boldsymbol{x}}_2 + \mathscr{N}(0.01, 0.02) \end{array}$$

The admittance score (Y) is based on a combination of the mean observation per feature. Let N be the number of observers.

$$\tilde{Y} = 0.4 \cdot (\frac{1}{N}\sum_{i=0}^{N} \tilde{\boldsymbol{x}}_0 \text{ Observer}_i) + 0.4 \cdot (\frac{1}{N}\sum_{i=0}^{N} \tilde{\boldsymbol{x}}_1 \text{ Observer}_i) + 0.2 \cdot (\frac{1}{N}\sum_{i=0}^{N} \tilde{\boldsymbol{x}}_2 \text{ Observer}_i)$$

Then, to incorporate direct discrimination, a factor γ is added to the admission score.

$$Y = \begin{cases} \tilde{Y} + \gamma, & \text{if } s = 1 \\ \tilde{Y} - \gamma, & \text{otherwise} \end{cases}$$

During our experiments, we set $\gamma = 0.02$.

We also model the final graduation grade. We model this as a binary label, 'good graduating grade' or 'not good graduating grade'. This is based on the unobserved score for each feature, and is different per subgroup to reflect that one measure of success need not be consistent across all of the population.

$$G = \begin{cases} 0.3\tilde{\boldsymbol{x}}_0 + 0.25\tilde{\boldsymbol{x}}_1 + 0.45\tilde{\boldsymbol{x}}_2, & \text{if } s = 1 \\ 0.1\tilde{\boldsymbol{x}}_0 + 0.7\tilde{\boldsymbol{x}}_1 + 0.2\tilde{\boldsymbol{x}}_2, & \text{otherwise} \end{cases}$$

To go from a score to a classification, we take a data-dependent threshold so that the top 20% of the candidates will be accepted. Across 10 seeds, this threshold is 0.585 ± 0.005. If the score is greater than this threshold value, an outcome of 1 is assigned. For the 'graduation' grade, we use a consistent threshold of 0.6 for all seeds.

4.9.2 *Model Training Techniques*

A number of techniques were used during training to make the results robust to dataset splits and random weight initialisation.

Sample Balancer: A 'memory bank' of samples is kept. During training, batches are balanced with regard to S for the Generator, and with regard to both S and Y for the classifier. This is achieved by 'upsampling' from the memory bank during training for groups that have fewer samples per mini-batch than the maximally represented group.

Mixup Data Augmentation: During training of the classifier mixup is used to provide richer target labels. The distribution for mixup is uniform in the range 0–0.49 so that the original sample remains the prominant component. The samples are mixed across S-groups. X_Mixed is a linear interpolation of samples from different S groups, so the X will be mixed, and the S will be mixed, and the Y may be, with the original sample the more dominant. Due to the samples only being partially interpolated, the S-label with the original sample remains dominant, and is used for indexing, and as the discriminator target.

4.9.3 *Expanded Outcome Comparator*

Table 4.10: Expanded version of Table 4.1. All combinations of ensemble outcomes are shown including those that directly violate our assumption that the minority group is at a disadvantage.

Ensemble Outcomes					Outcome Groups	
$y_{s_x=0,s_y=0}$	$y_{s_x=0,s_y=1}$	$y_{s_x=1,s_y=0}$	$y_{s_x=1,s_y=1}$	S-label	Selection Group	Outcome Label
0	0	0	0	0	5	Reject
0	0	0	0	1	6	Reject
0	0	0	1	0	3	PAC
0	0	0	1	1	4	Accept
0	0	1	0	0	-	Fallback
0	0	1	0	1	-	Fallback
0	0	1	1	0	3	PAC
0	0	1	1	1	4	Accept
0	1	0	0	0	-	Fallback
0	1	0	0	1	-	Fallback
0	1	0	1	0	1	Accept
0	1	0	1	1	4	Accept
0	1	1	0	0	-	Fallback
0	1	1	0	1	-	Fallback
0	1	1	1	0	1	Accept
0	1	1	1	1	2	Accept
1	0	0	0	0	-	Fallback
1	0	0	0	1	-	Fallback
1	0	0	1	0	-	Fallback
1	0	0	1	1	-	Fallback
1	0	1	0	0	-	Fallback
1	0	1	0	1	-	Fallback
1	0	1	1	0	-	Fallback
1	0	1	1	1	-	Fallback
1	1	0	0	0	-	Fallback
1	1	0	0	1	-	Fallback
1	1	0	1	0	-	Fallback
1	1	0	1	1	-	Fallback
1	1	1	0	0	-	Fallback
1	1	1	0	1	-	Fallback
1	1	1	1	0	1	Accept
1	1	1	1	1	2	Accept

5 PAPER 2: DISCOVERING FAIR REPRESENTATIONS IN THE DATA DOMAIN

AUTHORS: Novi Quadrianto[1,2], Viktoriia Sharmanska[3] and Oliver Thomas[1]

AFFILIATIONS:

[1] Predictive Analytics Lab (PAL), University of Sussex, Brighton, UK

[2] Higher School of Economics, Moscow, Russia

[3] Department of Computing, Imperial College London, UK

CONFERENCE: *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019

DOI: 10.1109/cvpr.2019.00842

5.1 ABSTRACT

Interpretability and fairness are critical in computer vision and machine learning applications, in particular when dealing with human outcomes, e.g. inviting or not inviting for a job interview based on application materials that may include photographs. One promising direction to achieve fairness is by learning data representations that remove the semantics of protected characteristics, and are therefore able to mitigate unfair outcomes. All available models however learn latent embeddings which comes at the cost of being uninterpretable. We propose to cast this problem as data-to-data translation, i.e. learning a mapping from an input domain to a fair target domain, where a fairness definition is being enforced. Here the data domain can be images, or any tabular data representation. This task would be straightforward if we had fair target data available, but this is not the case. To overcome this, we learn a highly unconstrained mapping by exploiting statistics of residuals – the difference between input data and its translated version – and the protected characteristics. When applied to the CelebA dataset of face images with gender attribute as the protected characteristic, our model enforces equality of opportunity by adjusting the eyes and lips regions. Intriguingly, on the same dataset we arrive at similar conclusions when using semantic attribute representations of images for translation. On face images of the recent DiF dataset, with

the same gender attribute, our method adjusts nose regions. In the Adult income dataset, also with protected gender attribute, our model achieves equality of opportunity by, among others, obfuscating the wife and husband relationship. Analyzing those systematic changes will allow us to scrutinize the interplay of fairness criterion, chosen protected characteristics, and prediction performance.

5.2 Introduction

Machine learning systems are increasingly used by government agencies, businesses, and other organisations to assist in making life-changing decisions such as whether or not to invite a candidate to a job interview, or whether to give someone a loan. The question is how can we ensure that those systems are *fair*, i.e. they do not discriminate against individuals because of their gender, disability, or other personal ('protected') characteristics? For example, in building an automated system to review job applications, a photograph might be used in addition to other features to make an invite decision. By using the photograph as is, a discrimination issue might arise, as photographs with faces could reveal certain protected characteristics, such as gender, race, or age (e.g. Brown and Perrett, 1993; Bruce et al., 1993; Fu et al., 2014; Levi and Hassncer, 2015). Therefore, any automated system that incorporates photographs into its decision process is at risk of indirectly conditioning on protected characteristics (indirect discrimination). Recent advances in learning fair representations suggest adversarial training as the means to hide the protected characteristics from the decision/prediction function (Beutel et al., 2017; Madras et al., 2018a; Zhang et al., 2018a). All fair representation models, however, learn *latent embeddings*. If we want to encourage public conversations and productive public debates regarding fair machine learning systems (Global Future Council on Human Rights 2016-18, 2018), interpretability in how fairness is met is an integral yet overlooked ingredient.

In this paper we focus on representation learning models that can transform inputs to their fair representations and retain the semantics of the input domain in the transformed space. When we have image data, our method will make a semantic change to the appearance of an image to deliver a certain fairness criterion[1]. To achieve this, we perform *a data-to-data translation* by learning a mapping from data in a source domain to a target domain. Mapping from source

[1] Examples of fairness criteria are equality of true positive rates (TPR), also called equality of opportunity (Hardt et al., 2016; Zafar et al., 2017a), between males and females.

to target domain is a standard procedure, and many methods are available. For example, in the image domain, if we have aligned source/target as training data, we can use the pix2pix method of Isola et al. (2017), which is based on conditional generative adversarial networks (cGANs) (Mirza and Osindero, 2014). Zhu et al. (2017)'s CycleGAN and Choi et al. (2018)'s StarGAN solve a more challenging setting in which only *un*aligned training examples are available. However, we can not simply reuse existing methods for source-to-target mapping because we do *not have data in the target domain* (e.g. fair images are not available; images by themselves can not be fair or unfair, it is only when they are coupled with a particular task that the concern of fairness arises).

To illustrate the difficulty, consider our earlier example of an automated job review system that uses photographs as part of an input. For achieving fairness, it is tempting to simply use GAN-driven methods to *translate female face photos to male*. We would require training data of female faces (source domain) and male faces (target domain), and only unaligned training data would be needed. This solution is however fundamentally flawed; who gets to decide that we should translate in this direction? Is it fairer if we translate male faces to female instead? An ethically grounded approach would be to translate both male and female face photos (source domain) to appropriate middle ground face photos (target domain). This challenge is actually multi-dimensional, it contains at least *two sub-problems*: a) how to have a general approach that can handle image data as well as tabular data (e.g. work experience, education, or even semantic attribute representations of photographs), and b) how to find a middle-ground with a multi-value (e.g. race) or continuous value (e.g. age) protected characteristic or even multiple characteristics (e.g. race and age).

We propose a solution to the multi-dimensional challenge described above by exploiting statistical (in)dependence between translated images and protected characteristics. We use the Hilbert-Schmidt norm of the cross-covariance operator between reproducing kernel Hilbert spaces of image features and protected characteristics (Hilbert-Schmidt independence criterion Gretton et al., 2005) as an empirical estimate of statistical independence. This flexible measure of independence allows us to take into account higher order independence, and handle a multi-/continuous value and multiple protected characteristics.

Related work We focus on expanding the related topic of learning fair, *albeit uninterpretable*, representations. The aim of fair representation learning is to learn an intermediate representation of the data that preserves as much information about the data as possible, while simultaneously removing protected characteristic information such as age and gender. Zemel et al. (2013) learn a probabilistic mapping of the data point to a set of latent prototypes that is independent of protected characteristic (equality of acceptance rates, also called a statistical parity criterion), while retaining as much class label information as possible. Louizos et al. (2016) extend this by employing a deep variational auto-encoder (VAE) framework for finding the fair latent representation. In recent years, we see increased adversarial learning methods for fair representations. Ganin et al. (2016) propose adversarial representation learning for domain adaptation by requiring the learned representation to be indiscriminate with respect to differences in the domains. Multiple data domains can be translated into multiple demographic groups. Edwards and Storkey (2016) make this connection and propose adversarial representation learning for the statistical parity criterion. To achieve other notions of fairness such as equality of opportunity, Beutel et al. (2017) show that the adversarial learning algorithm of Edwards and Storkey (2016) can be reused but we only supply training data with positive outcome to the adversarial component. Madras et al. (2018a) use a label-aware adversary to learn fair and transferable latent representations for the statistical parity as well as equality of opportunity criteria.

None of the above learn fair representations while simultaneously retaining the semantic meaning of the data. There is an orthogonal work on feature selection using human perception of fairness (e.g. Grgic-Hlaca et al., 2018), while this approach undoubtedly retains the semantic meaning of tabular data, it has not been generalized to image data. In an independent work to ours, Sattigeri et al. (2019) describe a similar motivation of producing fair representations in the input image domain; their focus is on creating a whole new image-like dataset, rather than conditioning on each input image. Hence it is not possible to visualise a fair version for a given image as provided by our method (refer to figure 5.2 and figure 5.3).

5.3 Interpretability in Fairness by Residual Decomposition

We will use the illustrative example of an automated job application screening system. Given input data (photographs, work experience, education and training, personal skills, etc.) $\mathbf{x}^n \in \mathcal{X}$, output labels of performed well or not well $y^n \in \mathcal{Y} = \{+1, -1\}$, and protected characteristic values, such as *race* or *gender*, $s^n \in \{A, B, C, D, \ldots\}$, or *age*, $s^n \in \mathbb{R}$, we would like to train a classifier f that decides whether or not to invite a person for an interview. We want the classifier to predict outcomes that are accurate with respect to y^n but fair with respect to s^n.

5.3.1 *Fairness definitions*

Much work has been done on mathematical definitions of fairness (e.g. Chouldechova, 2017; Kleinberg et al., 2017). It is widely accepted that no single definition of fairness applies in all cases, but will depend on the specific context and application of machine learning models (Global Future Council on Human Rights 2016-18, 2018). In this paper, we focus on the *equality of opportunity* criterion that requires the classifier f and the protected characteristic s be independent, conditional on the label being positive [2], in shorthand notation $f \perp\!\!\!\perp s \mid y = +1$. Expressing the shorthand notation in terms of a conditional distribution, we have $\mathbb{P}(f(\mathbf{x})|s, y = +1) = \mathbb{P}(f(\mathbf{x})|y = +1)$. With binary protected characteristic, this reads as equal true positive rates across the two groups, $\mathbb{P}(f(\mathbf{x}) = +1|s = A, y = +1) = \mathbb{P}(f(\mathbf{x}) = +1|s = B, y = +1)$. Equivalently, the shorthand notation can also be expressed in terms of joint distributions, resulting in $\mathbb{P}(f(\mathbf{x}), s|y = +1) = \mathbb{P}(f(\mathbf{x})|y = +1)\mathbb{P}(s|y = +1)$. The advantage of using the joint distribution expression is that the variable s does not appear as a conditioning variable, making it straightforward to use the expression for a multi- or continuous value or even multiple protected characteristics.

5.3.2 *Residual decomposition*

We want to learn a data representation $\tilde{\mathbf{x}}^n$ for each input $\mathbf{x}^n$ such that: a) it is able to predict the output label y^n, b) it protects s^n according to a certain fairness criterion, c) it lies in the same space as $\mathbf{x}^n$, that is $\tilde{\mathbf{x}}^n \in \mathcal{X}$. The third requirement ensures the learned representation to have the same *semantic meaning* as the input. For example, for images of people faces, the goal is to modify facial appearance in order to remove the protected characteristic information. For tabular data, we desire systematic changes in values of categorical features such as education (bachelors, masters, doctorate, etc.). Visualizing those systematic changes will give evidence on how our algorithm enforces a certain fairness criterion. This will be a powerful tool, albeit all the powers hinge on *observational data*, to scrutinize the interplay between fairness criterion, protected characteristics, and classification accuracy. We proceed by making the following decomposition assumption on $\mathbf{x}$:

$$\phi(\mathbf{x}) = \phi(\tilde{\mathbf{x}}) + \phi(\hat{\mathbf{x}}), \tag{5.1}$$

[2] With binary labels, it is assumed that positive label is a desirable/advantaged outcome, e.g. expected to perform well at the job.

with $\tilde{\mathbf{x}}$ to be the component that is independent of s, $\hat{\mathbf{x}}$ denoting the component of $\mathbf{x}$ that is dependent on s, and $\phi(\cdot)$ is some *pre-trained* feature map. We will discuss about the specific choice of this pre-trained feature map for both image and tabular data later in the section. What we want is to learn a mapping from a source domain (input features) to a target domain (fair features with the semantics of the input domain), i.e. $T : \mathbf{x} \to \tilde{\mathbf{x}}$, and we will parameterize this mapping $T = T_\omega$ where ω is a class of autoencoding transformer network. For our architectural choice of transformer network, please refer to section 5.4.

To enforce the decomposition structure in equation (5.1), we need to satisfy two conditions: a) $\tilde{\mathbf{x}}$ to be independent of s, and b) $\hat{\mathbf{x}}$ to be dependent of s. Given a particular statistical dependence measure, the first condition can be achieved by *minimizing* the dependence measure between $P = \{\phi(\tilde{\mathbf{x}}^1), \dots, \phi(\tilde{\mathbf{x}}^N)\} = \{\phi(T_\omega(\mathbf{x}^1)), \dots, \phi(T_\omega(\mathbf{x}^N))\}$ and $S = \{s^1, \dots, s^N\}$; N is the number of training data points. For the second condition, we first define a *residual*:

$$\phi(\mathbf{x}) - \phi(\tilde{\mathbf{x}}) = \phi(\mathbf{x}) - \phi(T_\omega(\mathbf{x})) = \phi(\hat{\mathbf{x}}), \tag{5.2}$$

where the last term is the data component that is *dependent* on a protected characteristic s. We can then enforce the second condition by *maximizing* the dependence measure between $R = \{\phi(\hat{\mathbf{x}}^1), \dots, \phi(\hat{\mathbf{x}}^N)\} = \{\phi(\mathbf{x}^1) - \phi(T_\omega(\mathbf{x}^1)), \dots, \phi(\mathbf{x}^N) - \phi(T_\omega(\mathbf{x}^N))\}$ and S. We use the decomposition property as a guiding mechanism to learn the parameters ω of the transformer network T_ω.

In the fair and interpretable representation learning task, we believe using residual is well-motivated because we know that our generated fair features should be somewhat similar to our input features. Residuals will make learning the transformer network easier. Taking into consideration that we do not have training data about the target fair features $\tilde{\mathbf{x}}$, we should not desire the transformer network to take the input feature $\mathbf{x}$ and *generate* a new output $\tilde{\mathbf{x}}$. Instead, it should just learn how to *adjust* our input $\mathbf{x}$ to produce the desired output $\tilde{\mathbf{x}}$. The concept of residuals is universal, for example, a residual block has been used to speed up and to prevent over-fitting of a very deep neural network (He et al., 2016), and a residual regression output has been used to perform causal inference in additive noise models (Mooij et al., 2009).

Formally, given the N training triplets (X, S, Y), to find a fair and interpretable representation $\tilde{\mathbf{x}} = T_\omega(\mathbf{x})$, our optimization problem is given by:

$$\underset{T_\omega}{\text{minimize}} \underbrace{\sum_{n=1}^{N} \mathscr{L}(T_\omega(\mathbf{x}^n), y^n)}_{\text{prediction loss}} + \lambda_1 \underbrace{\sum_{n=1}^{N} \|\mathbf{x}^n - T_\omega(\mathbf{x}^n)\|_2^2}_{\text{reconstruction loss}} + \\ + \lambda_2 \left(\underbrace{- \operatorname{HSIC}(R, S|Y = +1) + \operatorname{HSIC}(P, S|Y = +1)}_{\text{decomposition loss}} \right) \tag{5.3}$$

where $\operatorname{HSIC}(\cdot, \cdot)$ is the statistical dependence measure, and λ_i are trade-off parameters. HSIC is the Hilbert-Schmidt norm of the cross-covariance operator between reproducing kernel Hilbert spaces. This is equivalent to a non-parametric distance measure of a joint distribution and the product of two marginal distributions using the Maximum Mean Discrepancy (MMD) criterion (Gretton et al., 2012); MMD has been successfully used in fairnesss literature in it's own right (Louizos et al., 2016; Quadrianto and Sharmanska, 2017). section 5.3.1 discusses defining statistical independence based on a joint distribution, contrasting this with a conditional distribution. We use the biased estimator of HSIC (Gretton et al., 2005; Song et al., 2012): $\operatorname{HSIC}_{\text{emp.}} = (N - 1)^{-2} \operatorname{tr} HKHL$, where $K, L \in \mathbb{R}^{N \times N}$ are the kernel matrices for the *residual* set R and the protected characteristic set S respectively, i.e. $K_{ij} = k(r^i, r^j)$ and $L_{ij} = l(s^i, s^j)$ (similar definition for measuring independence between sets P and S). We use a Gaussian RBF kernel function for both $k(\cdot, \cdot)$ and $l(\cdot, \cdot)$. Moreover, $H_{ij} = \delta_{ij} - N^{-1}$ centres the observations of set R and set S in RKHS feature space. The prediction loss is defined using a softmax layer on the output of the transformer network. While in image data we add the total variation (TV) penalty (Mahendran and Vedaldi, 2015) on the fair representation to ensure spatial smoothness, we do not enforce any regularization term for tabular data. In summary, we learn a new representation $\tilde{\mathbf{x}}$ that removes statistical dependence on the protected characteristic s (by minimizing $\operatorname{HSIC}(P, S|Y = +1)$) and enforces the dependence of the residual $\mathbf{x} - \tilde{\mathbf{x}}$ and s (by maximizing $\operatorname{HSIC}(R, S|Y = +1)$). We can then train any classifier f using this new representation, and it will inherently satisfy the fairness criterion (Madras et al., 2018a).

NEURAL STYLE TRANSFER AND PRE-TRAINED FEATURE SPACE Neural style transfer (e.g. Gatys et al., 2016; Johnson et al., 2016) is a popular approach to perform an image-to-image translation. Our decomposition loss in equation (5.3) is reminiscent of a style loss used in neural style transfer models. The style loss is defined as the distance between second-order statistics of a style image and the translated image. Excellent results (Gatys et al., 2016; Johnson et al., 2016; Ulyanov et al., 2016, 2017) on neural style transfer rely on pre-trained features. Following this spirit, we also use a 'pre-trained' feature mapping $\phi(\cdot)$ in defining our decomposition loss. For image data, we take advantage of the powerful representation of deep Convolutional Neural Network (CNN) to define the mapping function (Gatys et al., 2016). The feature maps of $\mathbf{x}$ in the layer l of a CNN are denoted by $F^l_{\mathbf{x}} \in R^{N_l \times M_l}$ where N_l is the number of the feature maps in the layer l and M_l is the height times the width of the feature map. We use the vectorization of $F^l_{\mathbf{x}}$ as the required mapping $\phi(\mathbf{x}) = \mathrm{vec}(F^l_{\mathbf{x}})$. Several layers of a CNN will be used to define the full mapping (see section 5.4). For tabular data, we use the following random Fourier feature (Rahimi and Recht, 2007) mapping $\phi(\mathbf{x}) = \sqrt{2/D}\cos(\langle\theta, \mathbf{x}\rangle + b)$ with a bias vector $b \in \mathbb{R}^D$ that is uniformly sampled in $[0, 2\pi]$, and a matrix $\theta \in \mathbb{R}^{d \times D}$ where θ_{ij} is sampled from a Gaussian distribution. We have assumed the input data lies in a d-dimensional space, and we transform them to a D-dimensional space.

5.4 Experiments

We gave an illustrative example about screening job applications, however, no such data is publicly available. We will instead use publicly available data to simulate the setting. We conduct the experiments using three datasets: the CelebA image dataset[3] (Liu et al., 2015), the Diversity in Faces (DiF) dataset [4] (Merler et al., 2019), and the Adult income dataset[5] from the UCI repository (Dheeru and Karra Taniskidou, 2017). The CelebA dataset has a total of 202, 599 celebrity images. The images are annotated with 40 attributes that reflect appearance (hair color and style, face shape, makeup, for example), emotional state (smiling), gender, attractiveness, and age. For this dataset, we use gender as a binary protected characteristic, and attractiveness as the proxy measure of getting invited for a job interview in the world of fame. We randomly select 20K images for testing and use the rest for training the model. The DiF dataset has only been introduced very recently and contains nearly a million human face images reflecting diversity in ethnicity, age and gender. We include preliminary results using 200K images for training and 200K images for testing our model on this dataset. The images are annotated with attributes such as race, gender and age (both continual and discretized into seven age groups) as well as facial landmarks and facial symmetry features. For this dataset, we use gender as a binary protected characteristic, and the discretized age groups as a predictive task. The Adult income dataset is frequently used

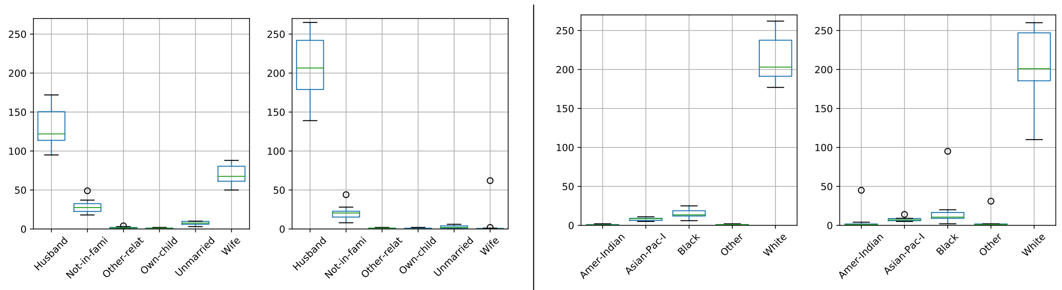
[3] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[4] https://www.research.ibm.com/artificial-intelligence/trusted-ai/diversity-in-faces/

[5] https://archive.ics.uci.edu/ml/datasets/adult

Table 5.1: Results of training multiple classifiers (rows 1–7) on 3 different representations, $\mathbf{x}$, $\tilde{\mathbf{x}}$, and $\mathbf{z}$. $\mathbf{x}$ is the original input representation, $\tilde{\mathbf{x}}$ is the interpretable, fair representation introduced in this paper, and $\mathbf{z}$ is the latent embedding representation of Beutel et al. (Beutel et al., 2017). We *boldface* Eq. Opp. since this is the fairness criterion (the lower the better). *The solver of `Zafar et al.` fails to converge in 4 out of 10 repeats. Our learned representation $\tilde{\mathbf{x}}$ achieves comparable fairness level to the latent representation $\mathbf{z}$, while maintaining the constraint of being in the same space as the original input.

	original $\mathbf{x}$		fair interpretable $\tilde{\mathbf{x}}$		latent embedding $\mathbf{z}$	
	Accuracy ↑	Eq. Opp ↓	Accuracy ↑	Eq. Opp ↓	Accuracy ↑	Eq. Opp ↓
1: `LR`	85.1 ± 0.2	**9.2** ± **2.3**	84.2 ± 0.3	**5.6** ± **2.5**	81.8 ± 2.1	**5.9** ± **4.6**
2: `SVM`	85.1 ± 0.2	**8.2** ± **2.3**	84.2 ± 0.3	**4.9** ± **2.8**	81.9 ± 2.0	**6.7** ± **4.7**
3: `Fair Reduction LR`[6]	85.1 ± 0.2	**14.9** ± **1.3**	84.1 ± 0.3	**6.5** ± **3.2**	81.8 ± 2.1	**5.6** ± **4.8**
4: `Fair Reduction SVM`[7]	85.1 ± 0.2	**8.2** ± **2.3**	84.2 ± 0.3	**4.9** ± **2.8**	81.9 ± 2.0	**6.7** ± **4.7**
5: `Kamiran & Calders LR`[8]	84.4 ± 0.2	**14.9** ± **1.3**	84.1 ± 0.3	**1.7** ± **1.3**	81.8 ± 2.1	**4.9** ± **3.3**
6: `Kamiran & Calders SVM`[9]	85.1 ± 0.2	**8.2** ± **2.3**	84.2 ± 0.3	**4.9** ± **2.8**	81.9 ± 2.0	**6.7** ± **4.7**
7: `Zafar et al.`*[10]	85.0 ± 0.3	**1.8** ± **0.9**	—	—	—	—



('Relationship Status')

('Race')

Figure 5.1: *Left* Boxplots showing the distribution of the categorical feature 'Relationship Status' *Right* Boxplots showing the distribution of the categorical feature 'Race'. *Left of each*: original representation $\mathbf{x} \in \mathcal{X}$. *Right of each*: fair representation $\tilde{\mathbf{x}} \in \mathcal{X}$.

to assess fairness methods. It comes from the Census bureau and the binary task is to predict whether or not an individual earns more than \$50K per year. It has a total of 45, 222 data instances, each with 14 features such as gender, marital status, educational level, number of work hours per week. For this dataset, we follow Zemel et al. (2013) and consider gender as a binary protected characteristic. We use 28, 222 instances for training, and 15, 000 instances for testing. We enforce equality of opportunity as the fairness criteria throughout for the three experiments.

5.4.1 *The Adult Income dataset*

The focus is to investigate whether (*Q1*) our proposed fair and interpretable learning method performs on a par with state-of-the-art fairness methods, and whether (*Q2*) performing a tabular-to-tabular translation brings us closer to achieving interpretability in how fairness is being satisfied. We compare our method against an unmodified $\mathbf{x}$ using the following classifiers: 1) logistic regression (`LR`) and 2) support vector machine with linear kernel (`SVM`), We select the regularization parameter of `LR` and `SVM` over 6 possible values (10^i for $i \in [0, 6]$) using 3-fold cross validation. We then train classifiers 1–2 with the learned representation $\tilde{\mathbf{x}}$ and with the latent embedding $\mathbf{z}$ of a state-of-the-art adversarial model described in Beutel et al. (2017). We also apply methods which reweigh the samples to simulate a balanced dataset with regard to the protected characteristic; FairLearn (Agarwal et al., 2018) `Fair Reduction` 3-4 and Kamiran & Calders (Kamiran and Calders, 2012) `Kamiran & Calders` 5-6, optimized with both the cross-validated `LR` and `SVM` (1-2), giving (`Fair Reduction LR`), (`Fair Reduction SVM`), (`Kamiran & Calders LR`) and (`Kamiran & Calders SVM`) respectively. As a reference, we also compare with: 7) Zafar et al. (2017a)'s fair classification method (`Zafar et al.)` that adds equality of opportunity directly as a constraint to the learning objective function. It has been shown that applying fairness constraints in succession as 'fair pipelines' do not enforce fairness (Bower et al., 2017; Dwork and Ilvento, 2019), as such, we only demonstrate (fair) classifier 7 on the unmodified $\mathbf{x}$.

Benchmarking We train our model for $50,000$ iterations using a network with 1 hidden layer of 40 nodes for both the encoder and decoder, with the encoded representation being 40 nodes. The predictor acts on the decoded output of this network. We set the trade-off parameters of the reconstruction loss (λ_1) and decomposition loss (λ_2) to 10^{-4} and 100 respectively. We then use this model to translate 10 different training and test sets into $\tilde{\mathbf{x}}$. Using a modified version of the framework provided by Friedler et al. (2018) we evaluate methods 1–6 using $\mathbf{x}$ and $\tilde{\mathbf{x}}$ representations. To ensure consistency, we train the model of Beutel et al. (2017) with the same architecture and number of iterations as our model.

Table 5.1 shows the results of these experiments. Our interpretable representation, $\tilde{\mathbf{x}}$ achieves similar fairness level to Beutel's state-of-the-art approach (*Q1*). Consistently, our representation $\tilde{\mathbf{x}}$ promoted the *fairness* criterion (Eq. Opp. close to 0), with only a small penalty in accuracy.

Interpretability We promote equality of opportunity for the positive class (actual salary > \$50K). In section 5.4 we show the effect of learning a fair representation, showing changes in the 'Relationship Status' and 'Race' features of samples that were incorrectly classified by an SVM as earning < \$50K in $\mathbf{x}$, but were correctly classified in $\tilde{\mathbf{x}}$. The visualization can be used for understanding how representation methods adjust the data for fairness. For example in section 5.4 (left) we can see that our method deals with the notorious problem of a husband or wife relationship status being a direct proxy for gender (*Q2*). Our method recognises this across all repeats in

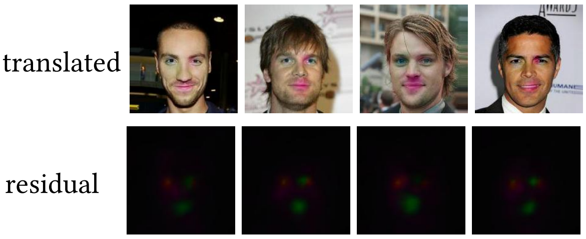


Figure 5.2: Examples of the translated and residual images on CelebA from the protected group of males (minority group) that have been classified correctly (as attractive) after transformation. These results are obtained with the transformer network for image-to-image translation. Best viewed in color.

Table 5.2: Results on CelebA dataset using a variety of input domains. Prediction performance is measured by accuracy, and we use equality of opportunity, TPRs difference, as the fairness criterion. Here, domain of fake images (last row) denotes images synthesized by the StarGAN(Choi et al., 2018) model from the original images and their fair attribute representations. We *emphasise* Eq. Opp. since this is the fairness criterion.

	domain	Acc.	Eq. Opp.	TPR	TPR
	$\mathcal{X}$	↑	↓	female	male
1: orig. $\mathbf{x}$	*images*	80.6	**33.8**	90.8	57.0
2: orig. $\mathbf{x}$	*attributes*	79.1	**39.9**	90.8	50.9
3: fair $\tilde{\mathbf{x}}$	*images*				
a: $\lambda_2 = 1.00$, biased HSIC		79.4	**23.8**	85.2	61.4
b: $\lambda_2 = 10.0$, biased HSIC		80.3	**22.8**	85.6	62.7
c: $\lambda_2 = 10.0$, unbiased HSIC		80.2	**18.7**	84.3	65.6
4: fair $\tilde{\mathbf{x}}$	*attributes*	75.9	**12.4**	87.2	74.8
5: fair $\tilde{\mathbf{x}}$	*fake images*	78.5	**23.0**	87.5	64.5

an unsupervised manner and reduces the wife category which is associated with a negative prediction. Other categories that have less correlation with the protected characteristic, such as race, largely remain unmodified (section 5.4 (right)).

5.4.2 *The CelebA dataset*

Our intention here is to investigate whether (*Q3*) performing an image-to-image translation brings us closer to achieving interpretability in how fairness is being satisfied, and whether (*Q4*) using semantic attribute representations of images reinforces similar interpretability conclusions as using image features directly.

Image-to-image translation Our autoencoder network is based on the architecture of the transformer network for neural style transfer (Johnson et al., 2016) with three convolutional layers, five residual layers and three deconvolutional/upsampling layers in combination with instance weight normalization (Ulyanov et al., 2017). The transformer network produces the residual image using a non-linear tanh activation, which is then subtracted from the input image to form the translated fair image $\tilde{\mathbf{x}}$. Similarly to neural style transfer (Gardner et al., 2016; Gatys et al., 2016; Johnson et al., 2016), for computing the loss terms, we use the activations in the deeper layers of the 19-layered VGG19 network (Simonyan and Zisserman, 2015) as feature representations of both input and translated images. Specifically, we use activations in the conv3_1, conv4_1 and conv5_1 layers for computing the decomposition loss, the conv3_1 layer activations for the reconstruction loss, and the activations in the last convolutional layer pool_5 for the prediction loss and when evaluating the performance. Given a 176x176 color input image, we compute the activations at each layer mentioned earlier after ReLU, then we flatten and l_2 normalize them to form features for the loss terms. In the HSIC estimates of the decomposition loss, we use a Gaussian RBF kernel $k(x_1, x_2) = \exp(-\gamma \|x_1 - x_2\|^2)$ width $\gamma = 1.0$ for image features, and $\gamma = 0.5$ for protected characteristics (as one over squared distance in the binary space). To compute the decomposition loss, we add the contributions across the three feature layers. We set the trade-off parameters λ_1 and λ_2 of the reconstruction loss and the decomposition loss, respectively, to 1.0 , and the TV regularization strength to 10^{-3}. Training was carried out for 50 epochs with a batch size of 80 images. We use minibatch SGD and apply the Adam solver (Kingma and Ba, 2015) with learning rate 10^{-3}; our TensorFlow implementation is publicly available[11].

Benchmarking and interpretability We enforce equality of opportunity as the fairness criterion, and we consider attractiveness as the positive label. Attractiveness is what could give someone a job opportunity or an advantaged outcome as defined in Hardt et al. (2016). To test the hypothesis that we have learned a fairer image representation, we compare the performance and fairness of a standard SVM classifier trained using original images and the translated fair images. We use activation in the pool_5 layer of the VGG19 network as features for training and evaluating the classifier[12].

We report the quantitative results of this experiment in table 5.2 (first and third rows) and the qualitative evaluations of image-to-image translations in figure 5.2. From the table 5.2 it is clear that the classifier trained on fair/translated images $\tilde{\mathbf{x}}$ has improved over the classifier trained on the original images $\mathbf{x}$ in terms of equality of opportunity (reduction from 33.8 to 23.8) while maintaining the prediction accuracy (79.4 comparing to 80.6). The reduction in equality of opportunity can be further improved by increasing the parameter λ_2 to 10.0 (third row (b)), and

[11] https://github.com/predictive-analytics-lab/Data-Domain-Fairness

[12] We deliberately evaluate the performance (accuracy and fairness) using an auxiliary classifier instead of using the predictor of the transformer network. Since the emphasis of this work is on representation learning, we should not prescribe what classifier the user chooses on top of learned representation.

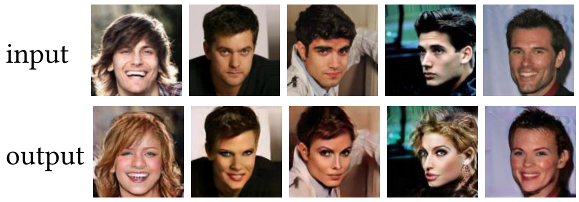


Figure 5.3: Results of our approach (image-to-image translation via attributes). Given N i.i.d. samples $\{(\mathbf{x}^n, y^n)\}_{n=1}^{N}$, our method transforms them into a new fair dataset $\{(\tilde{\mathbf{x}}^n, y^n)\}_{n=1}^{N}$ where $(\tilde{\mathbf{x}}^n, y^n)$ is the fair version of $(\mathbf{x}^n, y^n)$. The synthesized images are produced by the StarGAN model (Choi et al., 2018) conditioned on the original images and their fair attribute representation.

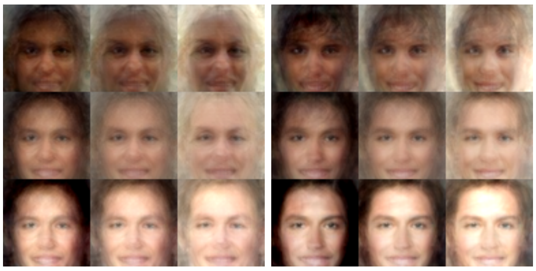


Figure 5.4: Results of Fainess GAN (Sattigeri et al., 2019) (Fig.2) of non-attractive (left) and attractive (right) males after pre-processing. Given N i.i.d. samples $\{(\mathbf{x}^n, y^n)\}_{n=1}^{N}$, Fainess GAN transforms them into a new fair dataset $\{(\tilde{\mathbf{x}}^n, \tilde{y}^n)\}_{n=1}^{N'}$ where $N' \neq N$ and $(\tilde{\mathbf{x}}^n, \tilde{y}^n)$ has no correspondence to $(\mathbf{x}^n, y^n)$.

by using unbiased estimator of HSIC (third row (c)). Looking at the TPR values across protected features (females and males), we can see that the male TPR value has increased, but it has an opposite effect for females. In the CelebA dataset, the proportion of attractive to unattractive males is around 30% to 70%, and it is opposite for females; male group is therefore the minority group in this problem. Our method achieves better equality of opportunity measure than the baseline by increasing the minority group TPR value while decreasing the majority group TPR value. To understand the balancing mechanism of TPR values (*Q3*), we visualize a subset of test male images that have been classified correctly as attractive after transformation (those examples were misclassified in the original domain) in figure 5.2.

We observe a consistent localized area in face, specifically *lips* and *eyes* regions. The CelebA dataset has a large diversity in visual appearance of females and males (hair style, hair color) and their ethnic groups, so more localized facial areas have to be discovered to equalize TPR values across groups. Lips are very often coloured in female (the majority group) celebrity faces, hence our method, to increase the minority group TPR value, colorizes the lip regions of the minority group (males). Interestingly, female faces without prominent lipstick often got this transformation as well, prompting the decrease in the majority group TPR value. Regarding eye regions, several studies (e.g. Brown and Perrett, 1993 and references therein) have shown their importance in

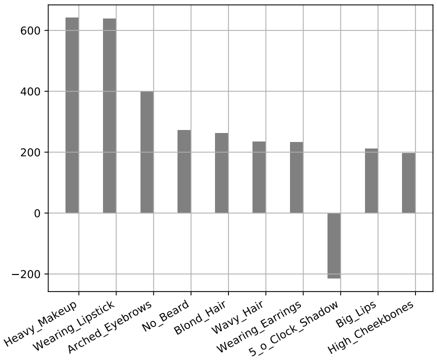


Figure 5.5: Top 10 semantic attribute features that have been changed in 647 males; those males were *incorrectly* predicted as not attractive, but are now correctly predicted as attractive. 641 and 639 males out of 647 are now with 'Heavy_Makeup' and 'Wearing_Lipstick' attributes, respectively, and 215 out of 647 males are now *without* a '5_o_Clock_Shadow' attribute.

gender identification. Also, a heavy makeup that is often applied to female celebrity eyes can also support our visualization in figure 5.2.

The image-to-image translation using transformer network learns to produce coarse-grained changes, i.e. masking/colorizing face regions. This is expected as we learn a highly unconstrained mapping from source to target domain, in which the target data is unavailable. To enable fine-grained changes and semantic transformation of the images, we now explore semantic attributes; attributes are well-established interpretable mid-level representations for images. We show how an attribute-to-attribute translation provides an alternative way in analysing and performing an image-to-image translation.

Attribute-to-attribute translation Images in the CelebA dataset come with 40 dimensional binary attribute annotations. We use all but two attributes (*gender* and *attractiveness*) as semantic attribute representation of images. We then perform attribute-to-attribute translation with the transformer network and consider the same attractive versus not attractive task and gender protected characteristic as with the image data. We report the results of this experiment in table 5.2 (second and forth rows correspond to the domain of attributes). First, we observe that the predictive performance of the classifier trained on attribute representation is only slightly lower than the performance of the classifier trained on the image data (79.1 versus 80.6), which enables sensible comparison of the results in these two settings. Second, we observe better gain in equality of opportunity when using the transformed attribute representation comparing to transformed images (12.4 is the best Eq. Opp. result in this experiment). This comes at the cost of a drop in accuracy performance. The TPR rates for both groups are higher when using translated attribute representation than when using translated image representation (third row versus fourth row). The largest improvement of the TPR is observed in the group of males (from 50.9 in the original attribute to 74.8 in the translated attribute space). Further analysis of changes in attribute representation reveals that equality of opportunity is achieved by putting *lipstick* and *heavy-makeup* to the male group (figure 5.5). These top 2 features have been mostly changed in the group of *males*. Very few changes happened in the group of females. This is encouraging as we have just arrived at the same conclusion (figure 5.2 and figure 5.5), be it using images or using semantic attributes (*Q4*).

Image-to-image translation via attributes Given the remarkable progress that has been made in the field towards image synthesis with the conditional GAN models, we attempt to synthesize images with respect to the attribute description. Specifically, we use the StarGAN model (Choi et al., 2018), the state-of-the-art model for image synthesis with multi-attribute transformation, to synthesize images with our learned fair attribute representation. For this, we pre-train the StarGAN model to perform image transformations with 38 binary attributes (excluding gender and attractive attributes) using training data. We then translate all images in CelebA with respect to their fair attribute representation. We evaluate the performance of this approach and report the results in table 5.2 (last row). We also include the qualitative evaluations of image-to-image translations via attributes in figure 5.3. These visualizations essentially generalize counterfactual explanations in the sense of Wachter et al. (2018) to the image domain. We have just shown the 'closest synthesized world', i.e. the smallest change to the world that can be made to obtain a desirable outcome. Overall, the classifier trained using this fair representation shows similar Eq. Opp. performance and comparable accuracy to the classifier trained on representation learned with the transformer network. However, the TPR rates for both protected groups are higher (last row versus third row), especially in the group of males, when using this representation.

Pre-processing approaches The aim of the pre-processing approaches such as Calmon et al. (2017) and Sattigeri et al. (2019) is to transform the given dataset of N i.i.d. samples $\{(\mathbf{x}^n, y^n)\}_{n=1}^{N}$ into a *new* fair dataset $\{(\tilde{\mathbf{x}}^n, \tilde{y}^n)\}_{n=1}^{N'}$. It is important to note that N' is not necessarily equal to N, and therefore $(\tilde{\mathbf{x}}^n, \tilde{y}^n)$ has no correspondence to $(\mathbf{x}^n, y^n)$. Calmon et al. (2017) has proposed this approach for tabular (discrete) data, while Sattigeri et al. (2019) has explored image data. Here, we offer a unified framework for tabular (continuous and discrete) and image data that transforms the given dataset $\{(\mathbf{x}^n, y^n)\}_{n=1}^{N}$ into a new fair dataset $\{(\tilde{\mathbf{x}}^n, y^n)\}_{n=1}^{N}$ where $(\tilde{\mathbf{x}}^n, y^n)$ is the fair version of $(\mathbf{x}^n, y^n)$. *What is the advantage of creating a fair representation per sample (our method) rather than on the whole dataset at once* (Calmon et al., 2017; Sattigeri et al., 2019)? The first can be used to provide an *individual*-level explanation of fair systems, while the latter can only be used to provide a *system*-level explanation. For comparison, we include here a snapshot of results presented in Sattigeri et al. (2019) using the CelebA dataset in figure 5.4. The figure shows eigenfaces/eigensketches with *the mean image* of the new fair dataset $\{(\tilde{\mathbf{x}}^n)\}_{n=1}^{N'}$ (in the center) of the 3×3 grid. No per sample visualisation ($\tilde{\mathbf{x}}^n$) was provided. Left/right/top/bottom images in figure 5.4 show variations along the first/second principal components. In contrast, figure 5.3 shows a per sample visualisation ($\tilde{\mathbf{x}}^n$) using our proposed method.

5.4.3 *The Diversity in Faces dataset*

We extract and align face crops from the images and use 128x128 facial images as the inputs. Our preliminary experiment has similar setup to the image-to-image translation on the CelebA dataset except that the prediction task has seven age groups to be classified. As the fairness criterion we enforce equality of opportunity considering the middle age group (31-45) to be desirable (as the positive label when conditioning). As before, to test the hypothesis that we have learned a fairer image representation, we compare the performance and fairness of the SVM classifier trained using original images and the translated fair images (with features as activations in the pool_5 layer of the VGG19 network). We achieve 52.85 as the overall classification accuracy over seven age groups when using original image features and an increased 60.26 accuracy when using translated images. The equality of opportunity improved from 27.21 using original image representation to 9.85 using fair image representation. Similarly to the CelebA dataset, the image-to-image translation using transformer network learns to produce coarse-grained changes, i.e. masking/colorizing nose regions (as opposed to lips and eyes regions on CelebA). These preliminary results are encouraging and further analysis will be addressed as a future extension.

5.5 Discussion and Conclusion

> *It is not clear if fairness and interpretability are conflicting requirements.*
>
> Reviewer #1

They are not, however interpretability in how fairness is enforced has so far been overlooked despite being an integral ingredient for encouraging productive public debates regarding fair machine learning systems. Interpretability in machine learning models can help to ascertain qualitatively *whether* fairness is met (Doshi-Velez and Kim, 2017; Working Group, 2017). This paper takes a step further and advocates interpretability to ascertain qualitatively *how* fairness is met, once we have agreed to enforce fairness (e.g. equality of opportunity) in machine learning models. We specifically focus on enforcing fairness in representation learning. Unlike other fair representation learning methods that learn *latent* embeddings, our method learns a representation that is in the same space as the original input data, therefore retaining the semantics of the input domain. Our method picks up consistently in 10 *out of* 10 *repeated experiments* whether a person is a husband or wife as a direct proxy for *gender*, and subsequently reduces the wife category which is associated with a negative prediction. In our experiments with people's faces, eyes and lips are considered to be the direct proxy for gender attractiveness, and nose regions for being in a certain age group. As a potential future direction, we plan to further analyze the interpretability in fairness using causal reasoning (Lopez-Paz et al., 2017).

5.6 Addendum

The prior text of this chapter was published at Conference on Computer Vision and Pattern Recognition (CVPR), 2019. The following sections within this chapter comprise advances on the method that have been made, but are yet to be published.

5.6.1 *Introduction*

Post-publication of 'Discovering Fair Representations in the Data Domain' there were a number of directions that could have been explored. The purpose of this addendum is to investigate three additions with respect to image inputs: 1) Making the model more general, 2) Improving the qualitative results, and 3) Evaluating the performance on an additional dataset.

5.6.2 *Making the model more general*

The work presented in this paper introduced a method at the intersection of fairness and interpretability, using feature translation to produce fair representations. While all variations of our proposed approach improved the fairness of the final outcome, the qualitative aspect varied.

When operating on tabular data, the translations consistently produced qualitatively plausible results (see section 5.4). A potential reason for this is that the more constrained space of tabular features (e.g., categorical features) benefits our proposed method. Results attained when operating directly on the less constrained domain of images were less plausible (see figure 5.2). To overcome this, a two-step procedure was followed, referred to as ‘image-to-image translation via attributes’ in section 5.4.2. This two-step procedure bore excellent results, both quantitatively (table 5.2) and qualitatively (figure 5.3). The aim of this addendum is to improve the robustness of the direct image-to-image translation results, so that this two-step procedure of translation via attributes is not needed for results that are visually appealing. To address this, in the following section, each of the component terms of the loss function shown in equation (5.3) is examined.

5.6.2.1 *Prediction Loss Term*

The original loss term of the paper included a ‘prediction loss’ element.

$$\sum_{n=1}^{N} \mathscr{L}(T_{\omega}(\mathbf{x}^n), y^n) \tag{5.4}$$

The presence of this term aligns our work with the previously proposed fair representation methods of Beutel et al. (2017) and Madras et al. (2018a), amongst others. In these works, the model operates across two spaces, producing both a feature embedding representation *and* a target prediction in a single step forward pass.

$$\overbrace{X \to Z}^{\text{embed}}\mathllap{\underbrace{}_{}}\!\!\!\!\underbrace{\to Y}_{\text{task}}$$

Recent works, however, have questioned if this task-prediction term is necessary. Works such as Madras et al. (2018a) and Oneto et al. (2020b) have argued that the ‘embedding’ stage of the model could be produced in isolation from the predictive stage, which can be trained after the fair representation has been produced. The argument is that once trained, the representation can be used for a multitude of tasks, allowing fair transfer learning. In addition, McNamara et al. (2019) and Oneto et al. (2020a) show that the task can be removed during training, with utility unrelated to a protected characteristic retained, although both are in the context of latent embedding. Intuitively, this aligns with our aim of ‘styling’ an image to be fairer. The result *should* remain close to the original input. The intention is only for features correlated with a protected characteristic to be obfuscated. To re-align our work with these recent proposals, the proposal in this addendum is that the prediction loss term can removed. The further benefit of this approach is that the fair representation model can be trained in an *unsupervised* manner.

5.6.2.2 *Reconstruction Loss Term*

Continuing through the components of the original training objective, we reach the reconstruction loss term.

$$\sum_{n=1}^{N} \|\mathbf{x}^n - T_\omega(\mathbf{x}^n)\|_2^2 \tag{5.5}$$

This term, along with the decomposition loss component, introduces the main tension in what we are trying to achieve. Our aim is for the translated image to be visually close to the input image, crucially, however, the translated images should also be sufficiently different that the protected characteristic is obfuscated. The reconstruction term penalises pixel values that are far from the original, whereas the decomposition loss may require that they are. If we take 'gender' as an example of an attribute that we want to become less clear in our translation, then this is a challenge when features, such as hair style, facial hair, or make-up, that are correlated with a protected characteristic, but also represent a large area of the pixel-space need to be changed. The decomposition loss may be trying to remove, add, or alter one of these features. Here, we would like the pixel values to change, in contradiction to the reconstruction loss term. One potential solution would be to remove the reconstruction loss term altogether, but this poses a problem – we want the result to retain meaning in pixel space. If the term were simply removed, then the representation would be able to collapse to a trivial solution, such as producing an image of uniform colour – this would satisfy the objective of being unable to retain the protected characteristic, but would no longer retain information useful for a downstream task. Instead, we propose two modifications. Firstly, extending the reconstruction loss to also evaluate the difference in the residual space $\phi(x)$ between the original image and the fair translation. This allows for similarity in more complex features, with less emphasis on particular values. Therefore, we retain the reconstruction loss, but include it as an average of a number of reconstruction losses. Secondly, we use a smooth-L_1 loss (Girshick, 2015).

$$\text{smooth}_{L_1}(x, y) = \begin{cases} 0.5(x-y)^2, & \text{if } \|x-y\| < 1 \\ \|x-y\| - 0.5, & \text{otherwise} \end{cases}$$

This is a relaxation of the $L1$ loss term where the loss function is replaced with a quadratic function if x and y are sufficiently close. We use this, to allow for greater freedom in the reconstructions. This results in the modified reconstruction loss, where M is the set of output layers from the feature extractor network (VGG in the main text), and SL_1 is the smooth-L_1 loss:

$$\frac{1}{N}\sum_{n=1}^{N}\Big(\frac{1}{M+1}\sum_{m\in M} SL_1(\mathbf{x}^n, T_\omega(\mathbf{x}^n)) + \frac{1}{M+1} SL_1(\phi_m(\mathbf{x}^n), \phi_m(T_\omega(\mathbf{x}^n)))\Big) \tag{5.6}$$

5.6.2.3 *Decomposition Loss Term*

The decomposition loss term used throughout is based on Hilbert-Schmidt Independence Criteria (HSIC) and is a powerful statistics-based dependency measure.

$$-\operatorname{HSIC}(R, S|Y = +1) + \operatorname{HSIC}(P, S|Y = +1) \tag{5.7}$$

Where S is the set of protected attributes present in the dataset, P is the set of fair transformed features, and R is the set of residuals between the features and the fair transformation. The aim of the decomposition loss is to encourage the input to be represented as a 'fair' and 'unfair' component. The fair component is then additionally required to be close to the original input.

The challenge, as with all kernel-based approaches, is the choice of hyperparameters to configure the kernel. In the case of HSIC, we use an Radial Basis Function (RBF) kernel which accepts a σ-term to control the bandwidth of the kernel. To make a reasonable attempt to ensure that our model doesn't learn to operate outside of the scope of the dependency metric bandwidth, we extend our existing approach and evaluate at a range of sigma values. For ease of notation, we use Φ to denote the set of σ values which we use to parametrise the RBF kernel used in HSIC. Letting $\Phi = \{0.5, 1.0, 2.0, 5.0, 10.0, 20.0\}$, we obtain:

$$\sum_{\sigma\in\Phi}(-\operatorname{HSIC}_\sigma(R, S|Y = +1) + \operatorname{HSIC}_\sigma(P, S|Y = +1)) \tag{5.8}$$

This results in the final unsupervised objective term equation (5.9).

$$\underset{T_\omega}{\text{minimize}} \underbrace{\frac{1}{N}\sum_{n=1}^{N}\Big(\frac{1}{N}\sum_{n=1}^{N}\Big(\frac{1}{M+1}\sum_{m\in M} SL_1(\mathbf{x}^n, T_\omega(\mathbf{x}^n)) + \frac{1}{M+1} SL_1(\phi_m(\mathbf{x}^n), \phi_m(T_\omega(\mathbf{x}^n)))\Big)\Big)}_{\text{reconstruction loss}} + \\ + \lambda_1 \left(\underbrace{\sum_{\sigma\in\Phi}(-\text{HSIC}_\sigma(R, S|Y = +1) + \text{HSIC}_\sigma(P, S|Y = +1))}_{\text{decomposition loss}} \right) \tag{5.9}$$

5.6.3 *Additional Experiments*

To evaluate the proposed adaptations, we reuse the CelebA dataset from the main text, but treat the feature attribute ‘smiling’, as the target label while promoting equality of opportunity, as opposed to the feature ‘attractive’ which was used in the main paper. This change is to bring the work into line with subsequent works that have used this dataset in an algorithmic fairness context such as Denton et al. (2019). In addition, we evaluate on the Colorised MNIST (cMNIST) dataset, which is fully introduced in chapter 6. Here we use a variation with three colours, and all ten outcome classes. The aim of the experiments is to demonstrate that a more visually pleasing image can be produced with the new objective.

5.6.3.1 *Qualitative Evaluation and Discussion*

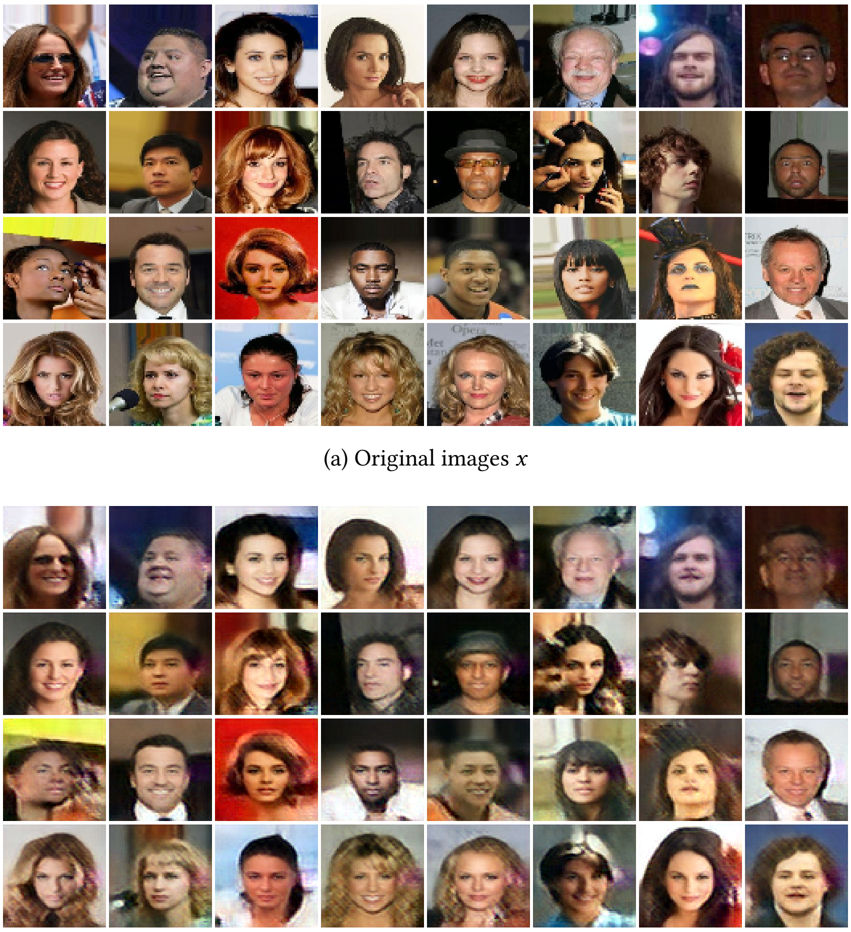
The main purpose of this evaluation is to demonstrate that more plausible transformations can be produced without the need to translate via attributes. Sample images from the main paper can be seen in figure 5.2 and results from the suggested modifications are shown in figure 5.6. In terms of the image quality, the results are certainly more plausible. Interestingly, the regions of change are most associated with facial hair and make-up, which focuses on the eyes and mouth regions. This is consistent with our previous results. One thing to consider is that these are all features that occupy only a few pixels. Instinctively, features such as ‘baldness’, or ‘hair length’ should also be associated with the protected characteristic, ‘gender’, but these are not as affected. One possibility for this is that these features occupy a large number of pixels and that changing them would be too costly in terms of reconstruction error. Ultimately, regardless of the actual reason, because the fair representation resides in the data domain, we are able to speculate as to what may be the cause. If the learned representation remained in a latent space, it would be

Table 5.3: Results on cMNIST dataset.

	domain $\mathcal{X}$.	Digit Acc. ↑	Colour Acc. → 33.33 ←
1: orig. $\mathbf{x}$	*images*	97.917	100.0
2: fair $\tilde{\mathbf{x}}$	*images*	87.5	18.75

very difficult to understand what features had been addressed and the limitations that these may present.

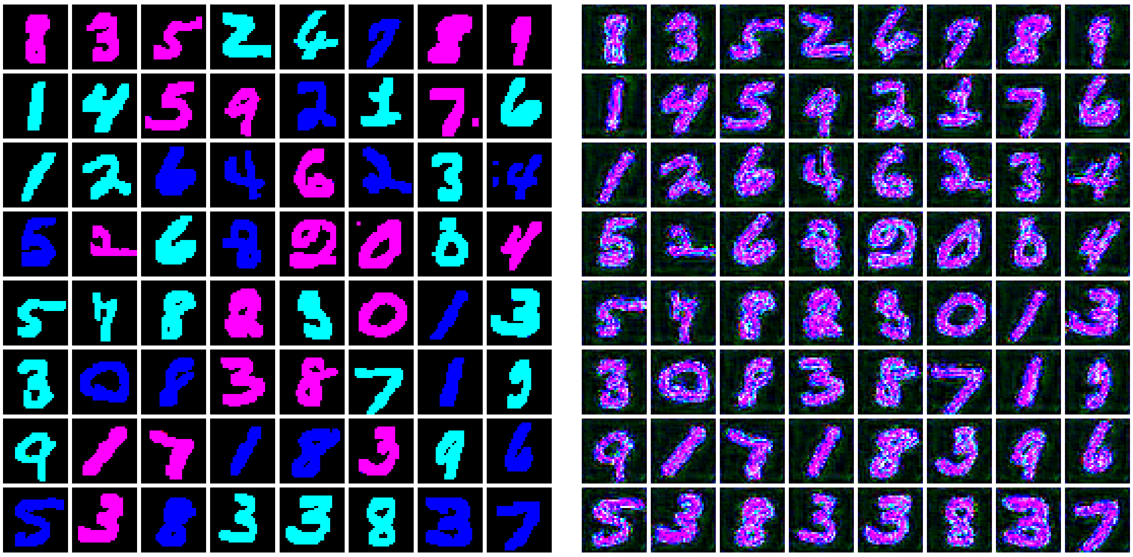
For the CMNIST dataset, the results are encouraging. The model produces a representation that is invariant to colour, while retaining much of the information relevant to digit classification. Although the images are a little *noisy*, the predominant transform is clear. To clarify this further, we train two additional CNN models to predict the digit class and the colour of the transformed images. Results are shown in table 5.3.



(a) Original images x

(b) Translated output $T_\omega(x)$

Figure 5.6: A comparison of the translated images produced with our amendments alongside the original images from the CelebA dataset. *Left*: A random selection of original images from a withheld evaluation set. *Right*: Translated versions of the original input which are now more plausible. The model was trained for Equality of Opportunity (EOpp), so the invariance measure was conditioned on images where the celebrity is smiling, with gender as the protected characteristic. While some images display artifacts, consistently across samples, lipstick and eye makeup have been removed and facial hair reduced. This aligns with our findings in the main text.



(a) Original images x

(b) Amended output $T_{\omega}(x)$

Figure 5.7: A comparison of the translated images produced with presented amendments alongside the original images in the cMNIST dataset. **Left**: A random selection of original images from a withheld evaluation set. **Right**: Translated versions of the original input which show that unique colours have been removed. The model was trained for Demographic Parity (DP), so the invariance measure was provided with all samples. The target is digit value, and the protected attribute to be invariant to is colour.

6 PAPER 3: NULL-SAMPLING FOR INTERPRETABLE AND FAIR REPRESENTATIONS

AUTHORS:

Thomas Kehrenberg, Myles Bartlett, Oliver Thomas and Novi Quadrianto

AFFILIATIONS:

Predictive Analytics Lab (PAL), University of Sussex, Brighton, UK

CONFERENCE: *European Conference on Computer Vision* (ECCV), 2020

DOI: `10.1007/978-3-030-58574-7_34`

NOTE: The appendix has been included as section 6.7.

6.1 ABSTRACT

We propose to learn invariant representations, in the data domain, to achieve interpretability in algorithmic fairness. Invariance implies a selectivity for high level, relevant correlations w.r.t. class label annotations, and a robustness to irrelevant correlations with protected characteristics such as race or gender. We introduce a non-trivial setup in which the training set exhibits a strong bias such that class label annotations are irrelevant and spurious correlations cannot be distinguished. To address this problem, we introduce an adversarially trained model with a *null-sampling* procedure to produce invariant representations in the data domain. To enable disentanglement, a partially-labelled *representative* set is used. By placing the representations into the data domain, the changes made by the model are easily examinable by human auditors. We show the effectiveness of our method on both image and tabular datasets: Coloured MNIST, the CelebA and the Adult dataset.

6.2 Introduction

Without due consideration for the data collection process, machine learning algorithms can exacerbate biases, or even introduce new ones if proper control is not exerted over their learning (Holstein et al., 2019). While most of these issues can be solved by controlling and curating data collection in a fairness-conscious fashion, doing so is not always an option, such as when working with historical data. Efforts to address this problem algorithmically have been centred on developing statistical definitions of fairness and learning models that satisfy these definitions. One popular definition of fairness used to guide the training of fair classifiers, for example, is *demographic parity*, stating that positive outcome rates should be equalised (or *invariant*) across protected groups.

In the typical setup, we have an input $\boldsymbol{x}$, a sensitive attribute s that represents some non-admissible information like gender and a class label y which is the prediction target. The idea of fair *representation* learning (Zemel et al., 2013; Edwards and Storkey, 2016; Madras et al., 2018a) is then to transform the input $\boldsymbol{x}$ to a representation $\boldsymbol{z}$ which is invariant to s. Thus, learning from $\boldsymbol{z}$ will not introduce a forbidden dependence on s. A good fair representation is one that preserves most of the information from $\boldsymbol{x}$ while satisfying the aforementioned constraints.

As unlabelled data is much more freely available than labelled data, it is of interest to learn the representation in an unsupervised manner. This will allow us to draw on a much more diverse pool of data to learn from. While annotations for y are often hard to come by (and often noisy; see Kehrenberg et al., 2020a), annotations for the sensitive attribute s are usually less so, as s can often be obtained from demographic information provided by census data. We thus consider the setting where the representation is learned from data that is only labelled with s and not y. This is in contrast to most other representation learning methods. We call the set used to learn the representation the *representative* set, because its distribution is meant to match the distribution of the deployment setting (and is thus representative).

Once we have learnt the mapping from $\boldsymbol{x}$ to $\boldsymbol{z}$, we can transform the *training* set which, in contrast to the representative set, has the y labels (and s labels). In order to make our method more widely applicable, we consider an *aggravated fairness problem* in which the training set contains a strong spurious correlation between s and y, which makes it impossible to learn from it a representation which is invariant to s but not invariant to y. Non-invariance to y is important in order to be able to predict y. The training set thus does *not* match the deployment setting, thereby rendering the representative set essential for learning the right invariance. From hereon, we will use the terms *spurious* and *sensitive* interchangeably, depending on the context, to refer to an attribute of the data we seek invariance to. We can draw a connection between learning in the presence of spurious correlations and what Kallus and Zhou (2018) call *residual unfairness*. Consider the Stop, Question and Frisk (SQF) dataset for example: the data was collected in New York City, but the demographics of the recorded cases do not represent the true demographics of NYC well. The demographic attributes of the recorded individuals might correlate so strongly with the prediction target that the two are nearly indistinguishable. This is the scenario that we are investigating: s and y are so closely correlated in the labelled dataset that they cannot be distinguished, but the learning of s is favoured due to being the 'path of least resistance'. The deployment setting (i.e. the test set) does not possess this strong correlation and thus a naïve approach will lead to very unfair predictions. In this case, a disentangled representation is insufficient; the representation needs to be explicitly invariant solely with respect to s. In our approach, we make use of the (partially labelled) representative set to learn this invariant representation.

While there is a substantial body of literature devoted to the problems of fair representation-learning, exactly how the invariance in question is achieved is often overlooked. When critical decisions, such as who should receive bail or be released from jail, are being deferred to an automated decision making system, it is critical that people be able to trust the logic of the model underlying it, whether it be via semantic or visual explanations. We build on the work of Quadrianto et al. (2019) and learn a decomposition ($f^{-1}: Z_s \times Z_{\neg s} \rightarrow X$) of the *data domain* ($X$) into independent subspaces *invariant* to s ($Z_{\neg s}$) and *indicative* of s (Z_s), which lends an interpretability that is absent from most representation-learning methods. While model interpretability has no strict definition (Zhang and Zhu, 2018), we follow the intuition of Adel et al. (2018) – *a simple relationship to something we can understand*, a definition which representations in the data domain naturally fulfil.

Whether as a result of the aforementioned sampling bias or simply because the features necessarily co-occur, it is not rare for features to correlate with one another in real-world datasets. Lipstick and gender for example, are two attributes that we expect to be highly correlated and to enforce invariance to gender can implicitly enforce invariance to makeup. This is arguably the desired behaviour. However, unforeseen biases in the data may engender cases which are less justifiable. By baking interpretability into our model (by having representations in the data domain), though we still have no better control over what is learned, we can at least diagnose such pathologies.

To render our representations interpretable, we rely on a simple transformation we call *null-sampling* to map invariant representations in the data domain. Previous approaches to fair representation learning (Edwards and Storkey, 2016; Louizos et al., 2016; Beutel et al., 2017; Madras et al., 2018a) predominantly rely upon autoencoder models to jointly minimise reconstruction loss and invariance. We discuss first how this can be done with such a model that we refer to as cVAE (conditional VAE), before arguing that the bijectivity of invertible neural networks (INNs) (Dinh et al., 2014) makes them better suited to this task. We refer to the variant of our method based on these as cFlow (conditional Flow). INNs have several properties that make them appealing for unsupervised representation learning. The focus of our approach is on creating invariant representations that preserve the non-sensitive information maximally, with only knowledge of s and not of the target y, while at the same time having the ability to easily probe what has been learnt.

Our contribution is thus two-fold: 1) We propose a simple approach to generating representations that are invariant to a feature s, while having the benefit of interpretability that comes with being in the data domain. We call our model *NIFR* (*N*ull-sampling for *I*nterpretable and *F*air *R*epresentations). 2) We explore a setting where the labelled training set suffers from varying levels of sampling bias, demonstrating an approach based on transferring information from a more diverse representative set, with guarantees of the non-spurious information being preserved.

6.3 Background

6.3.1 *Learning fair representations.*

Given a sensitive attribute s (for example, gender or race) and inputs $\boldsymbol{x}$, a fair representation $\boldsymbol{z}$ of $\boldsymbol{x}$ is then one for which $\boldsymbol{z} \perp s$ holds, while ideally also being predictive of the class label y. Zemel et al. (2013) was the first to propose the learning of fair representations which allow for transfer to new classification tasks. More recent methods are often based on variational autoencoders (VAEs) (Kingma and Welling, 2014; Edwards and Storkey, 2016; Louizos et al., 2016; Beutel et al., 2017). The achieved fairness of the representation can be measured with various fairness metrics. These measure, however, usually how fair the predictions of a classifier are and not how fair a representation is.

The appropriate measure of fairness for a given task is domain-specific (Liu et al., 2019) and there is often not a universally accepted measure. However, *Demographic Parity* is the most widely used (Edwards and Storkey, 2016; Louizos et al., 2016; Beutel et al., 2017). Demographic Parity demands $\hat{y} \perp s$ where $\hat{y}$ refers to the predictions of the classifier. In the context of fair representations, we measure the Demographic Parity of a downstream classifier, $f(\cdot)$, which is trained on the representation $\boldsymbol{z}$, i.e. $f\colon\ Z \rightarrow \hat{Y}$.

A core principle of all fairness methods is the *accuracy-fairness trade-off.* As previously stated, the fair representation should be invariant to s (→ fairness) but still be predictive of y (→ accuracy). These desiderata cannot, in general, be simultaneously satisfied if s and y are correlated.

The majority of existing methods for fair representations also make use of y labels during training, in order to ensure that $\boldsymbol{z}$ remains predictive of y. This aspect can, in theory, be removed from the methods, but then there is no guarantee that information about y is preserved (Louizos et al., 2016).

6.3.2 *Learning fair, transferrable representations*

In addition to producing fair representations, Madras et al. (2018a) want to ensure the representations are transferrable. Here, an adversary is used to remove sensitive information from a representation $\boldsymbol{z}$. Auxiliary prediction and reconstruction networks, to predict class label y and reconstruct the input $\boldsymbol{x}$ respectively, are trained on top of $\boldsymbol{z}$, with s being ancillary input to the reconstruction.

Also related is Creager et al. (2019) who employ a FactorVAE (Kim and Mnih, 2018) regularised for fairness. The idea is to learn a representation that is both disentangled and invariant to multiple sensitive attributes. This factorisation makes the latent space easily manipulable such that the different subspaces can be freely removed and composed at test time. Zeroing out the dimensions or replacing them with independent noise imparts invariance to the corresponding sensitive attribute. This method closely resembles ours when we use an invertible encoder. However, the emphasis of our approach is on interpretability, information-preservation, and coping with sampling bias - especially extreme cases where $|\mathrm{supp}(S_{tr} \times Y_{tr})| < |\mathrm{supp}(S_{te} \times Y_{te})|$.

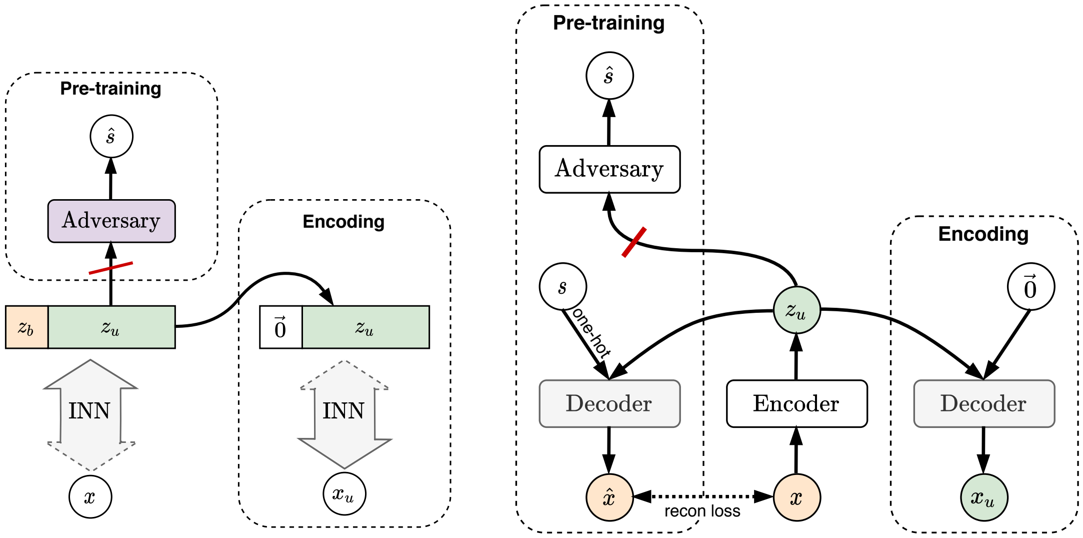
Attempts were made by Quadrianto et al. (2019) prior to this work to learn fair representations in the data domain in order to make it interpretable and transferable. In their work, the input is assumed to be additively decomposable in the feature space into a *fair* and *unfair* component, which together can be used by the decoder to recover the original input. This allows us to examine representations in a human-interpretable space and confirm that the model is not learning a relationship reliant on a sensitive attribute. Though a first step in this direction, we believe such a linear decomposition is not sufficiently expressive to fully capture the relationship between the sensitive and non-sensitive attributes. Our approach allows for the modelling of more complex relationships.

6.3.3 *Learning in the presence of spurious correlations*

Strong spurious correlations make the task of learning a robust classifier challenging: the classifier may learn to exploit correlations unrelated to the true causal relationship between the features and label, and thereby fail to generalise to novel settings. This problem was recently tackled by Kim et al. (2019) who apply a penalty based on the mutual information between the feature embedding and the spurious variable. While the method is effective under mild biasing, we show experimentally that it is not robust to the range of settings we consider.

Jacobsen et al. (2019) explore the vulnerability of traditional neural networks to spurious variables – e.g., textures, in the case of ImageNet (Geirhos et al., 2019) – and propose a INN-based solution akin to ours. The INN's encoding is split such that one partition, $\boldsymbol{z}_b$ is encouraged to be predictive of the spurious variable while the other serves as the logits for classification of the semantic label. Information related to the nuisance variable is 'pulled out' of the logits as a result of maximising $\log p(s|\boldsymbol{z}_n)$. This specific approach, however, is incompatible with the settings we consider, due to its requirement that both s and y be available at training time.

Viewing the problem from a causal perspective, Arjovsky et al. (2019) develop a variant of empirical risk minimisation called invariant risk minimisation (IRM). The goal of IRM is to train a predictor that generalises across a large set of unseen environments; because variables with spurious correlations do not represent a stable causal mechanism, the predictor learns to be invariant to them. IRM assumes that the training data is not *iid* but is partitioned into distinct environments, $e \in E$. The optimal predictor is then defined as the minimiser of the sum of the empirical risk R_e over this set. In contrast, we assume possession of only a single source of *labelled*,



(a) cFlow model.

(b) cVAE model.

Figure 6.1: Training procedure for our models. $\boldsymbol{x}$: input, s: sensitive attribute, $\boldsymbol{z}_u$: de-biased representation, $\boldsymbol{x}_u$: de-biased version of the input in the data domain. The red bar indicates a gradient reversal layer, and $\mathbf{0}$ the null-sampling operation.

albeit spuriously-correlated, data, but that we have a second source of data that is free of spurious correlations, with the benefit being that it only needs to be labelled *with respect to s*.

6.4 Interpretable Invariances by Null-Sampling

6.4.1 *Problem Statement*

We assume we are given inputs $\boldsymbol{x} \in \mathcal{X}$ and corresponding labels $y \in \mathcal{Y}$. Furthermore, there is some spurious variable $s \in \mathcal{S}$ associated with each input $\boldsymbol{x}$ which we do *not* want to predict. Let X, S and Y be random variables that take on the values $\boldsymbol{x}$, s and y, respectively. The fact that both y and s are predictive of $\boldsymbol{x}$ implies that $I(X;Y), I(X;S) > 0$, where $I(\cdot;\cdot)$ is the mutual information. Note, however, that the conditional entropy is non-zero: $H(S|X) \neq 0$, i.e., S is not completely determined by X.

The difficulty of this setup emerges in the training set: there is a close correspondence between S and Y, such that for a model that sees the data through the lens of the loss function, the two are indistinguishable. Furthermore, we assume that this is *not* the case in the test set, meaning the model cannot rely on shortcuts provided by S if it is to generalise from the training set.

We call this scenario where we only have access to the labels of a biasedly-sampled subpopulation an *aggravated fairness problem*. These are not uncommon in the real-world. For instance, in long-feedback systems such as mortgage-approval where the demographics of the subpopulation with observed outcomes is *not* representative of the subpopulation on which the model has been deployed. In this case, s has the potential to act as a false (or *spurious*) indicator of the class label and training a model with such a dataset would limit generalisability. Let (X^{tr}, S^{tr}, Y^{tr}) then be the random variables sampled for the training set and (X^{te}, S^{te}, Y^{te}) be the random variables for the test set. The training and test sets thus induce the following inequality for their mutual information: $I(S^{tr}; Y^{tr}) \gg I(S^{te}; Y^{te}) \approx 0$.

Our goal is to learn a representation $\boldsymbol{z}_u$ that is independent of s and transferable between downstream tasks. Complementary to $\boldsymbol{z}_u$, we refer to some abstract component of the model that absorbs the unwanted information related to s as $\mathscr{B}$, the realisation of which we define with respect to each of the two models to be described. The requirement for $\boldsymbol{z}_u$ can be expressed via mutual information:

$$I(\boldsymbol{z}_u; s) \overset{!}{=} 0 \; . \tag{6.1}$$

However, for the representation to be useful, we need to capture as much relevant information in the data as possible. Thus, the combined objective function:

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}\sim X}[-\log p_\theta(\boldsymbol{x})] + \lambda I(f_\theta(\boldsymbol{x}); s) \tag{6.2}$$

where θ refers to the trainable parameters of our model f_θ and $p_\theta(\boldsymbol{x})$ is the likelihood it assigns to the data.

We optimise this loss in an adversarial fashion by playing a min-max game, in which our encoder acts as the generative component. The adversary is an auxiliary classifier g, which receives $\boldsymbol{z}_u$ as input and attempts to predict the spurious variable s. We denote the parameters of the adversary as ϕ; for the parameters of the encoder we use θ, as before. The objective from equation 6.2 is then

$$\min_{\theta\in\Theta} \max_{\phi\in\Phi} \mathbb{E}_{\boldsymbol{x}\sim X}[\log p_\theta(\boldsymbol{x}) - \lambda \mathscr{L}_c(g_\phi(f_\theta(\boldsymbol{x})); s)] \tag{6.3}$$

where $\mathscr{L}_c$ is the cross-entropy between the predictions for s and the provided labels. In practice, this adversarial term is realised with a gradient reversal layer (GRL) (Ganin et al., 2016) between $\boldsymbol{z}_u$ and g as is common in adversarial approaches (Edwards and Storkey, 2016).

6.4.2 *The Disentanglement Dilemma*

The objective in equation 6.3 balances the two desiderata: predicting y and being invariant to s. However, in the training set (X^{tr}, S^{tr}, Y^{tr}), y and s are so strongly correlated that removing information about s inevitably removes information about y. This strong correlation makes existing methods fail under this setting. In order to even define the right learning goal, we require another source of information that allows us to disentangle s and y. For this, we assume the existence of another set of samples that follow a similar distribution to the test set, but whilst the sensitive attribute is available, the class labels are not. In reality, this is not an unreasonable assumption, as, while properly annotated data is scarce, unlabelled data can be obtained in abundance (with demographic information from census data, electoral rolls, etc.). Previous work has also considered treated 'unlabelled data' as still having s labels (Wick et al., 2019). We are restricted only in the sense that the spurious correlations we want to sever are indicated in the features. We call this the *representative set*, consisting of X^{rep} and S^{rep}. It fulfils $I(S^{rep}; Y^{rep}) \approx 0$ (or rather, it would, if the class labels Y^{rep} were available).

We now summarise the training procedure; an outline for the invertible network model (cFlow) can be seen in figure 6.1a. First, the encoder network f is trained on (X^{rep}, S^{rep}), during the first phase. The trained network is then used to encode the training set, taking in $\boldsymbol{x}$ and producing the representation, $\boldsymbol{z}_u$, decorrelated from the spurious variable. The encoded dataset can then be used to train any off-the-shelf classifier safely, with information about the spurious variable having been absorbed by some auxiliary component $\mathscr{B}$. In the case of the conditional VAE (cVAE) model, $\mathscr{B}$ takes the form of the decoder subnetwork, which reconstructs the data conditional on a one-hot encoding of s, while for the invertible network $\mathscr{B}$ is realised as a partition of the feature map $\boldsymbol{z}$ (such that $\boldsymbol{z} = [\boldsymbol{z}_u, \boldsymbol{z}_b]$), given the bijective constraint. Thus, the classifier cannot take the shortcut of learning s and instead must learn how to predict y directly. Obtaining the s-invariant representations, $\boldsymbol{x}_u$, in the data domain is simply a matter of replacing the $\mathscr{B}$ component of the decoder's input for the cVAE, and $\boldsymbol{z}_b$ for cFlow, with a zero vector of equivalent size. We refer to this procedure used to generate $\boldsymbol{x}_u$ as *null-sampling* (here, with respect to $\boldsymbol{z}_b$).

Null-sampling resembles the *annihilation* operation described in Xiao et al. (2018), however we note that the two serve very different roles. Whereas the annihilation operation serves as a regulariser to prevent trivial solutions (similar to Jaiswal et al., 2018a), null-sampling is used to generate the invariant representations post-training.

6.4.3 *Conditional Decoding*

We first describe a VAE-based model similar to that proposed in Madras et al. (2018a), before highlighting some of its shortcomings that motivate the choice of an invertible representation learner.

The model takes the form of a class conditional β-VAE (Higgins et al., 2017), in which the decoder is conditioned on the spurious attribute. We use $\theta_{enc}, \theta_{dec} \in \theta$ to denote the parameters of the encoder and decoder sub-networks, respectively. Concretely, the encoder component performs the mapping $\boldsymbol{x} \to \boldsymbol{z}_u$, while $\mathscr{B}$ is instantiated as the decoder, $\mathscr{B} := p_{\theta_{dec}}(\boldsymbol{x}|\boldsymbol{z}_u, s)$, which takes in a concatenation of the learned non-spurious latent vector $\boldsymbol{z}_u$ and a one-hot encoding of the spurious label s to produce a reconstruction of the input $\hat{\boldsymbol{x}}$. Conditioning on a one-hot encoding of s, rather than a single value, as done in Madras et al. (2018a) is the key to visualising invariant representations in the data domain. If $I(\boldsymbol{z}_u; s)$ is properly minimised, the decoder can only derive its information about s from the label, thereby freeing up $\boldsymbol{z}_u$ from encoding the unwanted information while still allowing for reconstruction of the input. Thus, by feeding a zero-vector to the decoder we achieve $\hat{\boldsymbol{x}} \perp s$. The full learning objective for the cVAE is given as

$$\begin{aligned}\mathscr{L}_{\text{cVAE}} =& \mathbb{E}_{q_{\theta_{enc}}(\boldsymbol{z}_u, b|\boldsymbol{x})}[\log p_{\theta_{dec}}(\boldsymbol{x}|\boldsymbol{z}, b) - \log p_{\theta_{dec}}(s|\boldsymbol{z}_u)] \\ & - \beta D_{KL}(q_{\theta_{enc}}(\boldsymbol{z}_u|\boldsymbol{x}) \| p(\boldsymbol{z}_u))\end{aligned} \tag{6.4}$$

where β is a hyperparameter that determines the trade-off between reconstruction accuracy and independence constraints, and $p(\boldsymbol{z}_u)$ is the prior imposed on the variational posterior. For all our experiments, $p(\boldsymbol{z}_u)$ is realised as an Isotropic Gaussian. figure 6.1b summarises the procedure as a diagram.

While we show this setup can indeed work for simple problems, as Madras et al. (2018a) before us have, we show that it lacks scalability due to disagreement between the components of the loss. Since information about s is only available to the decoder as a binary encoding, if the relationship between s and $\boldsymbol{x}$ is highly non-linear and cannot be summarised by a simple on/off mechanism, as is the case if s is an attribute such as gender, off-loading information to the decoder by conditioning is no longer possible. As a result, $\boldsymbol{z}_u$ is forced to carry information about s in order to minimise the reconstruction error.

The obvious solution to this is to allow the encoder to store information about s in a partition of the latent space as in Creager et al. (2019). However, we question whether an autoencoder (AE) is the best choice for this setup, with the view that an invertible model is the better tool for the task. Using an invertible model has several guarantees, namely complete information-preservation and freedom from a reconstruction loss, the importance of which we elaborate on below.

6.4.4 *Conditional Flow*

INVERTIBLE NEURAL NETWORKS. Invertible neural networks are a class of neural network architecture characterised by a bijective mapping between their inputs and output (Dinh et al., 2014). The transformations are designed such that their inverses and Jacobians are efficiently computable. These flow-based models permit *exact* likelihood estimation (Rezende and Mohamed, 2015) through the warping of a base density with a series of invertible transformations and computing the resulting, highly multi-modal, but still normalised, density, using the change of variable theorem:

$$\log p(\boldsymbol{x}) = \log p(\boldsymbol{z}) + \sum \log \left| \det \left(\frac{\mathrm{d}h_i}{h_{i-1}} \right) \right|, \quad p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \mathbb{I}) \tag{6.5}$$

where h_i refers to the outputs of the layers of the network and $p(\boldsymbol{z})$ is the base density, specifically an Isotropic Gaussian in our case. Training of the invertible neural network is then reduced to maximising $\log p(\boldsymbol{x})$ over the training set, i.e. maximising the probability the network assigns to samples in the training set.

THE BENEFITS OF BIJECTIVITY. Using an invertible network to generate our encoding, $\boldsymbol{z}_u$, carries a number of advantages over other approaches. Ordinarily, the main benefit of flow-based models is that they permit exact density estimation. However, since we are not interested in sampling from the model's distribution, in our case the likelihood term serves as a regulariser, as it does for Jacobsen et al. (2018). Critically, this forces the mean of each latent dimension to zero enabling null-sampling. The invertible property of the network guarantees the preservation of all information relevant to y which is independent of s, regardless of how it is allocated in the output space. Secondly, we conjecture that the encodings are more robust to out-of-distribution data. Whereas an autoencoder (AE) could map a previously seen input and a previously unseen input to the same representation, an invertible network sidesteps this due to the network's bijective property, ensuring all relevant information is stored somewhere. This opens up the possibility of transfer learning between datasets with a similar manifestation of s, as we demonstrate in section 6.7.8.

Under our framework, the invertible network f maps the inputs $\boldsymbol{x}$ to a representation $\boldsymbol{z}_u$: $f(\boldsymbol{x}) = \boldsymbol{z}$. We interpret the embedding $\boldsymbol{z}$ as being the concatenation of two smaller embeddings: $\boldsymbol{z} = [\boldsymbol{z}_u, \boldsymbol{z}_b]$. The dimensionality of $\boldsymbol{z}_b$, and $\boldsymbol{z}_u$, by complement, is a free parameter (see section 6.7.4 for tuning strategies). As f is invertible, $\boldsymbol{x}$ can be recovered like so:

$$\boldsymbol{x} = f^{-1}([\boldsymbol{z}_u, \boldsymbol{z}_b]) \tag{6.6}$$

where $\boldsymbol{z}_b$ is required for equality of the output dimension and input dimension to satisfy the bijectivity of the network – we cannot output $\boldsymbol{z}_u$ alone, but have to output $\boldsymbol{z}_b$ as well. In order to generate the pre-image of $\boldsymbol{z}_u$, we perform null-sampling with respect to $\boldsymbol{z}_b$ by zeroing-out the elements of $\boldsymbol{z}_b$ (such that $\boldsymbol{x}_u = f^{-1}([\boldsymbol{z}_u, \boldsymbol{0}])$), i.e. setting them to the mean of the prior density, $\mathcal{N}(\boldsymbol{z}; 0, I)$.

How can we be sure that $\boldsymbol{z}_u$ contains enough information about y? The importance of the invertible architecture bears out from this consideration. As long as $\boldsymbol{z}_b$ does not contain the information about y, $\boldsymbol{z}_u$ necessarily must. We can raise or lower the information capacity of $\boldsymbol{z}_b$ by adjusting its size; this should be set to the smallest size sufficient to capture all information about s, so as not to sacrifice class-relevant information. section 6.7.3 explores the effects of the size further.

6.5 Experiments

We present experiments to demonstrate that the null-sampled representations are in fact invariant to s while still allowing a classifier to predict y from them. We run our cVAE and cFlow models on the coloured MNIST (cMNIST) and CelebA dataset, which we artificially bias, first describing the sampling procedure we follow to do so for non-synthetic datasets. As baselines we have the model of Kim et al. (2019) (Ln2L) and the same CNN used to evaluate the cFlow and cVAE models but with the unmodified images as input (CNN). For the cFlow model we adopt a Glow-like architecture (Kingma and Dhariwal, 2018), while both subnetworks of the cVAE model comprise gated convolutions (Oord et al., 2016), where the encoding size is 256. For cMNIST, we construct the Ln2L baseline according to its original description, for CelebA, we treat it as an augmentation

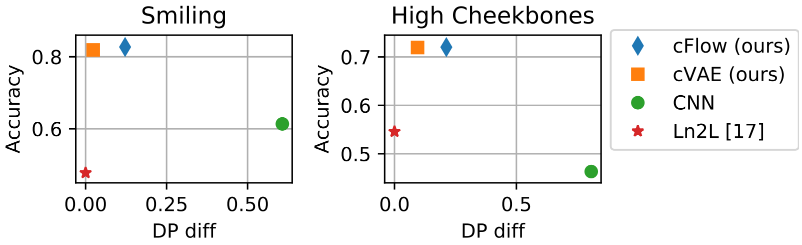


Figure 6.2: Performance of our model for different targets (mixing factor $\eta = 0$). Left: *Smiling* as target, right: *high cheekbones*. *DP diff* measures fairness with respect to demographic parity. A perfectly fair model has a *DP diff* of 0.

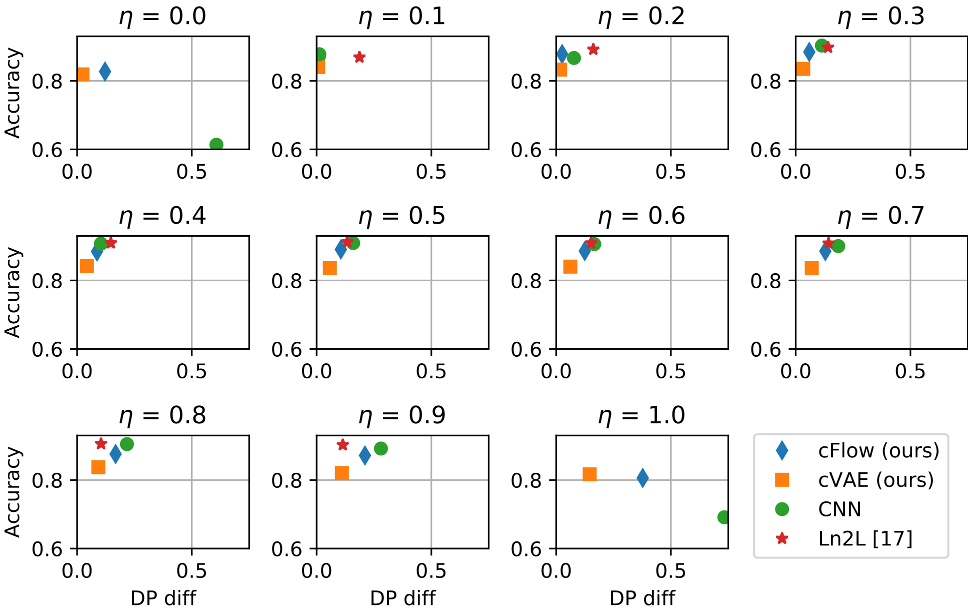


Figure 6.3: Performance of our model for the target 'smiling' for different mixing factors η. *DP diff* measures fairness with respect to demographic parity. A perfectly fair model has a *DP diff* of 0, thus the closer to top-left the better it is in terms of we accuracy-fairness trade-off. Only values $\eta = 0$ and $\eta = 1$ correspond to the scenario of a strongly biased training set. The results for $0.1 \leq \eta \leq 0.9$ are to confirm that our model does not harm performance for non-biased training sets.

of the baseline CNN's objective function. Detailed information regarding model architectures can be found in section 6.7.1 and section 6.7.4.[1]

6.5.1 *Synthesising Dataset Bias*

For our experiments, we require a training set that exhibits a strong spurious correlation, together with a test set that does not. For cMNIST, this is easily satisfied as we have complete control over the data generation process. For CelebA and UCI Adult, on the other hand, we have to generate the split from the existing data. To this end, we first set aside a randomly selected portion of the

[1] The code can be found at `https://github.com/predictive-analytics-lab/nifr`.

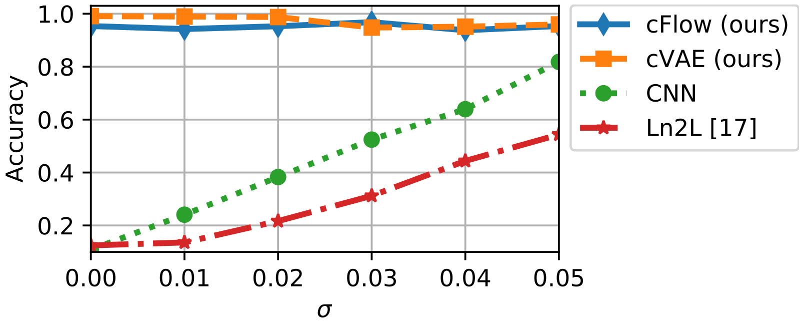


Figure 6.4: Accuracy of our approach in comparison with other baseline models on the cMNIST dataset, for different standard deviations (σ) for the colour sampling.

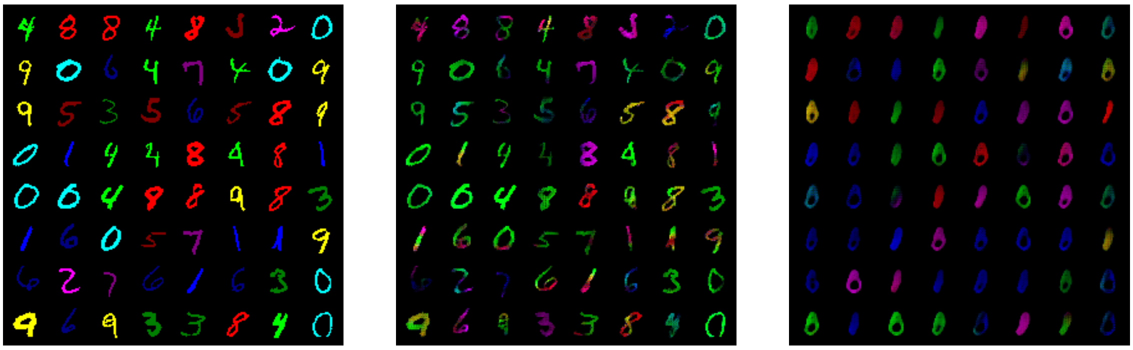
dataset from which to sample the biased dataset The portion itself is then split further into two parts: one in which $(s = -1 \wedge y = -1) \vee (s = +1 \wedge y = +1)$ holds true for all samples, call this part $\mathscr{D}_{eq}$, and the other part, call it $\mathscr{D}_{opp}$, which contains the remaining samples. To investigate the behaviour at different levels of correlation, we mix these two subsets according to a mixing factor η. For $\eta \leq \frac{1}{2}$, we combine (all of) $\mathscr{D}_{eq}$ with a fraction of 2η from $\mathscr{D}_{opp}$. For $\eta > \frac{1}{2}$, we combine (all of) $\mathscr{D}_{opp}$ and a fraction of $2(1-\eta)$ from $\mathscr{D}_{eq}$. Thus, for $\eta = 0$, the biased dataset is just $\mathscr{D}_{eq}$, for $\eta = 1$ it is just $\mathscr{D}_{opp}$ and for $\eta = \frac{1}{2}$ the biased dataset is an ordinary subset of the whole data. The test set is simply the data remaining from the initial split.

6.5.2 *Evaluation protocol*

We evaluate our results in terms of accuracy and fairness. A model that perfectly decouples its predictions from s will achieve near-uniform accuracy across all biasing-levels. For binary s/y we quantify the fairness of a classifier's predictions using *demographic parity* (DP): the absolute difference in the probability of a positive prediction for each sensitive group.

6.5.3 *Experimental results*

We report the results from two image datasets. cMNIST, a synthetic dataset, is a good starting point for evaluating our model due to the direct control we have over the biasing. CelebA, on the other hand, is a more practical and challenging example. We also test our method on a tabular dataset, the Adult dataset.



(a) Samples from the cMNIST training set, $\sigma = 0$.

(b) $\boldsymbol{x}_u$ null-samples from the cFlow model.

(c) $\boldsymbol{x}_b$ null-samples from the cFlow model.

Figure 6.5: Sample images from the coloured MNIST dataset problem with 10 predefined mean colours. (a): Images from the spuriously correlated subpopulation where colour is a reliable signal of the digit class-label. (b-c): Results of running our approach realised with cFlow on the cMNIST dataset. The model learns to retain the shape of the digit shape while removing the relationship with colour. A downstream classifier is now less prone to exploiting correlations between colour and the digit label class.

cMNIST. The coloured MNIST (cMNIST) dataset is a variant of the MNIST dataset in which the digits are coloured. In the training set, the colours have a one-to-one correspondence with the digit class. In the test set (and the representative set), colours are assigned randomly. The colours are drawn from Gaussians with 10 different means. We follow the colourisation procedure outlined by Kim et al. (2019), with the mean colour values selected so as to be maximally dispersed. The full list of such values can be found in section 6.7.5. We produce multiple variants of the cMNIST dataset corresponding to different standard deviations σ for the colour sampling: $\sigma \in \{0.00, 0.01, ..., 0.05\}$.

For this specific dataset, we can establish an additional baseline by simply grey-scaling the dataset which only leaves the luminosity as spurious information. We also evaluate the model, with all the associated hyperparameters, from Kim et al. (2019). The only difference between the setups is the dataset creation, including the range of σ values we consider. Our versions of the dataset, on the whole, exhibit much stronger colour bias, to the point of the mapping the digit's colour and class being bijective. figure 6.4 shows that the model significantly underperforms even the naïve baseline, aside from at $\sigma = 0$, where they are on par.

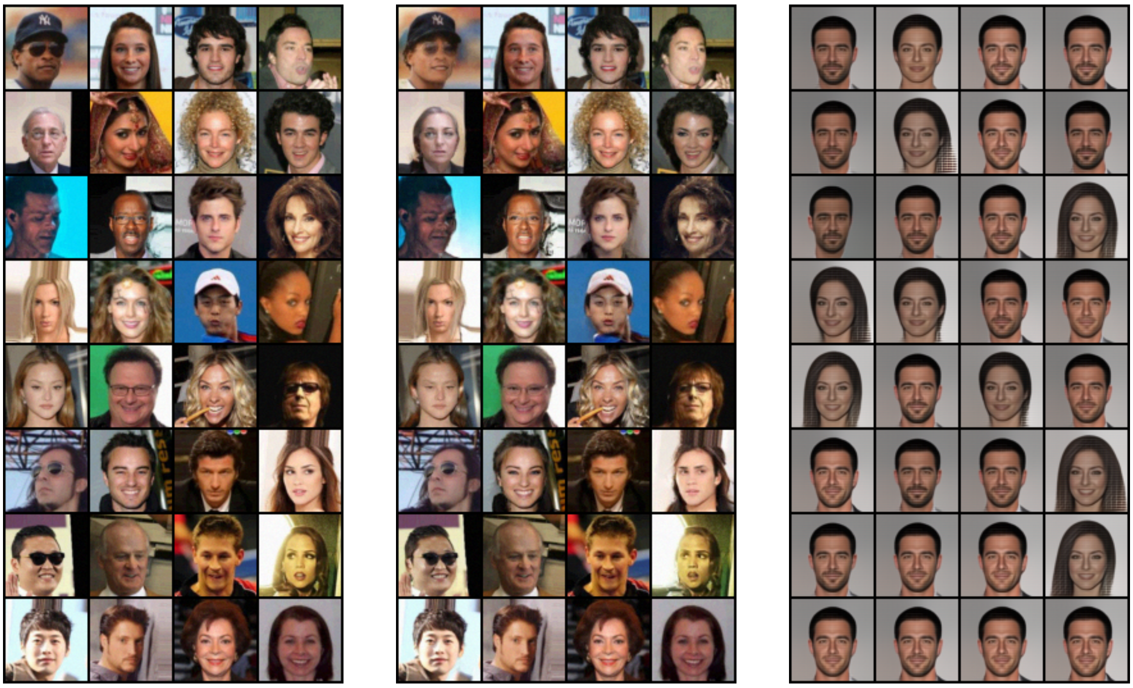
Inspection of the null-samples shows that both the cVAE and cFlow model succeed in removing almost all colour information, which is supported quantitatively by figure 6.4, and qualitatively by figure 6.5. While the cVAE outperforms cFlow marginally at low σ values, performance degrades as this increases. This highlights the problems with the conditional decoder we anticipated in section 6.4.3. The lower σ, and therefore the variation in sampled colour, is, the more reliably the

s label, corresponding to the mean of RGB distribution, encodes information about the colour. For higher σ values, the sampled colours can deviate far from the mean and so the encoder must incorporate information about s into its representation if it is to minimise the reconstruction loss. cFlow, on the other hand, is consistent across σ values.

CelebA. To evaluate the effectiveness of our framework on real-world image data we use the CelebA dataset (Liu et al., 2015), consisting of 202,599 celebrity images. These images are annotated with various binary physical attributes, including 'gender', 'hair colour', 'young', etc, from which we select our sensitive and target attributes. The images are centre cropped and resized to 64 × 64, as is standard practice. For our experiments, we designate 'gender' as the sensitive attribute, and 'smiling' and 'high cheekbones' as target attributes. We chose gender as the sensitive attribute as it a common sensitive attribute in the fairness literature. For the target attributes, we chose attributes that are harder to learn than gender and which do not correlate too strongly with gender in the dataset ('wearing lipstick' for example being an attribute too closely correlated with gender). The model is trained on the representative set (normal subset of CelebA) and is then used to encode the artificially biased training set and the test set. The results for the most strongly biased training set ($\eta = 0$) can be found in figure 6.2. Our method outperforms the baselines in accuracy and fairness.

We also assess performance for different mixing factors (η) which correspond to varying degrees of bias in the training set (see figure 6.3). This is to verify that the model does not *harm* performance when there is not much bias in the training set. For these experiments, the model is trained once on the representative set and is then used to encode different training sets. The results show that for the intermediate values of η, our model incurs a small penalty in terms of accuracy, but at the same time makes the results *fairer* (corresponding to an accuracy-fairness trade-off). Qualitative results can be found in figure 6.6 (images from cVAE can be found in section 6.7.7).

To show that our method can handle multinomial, as well as binary, sensitive attributes, we also conduct experiments with s = hair colour as a ternary attribute ('Blonde', 'Black', 'Brown'), excluding 'Red' because of the paucity of samples and the noisiness of their labels. The results for these experiments can be found in section 6.7.3.



(a) Original images.

(b) $\boldsymbol{x}_u$ null-samples from the cFlow model.

(c) $\boldsymbol{x}_b$ null-samples from the cFlow model.

Figure 6.6: CelebA null-samples learned by our cFlow model, with gender as the sensitive attribute. (a) The original, untransformed samples from the CelebA dataset (b) Reconstructions using only information unrelated to s. (c) Reconstruction using only information related to $\neg s$. The model learns to disentangle gender from the non-gender related information. Note that some attributes like skin tone seem to change along with gender due to the correlation between the attributes. This is especially visible in images (1,1) and (3,2). Only because our representations are produced in the data-domain can we easily spot such instances of entanglement.

Results for the UCI Adult dataset. The UCI Adult dataset consists of census data and is commonly used to evaluate models focused on algorithmic fairness. Following convention, we designate 'gender' as the sensitive attribute s and whether an individual's salary is \$50,000 or greater as y. We show the performance of our approach in comparison to baseline approaches in figure 6.7. We evaluate the performance of all models for mixing factors (η) 0 and 1. Results shown in figure 6.7 show that we match or exceed the baseline. In terms of fairness metrics, our approach generally outperforms the baseline models for both of η. Detailed results can be found in section 6.7.3.

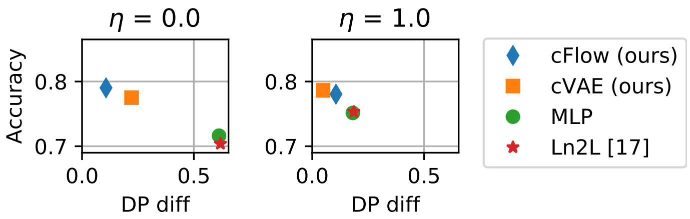


Figure 6.7: Results for the Adult dataset. The $\boldsymbol{x}$-axis corresponds to the difference in positive rates. An ideal result would occupy the top-left.

We also did experiments to show that the encoder transfers to other tasks. These transfer-learning experiments can be found in section 6.7.8.

6.6 Conclusion

We have proposed a general and straightforward framework for producing invariant representations, under the assumption that a representative but partially-labelled *representative* set is available. Training consists of two stages: an encoder is first trained on the representative set to produce a representation that is invariant to a designated spurious feature. This is then used as input for a downstream task-classifier, the training data for which might exhibit extreme bias with respect to that feature. We train both a VAE- and INN-based model according to this procedure, and show that the latter is particularly well-suited to this setting due to its losslessness. The design of the models allows for representations that are in the data domain and therefore exhibit meaningful invariances. We characterise this for synthetic as well as real-world datasets for which we develop a method for simulating sampling bias.

Acknowledgements

This work was in part funded by the European Research Council under the ERC grant agreement no. 851538. We are grateful to NVIDIA for donating GPUs.

6.7 Appendix

6.7.1 *Model Architectures*

For both cMNIST and CelebA we parameterise the coupling layers with the same convolutional architecture as in Kingma and Dhariwal (2018), consisting of 3 convolutional layers each with 512 filters of, in order, sizes 3×3, 1×1, and 3×3. Following Ardizzone et al. (2019), we Xavier initialise all but the last convolutional layer of the s and t sub-networks which itself is zero-initialised so that the coupling layers begin by performing an identity transform. We used a Glow-like architecture (Kingma and Dhariwal, 2018) (affine coupling layers together with checkerboard

Table 6.1: INN architecture used for each dataset.

Dataset	Levels	Level depth	Coupl. chan.	Input to discr.
UCI Adult	1	1	35	Null-samples
cMNIST	3	16	512	Encodings
CelebA	3	32	512	Encodings

Table 6.2: cVAE encoder architecture used for each dataset. The decoder architecture in each case mirrors that of its encoder counterpart through use of transposed convolutions. For the adult dataset we apply ℓ_2 and cross-entropy losses to the reconstructions of the continuous features and discrete features, respectively.

Dataset	Initial channels	Levels	β	Recon. loss
UCI Adult	35	–	0	ℓ_2 + CE
cMNIST	32	4	0.01	ℓ_2
CelebA	32	5	1	ℓ_1

reshaping and invertible 1×1 convolutions) for the convolutional INNs. Table 6.1 summarises the INN architectures used for each dataset.

For the image datasets each level of the cVAE encoder consists of two gated convolutional layers (Oord et al., 2016) with ReLU activation. At each subsequent level, the number of filters is doubled, starting with an initial value 32 and 64 in the case of CelebA and cMNIST respectively. In the case of the Adult dataset, we use an encoder with one fully-connected hidden layer of width 35, followed by SeLU activation (Klambauer et al., 2017). For both cMNIST and CelebA, we downsample to a feature map with spatial dimensions 8×8, but with 3 and 16 channels respectively. For the Adult dataset, the encoding is a vector of size 35. The output layer specifies both the parameters (mean and variance) of the representation's distribution. In all cases the KL-divergence is computed with respect to a standard isotropic Gaussian prior. Details of the encoder architectures can be found in table 6.2. The loss pre-factors were sampled from a logarithmic scale; without proper balancing the networks can exhibit instability, especially during the early stages of training.

6.7.2 *Instructions for potential users*

The first question a potential user has to ask themselves is whether the method is a good fit: is the problem that the user faces one of strong spurious correlation and is there non-spurious data available that has labels for the spurious variable? To investigate the first part of the question, the user should first try to train a standard neural network classifier and observe the test-set performance. Furthermore, one should check whether the spurious variable can be removed with data augmentations alone.

If the features of the data are categorical instead of continuous, it is best to first produce a continuous representation with an autoencoder. This step only has to be done once at the beginning.

The next question is whether to use the cFlow or cVAE variant of the method. For initial experiments, we would recommend the cVAE model as it is quicker to train, and will lead to shorter feedback cycles when validating the code. If the computational budget allows it, we would recommend switching to the cFlow model once cVAE is working as it provides better guarantees regarding the retention of information from the input data.

For choosing the network architecture, the only advice we have is to look at what architectures other people have used for similar data. Note, however, that encoder-decoder architectures usually differ in some ways from classification architectures, due to their different goals: the goal of the former is primarily to compress and disentangle, while the latter aims to *discard* information unrelated to the prediction task. As such, certain layer types, like pooling layers and batch normalisation, are only suitable for classifiers and not encoder-decoders. For the Invertible Neural Network (INN) architecture, the most import advice is to keep in mind that each individual layer is much less expressive than non-invertible layers, and so the number of layers required in INNs is much higher. However, the number also should not be *too high* or the model will overfit. It is likely that the architecture needs to be adapted during training. See also section 6.7.4 and the code we published alongside this paper to get inspiration for architectures.

During training, the user should mostly keep an eye on two variables: the reconstruction loss and the degree of invariance of $\boldsymbol{z}_u$ w.r.t. s, which can either be gleaned from looking at reconstructions of $\boldsymbol{z}_u$ or from computing the accuracy of a downstream classifier trained on $\boldsymbol{z}_u$. The information inherent in the reconstruction loss can also be obtained by looking at full reconstructions of $\boldsymbol{z}$. If the reconstruction loss does not go down during training, some possible reasons are: the dimension of the representation is too small, the reconstruction loss weight is too small or the network just needs to be trained for longer. If the degree of invariance does not increase during training, some possible reasons are: the network is not expressive enough (e. g. not deep enough) to disentangle $\boldsymbol{z}_u$ and $\boldsymbol{z}_b$, the adversary is not powerful enough, the adversarial loss weight is too small or the network just needs to be trained for longer.

For INN training, there is the additional complication that it can become non-invertible due to numerical problems (e. g. division by zero). If this happens, the losses will quickly diverge and further training will become pointless. See section 6.7.6 for some ways of preventing this.

6.7.3 *Additional results*

DETAILED RESULTS FOR UCI ADULT DATASET. This census data is commonly used to evaluate models focused on algorithmic fairness. Following convention, we designate 'gender' as the s and whether an individual's salary is \$50,000 or greater as y. We show the performance of our approach in comparison to baseline approaches in figure 6.8. We evaluate the performance of all models for mixing factors (η) of value $\{0, 0.1, ..., 1\}$. Results shown in figure 6.8 show that whilst our model fails to surpass the baseline models in terms of accuracy for the balanced case (and those close to it), we match or exceed the baseline as η moves the dataset to a more imbalanced setting. In terms of fairness metrics, our approach generally outperforms the baseline models regardless of η.

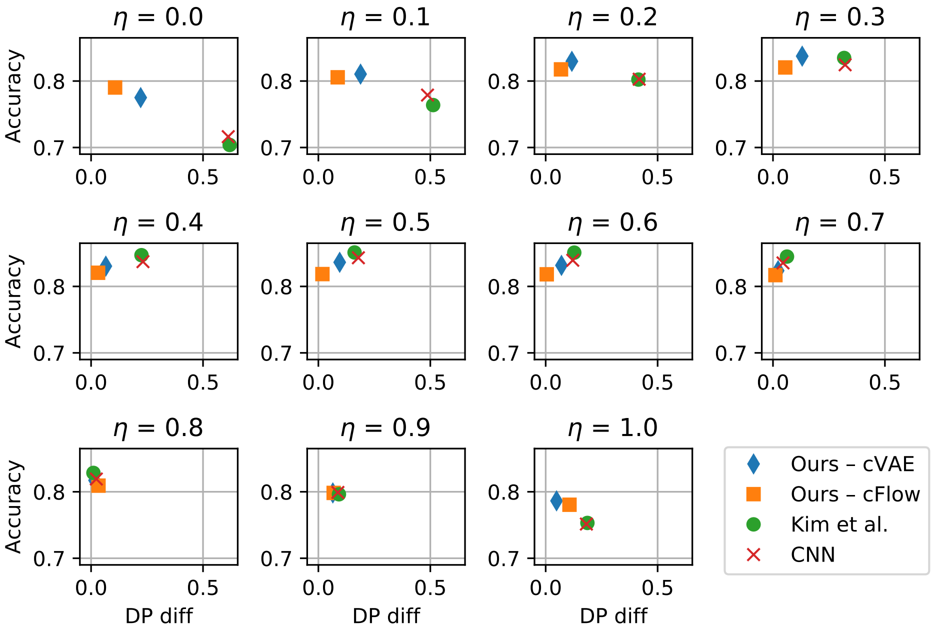


Figure 6.8: Results for the ADULT dataset. The x-axis corresponds to the difference in positive rates. An ideal result would occupy the TOP-LEFT.

MULTINOMIAL SENSITIVE ATTRIBUTES. In addition to binary sensitive attribute s, we also investigate multinomial s in the CelebA dataset. First, we do experiments with hair colour, where s

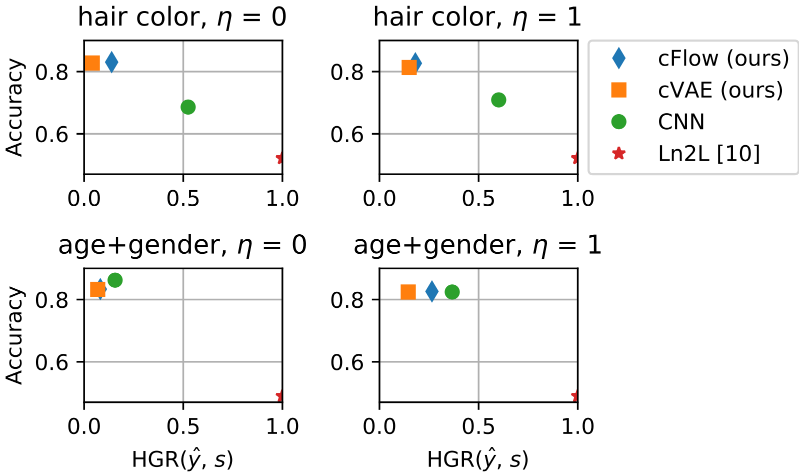


Figure 6.9: For *hair colour*, s takes on the values Blond, Brown and Black. For *age+gender*, s takes on the values Young/Female, Young/Male, Old/Female and Old/Male.

Table 6.3: Results on the CelebA dataset with different sizes of $\boldsymbol{z}_b$.

$\lvert\boldsymbol{z}_b\rvert$	$\lvert\boldsymbol{z}_b\rvert/\lvert\boldsymbol{z}\rvert$	Accuracy	DP diff
1	0.0082%	0.60	0.63
3	0.0245%	0.60	0.63
5	0.0410%	0.84	0.12
10	0.0820%	0.84	0.12
30	0.2442%	0.74	0.23
50	0.4070%	0.68	0.27

has three possible values: blond hair, brown hair and black hair. The other experiment is with a combination of age and gender, where s has four possible values, each of which is a combination of a gender and an age: Young/Female, Young/Male, Old/Female and Old/Male. To evaluate the fairness for multinomial s, we use the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient (HGR) (Mary et al., 2019) that is defined on the domain $[0, 1]$ and gives $\text{HGR}(Y, S) = 0$ iff $Y \perp S$ and 1 if there is a deterministic function to map between them. Results can be found in figure 6.9.

Investigation into the size of $\boldsymbol{z}_b$. In the cFlow model, the size of $\boldsymbol{z}_b$ is an important hyperparameter which can affect the result significantly. Here we investigate the sensitivity of the model to the choice of $\boldsymbol{z}_b$ size. Table 6.3 shows accuracy and fairness (as measured by *DP diff*) for different sizes of $\boldsymbol{z}_b$. The results show that both too large and too small $\boldsymbol{z}_b$ is detrimental. However, they also show that the model is not overly sensitive to this parameter: both sizes 5 and 10 achieve nearly identical results.

Table 6.4: Additional fairness metrics for the experiments on the CelebA dataset (figure 6.3 from the main text). *TPR diff.* refers to the difference in true positive rate. *TNR diff.* refers to the difference in true negative rate. Left: $\eta = 0$. Right: $\eta = 1$.

Method	Accuracy	DP diff	TPR diff	TNR diff
cFlow	0.83	0.10	0.15	0.25
cVAE	0.82	0.05	0.09	0.18
CNN	0.61	0.63	0.70	0.64
Ln2L	0.52	0.00	0.00	0.00

Method	Accuracy	DP diff	TPR diff	TNR diff
cFlow	0.82	0.33	0.28	0.21
cVAE	0.81	0.16	0.10	0.05
CNN	0.67	0.75	0.66	0.76
Ln2L	0.51	0.08	0.06	0.09

Additional fairness metrics. In addition to *DP diff*, we report here the result from other fairness measures. These results are from the same setup as those reported in the main paper. We report the difference in true positive rates (TPRs) between the two groups (male and female), which corresponds to a measure of Equality of Opportunity, and the difference in true negative rates (TNRs) between the two groups.

6.7.4 *Optimisation Details*

All our models were trained using the RAdam optimiser (Liu et al., 2020) with learning rates 3×10^{-4} and 1×10^{-3} for the encoder/discriminator pair and classifier respectively. A batch size of 128 was used for all experiments.

We now detail the optimisation settings, including the choice of adversary, specific to each dataset. Details of the cVAE and cFlow architectures can be found in table 6.2 and table 6.1, respectively.

UCI Adult. For this dataset our experiment benefited from using null-samples as inputs to the adversary of the cFlow model. Unlike for the image datasets, we found a single adversary to be sufficient. This was realised as a multi-layer perceptron (MLP) with one hidden layer, 256 units wide. The INN performs a bijection of the form $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$. However, the adult dataset is composed of mostly discrete (binary/categorical) features. To achieve good performance, we found it necessary to first pre-process the inputs with a pretrained autoencoder, using its encodings as the input to the cFlow model, as well as to the adversary. The learned representations were evaluated with a logistic regression model from scikit-learn (Pedregosa et al., 2011), using the standard settings. All baseline models were trained for 200 epochs. The Ln2L (Kim et al., 2019) and MLP baselines share the architecture of the cVAE's encoder, only with a classification layer affixed.

Coloured MNIST. Each level of the architecture used for the downstream classifier and naïve baseline alike consists of two convolutional layers, each with kernel size 3 and followed by Batch Norm (Ioffe and Szegedy, 2015) and ReLU activation. For the Ln2L baseline, we use an a setup identical to that described in Kim et al. (2019). Each level has twice the number of filters in its convolutional layer and half the spatial input dimensions as the last. The original input is downsampled to the point of the output being reduced to a vector, to which a fully-connected classification layer is applied.

To allow for an additional level in the INN (the downsampling operations requiring the number of spatial dimensions to be even), the data was zero-padded to a size of 32 × 32. The cVAE and cFlow models were trained for 50 and 200 epochs respectively, using ℓ_2 reconstruction loss for the former. The downstream classifier and all baselines were trained for 40 epochs. For both of our models, an ensemble of 5 adversaries was applied to the encodings, with each member taking the form of a fully-connected ResNet, 2 blocks in depth, with SeLU activation (Klambauer et al., 2017). The adversaries were reinitialised independently with probability 0.2 at the end of each epoch. While the adversaries could equally well take null-samples as input, as done for the Adult dataset, doing so requires the performing of both forward and inverse passes each iteration, which, for the convolutional INNs of the depths we require for the image datasets, introduces a large computational overhead, while also showing to be the less stable of the two approaches in our preliminary experiments.

CelebA. The downstream classifier and naïve baseline take the same form as described above for cMNIST, but with an additional level with 32 filters in each of its convolutions at the top of the network. For this dataset we adapt the Ln2L model by simply considering it as an augmentation the naïve baseline's objective function, with the entropy loss applied to the output of the final convolutional layer. These models were again trained for 40 epochs, which we found to be sufficient for convergence for the tasks in question. The cVAE and cFlow models were respectively trained for 100 epochs and 30 epochs, using ℓ_1 reconstruction loss for the former. Compared with cMNIST, the size of the adversarial ensemble was increased to 10, the reinitialisation probability to 0.33, but no changes were made to the architectures of its members.

The Pitfalls of Adversarial Training. Adversarial learning has become one of the go-to methods for enforcing invariance in fair representation learning (Ganin et al., 2016) with MMD

(Louizos et al., 2016) and HSIC (Quadrianto et al., 2019), being popular non-parametric alternatives. Ganin et al. (2016) proposed adversarial learning for domain adaptation problems, with Edwards and Storkey (2016) soon after making this and learning a representation promoting demographic parity. The adversarial approach carries the benefits of being both efficient and scalable to multi-class categorical variables, which many sensitive attributes are in practice, whereas the non-parametric methods only permit pair-wise comparison.

However, when realised as a neural network, the adversary is both sensitive to the values of the inputs as well as their ordering (though exchangeable architectures, such as Zaheer et al. (2017) do exist, but which sacrifice expressiveness). Thus, it can happen that the representation learner optimises for the surrogate objective of eluding the adversary rather than the real objective of expelling s-related information. Moreover, the non-stationarity of the dynamics can lead to cyclic-equilibria, irrespective of the capacity of the adversary.

When working with a partitioned latent space, this behaviour can be averted by instead encouraging $\boldsymbol{z}_b$ to be predictive of s, acting as a kind of information 'sink', as in Jacobsen et al. (2018). However, this does not have the guarantee of making $\boldsymbol{z}_u$ invariant to s - there are often many indicators for s, not all of which are needed to predict the label perfectly. Training the network to convergence before taking each gradient step with the representation learner is one way one to attempt to tame the unstable minimax dynamics (Feng et al., 2019). However, this does not prevent the emergence of the aforementioned cyclicity.

We try to mitigate the aforementioned degeneracies by maintaining a diverse set of adversaries, as has shown to be effective for GAN training (Durugkar et al., 2017), and by decorrelating the individual trajectories by intermittently re-initialising them with some small probability following each iteration.

Tuning the Partition Sizes. There are several ways of ensuring that the size of $\boldsymbol{z}_b$ is sufficient to capture all s dependencies, but minimal enough that information unrelated to s is maximally preserved We adopt the straightforward search strategy of, starting from some initial guess, calibrating the value according to accuracy attained by a classifier trained to predict s from $\boldsymbol{z}_b$ on a held-out subset of the representative set, which is measured whenever the adversarial loss plateaus. If the accuracy is above chance level then that suggests the size of the $\boldsymbol{z}_b$ partition, $|\boldsymbol{z}_b|$, needs to be increased to accommodate more information about s. If the accuracy is found to be at chance level then are two possibilities: 1) $|\boldsymbol{z}_b|$ is already optimal; 2) $|\boldsymbol{z}_b|$ is large enough

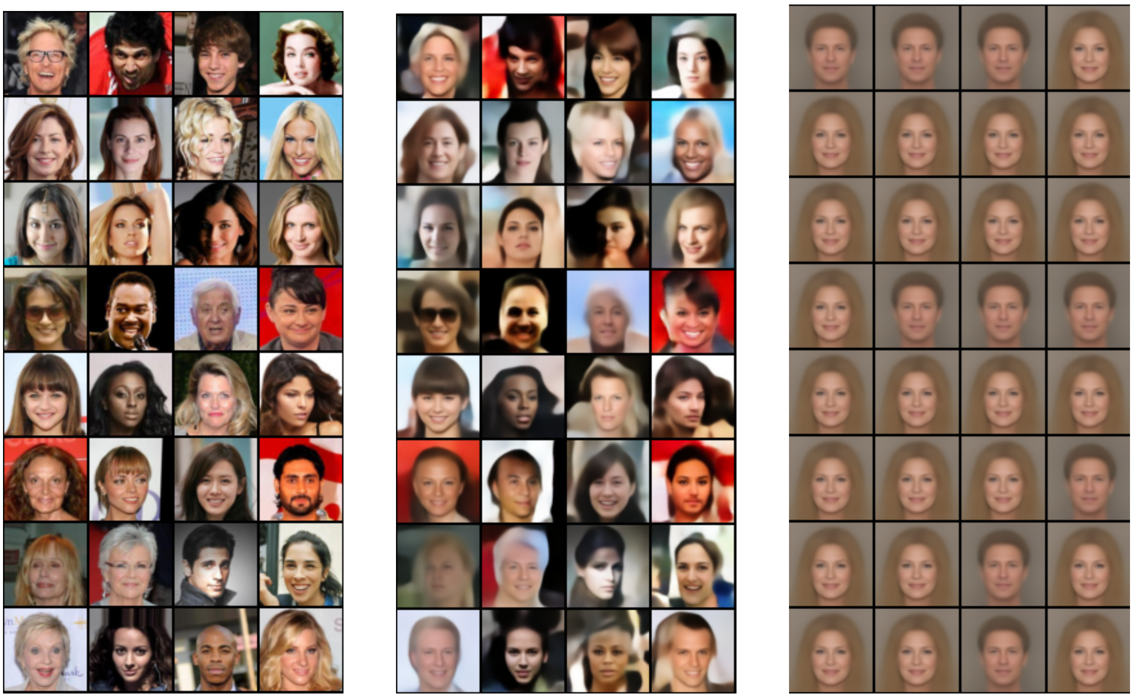
Table 6.5: Mean RGB values (in practice normalised to $[0, 1]$) parameterising the Multivariate Gaussian distributions from which each digit's colour is sampled in the biased (training) dataset. In the representative and test sets, the colour of each digit is sampled from one of the specified Gaussian distributions at random.

Digit	Colour Name	Mean RGB
0	Cyan	(0, 255, 255)
1	Blue	(0, 0, 255)
2	Magenta	(255, 0, 255)
3	Green	(0, 128, 0)
4	Lime	(0, 255, 0)
5	Maroon	(128, 0, 0)
6	Navy	(0, 0, 128)
7	Purple	(128, 0, 128)
8	Red	(255, 0, 0)
9	Yellow	(255, 255, 0)

that it fully contains both information s as well as that of a portion of y. If the former is true, then perturbations around the current value allow us to confirm this; if the latter is true then decreasing the value was indeed the correct decision.

6.7.5 *Synthesising Coloured MNIST*

We use a colourised version of MNIST as a controlled setting investigate learning from biased data in the image domain. In the biased training set, each digit is assigned a unique mean RGB value parameterising the multivariate Gaussian from which its colour is drawn. These values were chosen to be maximally dispersed across the 8-bit colour spectrum and are listed in table 6.5. By adjusting the standard deviation, σ, of the Gaussians, we adjust the degree of bias in the dataset. When $\sigma = 0$, there is a perfect and noiseless correspondence between colour and digit class which a classifier can exploit. The classifier can favour the learning of the low-level spurious feature over those higher level features constituent of the digit's class. As the standard deviation increases, the sampled RGB values are permitted to drift further from the mean, leading to overlap between the samples of the colour distributions and reducing their reliability as indicators of the digit class. In the test and representative sets alike, however, the colour of each sample is sampled from one of the 10 distributions randomly, such that colour can no longer be leveraged as a shortcut to predicting the digit's class.



(a) Original images.

(b) $\boldsymbol{x}_u$ null-samples generated by the cVAE model.

(c) $\boldsymbol{x}_b$ null-samples generated by the cVAE model.

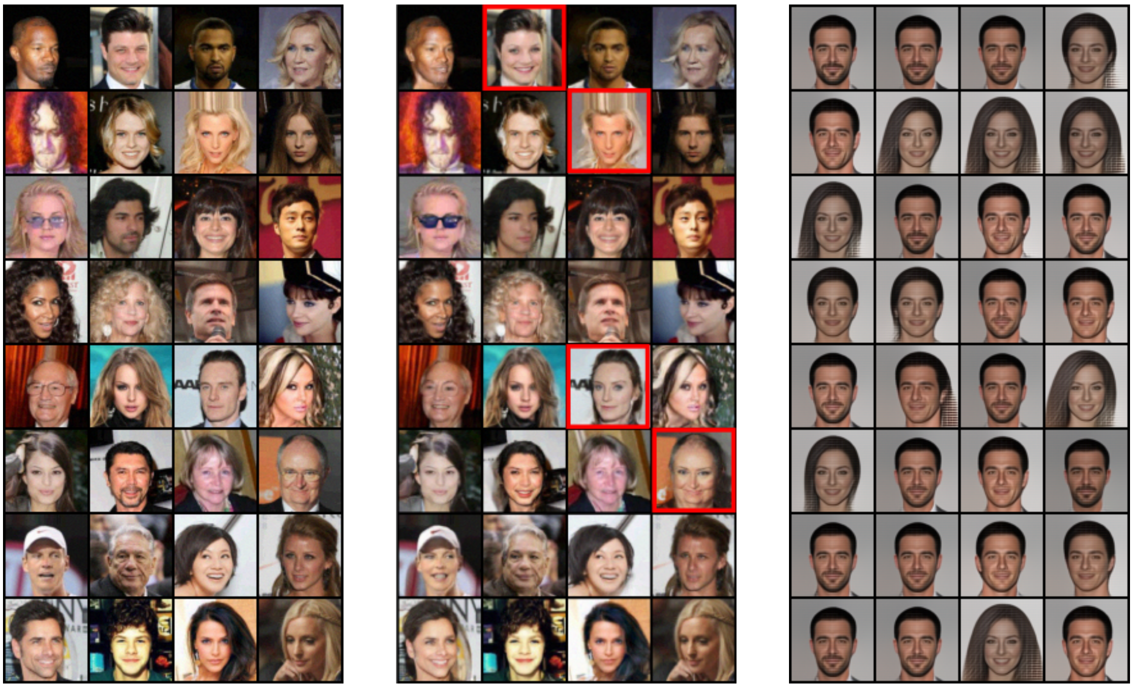
Figure 6.10: CelebA null-samples learned by our cVAE model, with gender as the sensitive attribute. (a) The original, untransformed samples from the CelebA dataset (b) Reconstructions using only information unrelated to s. (c) Reconstruction using only information related to $\neg s$. The model learns to disentangle gender from the non-gender related information. Compared with the cFlow model, there is a severe degradation in reconstruction quality due to the model trying to simultaneously satisfy conflicting objectives.

6.7.6 *Stabilising the Coupling layers*

Heuristically, we found that applying an additional nonlinear function to the scale coefficient of the form

$$s = \sigma(f(u)) + 0.5 \tag{6.7}$$

greatly improved the stability of the affine coupling layers. Here, σ is the logistic function, which we shift to be centred on 1 so that zero-initialising f results in the coupling layers initially performing an identity-mapping.



(a) Original images.

(b) $\boldsymbol{x}_u$ null-samples generated by the cFlow model.

(c) $\mathbf{x}_b$ null-samples generated by the cFlow model.

Figure 6.11: CelebA null-samples learned by our cFlow model, with gender as the sensitive attribute. (a) The original, untransformed samples from the CelebA dataset (b) Reconstructions using only information unrelated to s. (c) Reconstruction using only information related to $\neg s$. The model learns to disentangle gender from the non-gender related information. Attributes such as *makeup* and *hair length* are also often modified in the process (prime examples framed with red) due to inherent correlations between them and the sensitive attribute, which the interpretability of our representations allows us to easily identify.

6.7.7 *Qualitative Results for CelebA*

Learning a representation alongside its inverse mapping, be it approximate or exact, enables us to probe the behaviour of the model that produced it, and any biases it may have implicitly captured due to entanglement between the sensitive attribute and other attributes present in the data. We highlight a few examples of such biases manifesting in the cFlow model's CelebA null-samples in figure 6.11. In these cases, makeup and hair style have been inadvertently modified during the null-sampling due to the tight correlation between these two attributes and the sensitive attribute, gender, to which we had aimed to make our representations invariant. Additionally, in all highlighted images, the skin tone has changed: from male to gender-neutral, the skin becomes lighter and from female to gender-neutral, the skin becomes darker; in the change from male to gender-neutral, glasses are also often removed. As the model cannot know that the label is meant to only refer to gender, and not to these other (correlated) attributes, the links cannot be disentangled by the model. However, the advantage of our method is that we can at least identify such biases due to the interpretability that comes with the representations being in the data domain.

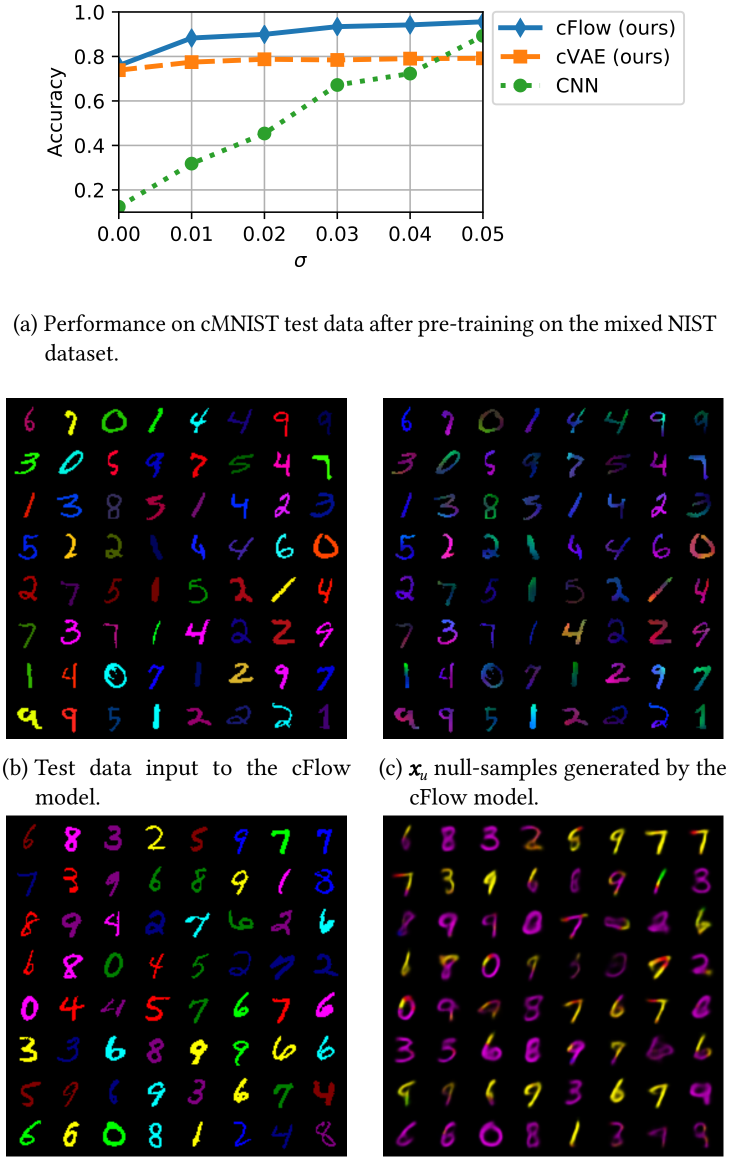
6.7.8 *Transfer Learning*

For our method, we require a representative set which follows the same distribution as that observed during deployment. Such a representative set might not always be available. In such a scenario, we can resort to using a set that is merely *similar* to that in the deployment setting and leverage transfer learning.

One of the advantages of using an invertible architecture over conventional, *surjective* ones that we stressed in the main text is its *losslessness*. Since the transformations are necessarily bijective, the information contained in the input can never be destroyed, only redistributed. This makes such models particularly well-suited, in our minds, for transferring learned invariances: even if the input is unfamiliar, no information should be lost when trying to transform it. This works as long as only the information about s ends up in the $\boldsymbol{z}_b$ partition. If s takes a form similar to that which we pre-trained on, and can thus be correctly partitioned in the latent space, by complement we have the information about $\neg s$ stored in the $\boldsymbol{z}_u$ partition, without presupposing similarity to the $\neg s$ observed during pre-training.

Transferring from mixed-NIST to MNIST. We test our hypothesis by comparing the performance of the cFlow and cVAE models pre-trained on a mixture of datasets belonging to the NIST family, colourised in the same way as cMNIST, while the downstream train and test sets remain the same as in the original cMNIST experiments. Specifically, we create this representative set by sampling 24,000 images (to match the cardinality of the original representative set) from EMNIST (letters only) (Cohen et al., 2017), FashionMNIST (Xiao et al., 2017) and KMNIST (Clanuwat et al., 2018), in equal proportion. We use the same architectures for the cVAE and cFlow models as we did in the non-transfer learning setting. In terms of hyperparameters, the only change made was to the KL-divergence's pre-factor, finding it necessary to increase it to 1 to guarantee stability.

The results for the range of σ values are shown in figure 6.12a. Unsurprisingly, the performance of both models suffers when the representative and test sets do not completely correspond. However, the cFlow model consistently outperforms the cVAE model, with the gap increasing as the bias decreases. Although some colour information is retained in the cFlow null-samples, symptomatic of an imperfect transfer, semantic information is almost entirely retained as well. Conversely, the cVAE is very much flawed in this respect; as can be seen in the bottom row of figure 6.12a, for some samples, semantic information is degraded to the point of the digit's identity being altered. As a result of this semantic degradation, the performance of the downstream classifier is curtailed by the noisiness of the digit's identity and is relatively unchanging across σ-values, in contrast to the monotonic improvement of that achieved on the cFlow null-samples.



(a) Performance on cMNIST test data after pre-training on the mixed NIST dataset.

(b) Test data input to the cFlow model.

(c) $\boldsymbol{x}_u$ null-samples generated by the cFlow model.

(d) Test data input to the cVAE model.

(e) $\boldsymbol{x}_u$ null-samples generated by the cVAE model.

Figure 6.12: Results for the transfer learning experiments in which the representative set consists of colourised samples from EMNIST, KMNIST, and FashionMNIST, while the downstream dataset remains as cMNIST. (a) Quantitative results for different σ-values. (b-c) Qualitative results for the cFlow model. (d-e) Qualitative results for the cVAE model. The qualitative results provide comparisons of the images before (left) and after (right) null-sampling. Note that for some of the cVAE samples, the clarity of the digits has clearly changed due to null-sampling, serving as an explanation for the non-increasing downstream performance.

Part III

CONCLUSION

7 DISCUSSION AND CONCLUSION

...if a machine is expected to be infallible, it cannot also be intelligent.
— Lecture to the London Mathematical Society, 20 February 1947
Alan Turing

In this thesis, I have investigated the problem of developing fair representations of data that are inherently interpretable. The purpose of this is to increase trust in machine learning (ML) systems with the population on which they will be deployed. The intention is that by making fair representations more interpretable, there will be higher levels of accountability in the decisions to use and maintain them (or, indeed, a decision to *not* use or maintain fair representations). To do this, I have explored three approaches based on *fair representations in the data domain*. These are documented in chapters 4 to 6.

In chapter 4 I built on existing adversarial-based fair representation methods to interrogate the interaction of a specific, predefined protected attribute on the remaining features and also the decision outcome. This was achieved by disentangling the protected attribute from the remaining features and manipulating its value during both feature reconstruction and prediction. I made a connection between this approach and a restricted form of counterfactual modelling, where only one predefined variable can be intervened on. Furthermore, I proposed a novel use-case for this approach, introducing an additional 'positive action' outcome. This approach has two goals: The first is to take advantage of a realistic and practical mechanism to promote greater equality in outcomes. The second goal is to provide more insight and facilitate discussion about quantifying different *types* of discrimination present in datasets.

In chapter 5 I introduced fair representations that are constrained to exist in the data domain. This was achieved by modelling fair representation learning as a problem of data-to-data translation. Uniquely, instead of translating between existing subdomains present in the data, e.g., male-to-female or vice versa, the target subdomain is not present in the dataset. To do this, I use the simple assumption that an input can be decomposed into a 'fair' and 'unfair' component. To encourage the fair component to be independent of a protected characteristic, a kernel-based independence measure was used. The work in this chapter was one of the first to apply fairness constraints to images and has since been credited as independently introducing 'controllable fair representations'[1].

Finally, in chapter 6 I extended the work of the previous chapter by allowing a more complex relationship between the fair and unfair components of a sample to be modelled. Using Invertible Neural Networks (INNs), a type of generative model that captures the exact likelihood of high-dimensional probability distributions, I produced a more detailed representation for each sample. One of the fundamental benefits of this model choice was the ability to capture the effect of specific parts of the embedding. Through the process of 'null-sampling', I was able to visualise not only the effect of a particular region of the embedding space on the reconstruction, but also the information that this specific region captured. Furthermore, the setting of a disparity in demographics between the training and deployment distributions was investigated, with a proposal to use additional unlabelled data to overcome cases of missing demographics.

7.1 Limitations and Intended Use

Given the nature of this thesis topic, it is natural to consider the limitations of the presented work from a practical, and to some extent, a more general anthropocentric perspective. Although the work presented in this thesis is intended to produce fairer outcomes, it is not claimed that by employing these techniques, *all* biases and discriminatory practices can be identified, accounted for, and remedied. That is not to disparage or minimise this work, or research into this topic — in fact quite the opposite. There is an argument that unfair behaviour is so ingrained in almost all aspects of life in cultures around the world that fairness concerns are ubiquitous. Addressing these challenges, therefore, such as by employing the methods described in this thesis, has significant potential impact. Furthermore, almost all new ML applications are likely to have a practical 'fair cold start' problem stemming from the challenge of training with imperfect data and focusing on utility for the majority of users to alleviate commercial pressures. When these applications have social impact, there will, of course, be greater scrutiny, and rightly so. Being vigilant then to the implications of decisions that we, as ML practitioners, make is important and is likely to become *fundamental* in the coming decades based on the rapid growth of ML applications. I

[1] The term 'controllable fair representations' was coined in parallel work (Song et al., 2019)

am optimistic that as the research community develops more approaches and becomes more precise in its characterisation and definition of different biases, greater improvements will be made. Meanwhile, an initial step towards fairer outcomes, regardless of the task, is taking time to understand and document the limitations of both the available data and the model. This may seem obvious in some way, but as the use of ML becomes more readily available and automated model training and deployment become more prevalent, it is worth reiterating that the insight of such an analysis can be incredibly beneficial.

An important element of this discussion is that not all tasks should be considered equally valid use-cases for predictive modelling. There are some tasks that may initially appear to be possible, but that does not mean that development in these areas should be pursued. For example, investigating predictions of sexuality (Wang and Kosinski, 2018), or predisposition to committing crime (Wu and Zhang, 2016; Harrisburg University, 2020) from an image of a subject's face are tasks that verge on pseudoscience. In addition, they are both unlikely to hold up to any serious scrutiny and, outside of a narrow, peer-reviewed academic context, are also likely to violate ethical guidelines such as the British Computer Society Code of Conduct (BCS, 2022). Although these unscrupulous use-cases have the potential to cause significant harm, very few new articles related to their development are released annually and these remain subject to intense scrutiny. While it may be tempting to stop development of new ML approaches due to the risk of harm, I am motivated by the number of positive use-cases where ML can improve the lives and well-being of users. If it is determined that the use-case presented for predictive modelling can be beneficial for all, there are methods available, such as those presented in this thesis, that can contribute to fairer outcomes. If applicable, these methods should be considered.

In terms of practical considerations, all of the methods presented in this thesis have some limitations. In comparison to a standard Empirical Risk Minimization model, they require more compute resource in terms of training time, and, as is often the case with new approaches, often more hours of human resource for development and parameter-tuning. However, this limitation is not unique to this work. Another computational aspect is that the INN model used in chapter 6 requires a very large amount of memory. This is due to the lossless transformations that the model performs on the high-dimensional inputs, as opposed to the lossy transforms that are standard in (artificial) neural network (NN) models. In all of these cases, it is reasonable to assume, given the novelty of the presented work, that methods and techniques will be developed, shared, and adopted making all of the approaches presented increasingly feasible.

7.2 Potential Extensions

The research topic of algorithmic fairness is growing at a rapid rate, and the number of developments is growing considerably. As a research community, it seems as though we are only just beginning to scrape the surface of what can be achieved. One of the arguments against adopting ML in practice is the lack of control that can be exerted over a model. Often a question is how to ensure that certain patterns or behaviours are adopted *not*. Algorithmic fairness provides a mechanism that allows us to address some of these issues. It is a topic that allows for a safety net, allowing some form of control over what should and should not be learnt by ML systems. However, there are a significant number of avenues still to investigate.

One of the most significant advances in recent years in improving model performance is the growing number of data augmentation strategies. These are methods by which alterations can be made to the data available for training so that generalisation of the model to the deployment setting is improved. A popular approach is to use predefined transformations of data with the aim of learning a more robust ML model. Data transformation methods such as Mixup (Zhang et al., 2018b) and CutMix (Yun et al., 2019) have been shown to be generally applicable and improve out-of-distribution generalisation performance. Recent work has investigated the automated selection of data transforms that, when applied during training, will produce a model that generalises well using reinforcement learning (Cubuk et al., 2019) and meta-learning (Hataya et al., 2022) to achieve this. A reasonable research direction then may be to investigate whether some data transform, or combination of transforms, can be used to produce a model that is less sensitive to protected attributes for a given dataset and task.

In the same theme of training with more data, so-called foundation models (Bommasani et al., 2021) are becoming inescapable in the world of deep learning. These are models that are phenomenally large with a number of parameters in the order of several billion. They are then trained on a wide-variety of datasets, presenting two related, but distinct possible extensions. First, recent work such as Goyal et al. (2022) shows that training a foundation model with more diverse uncurated images can improve performance on fairness benchmarks related to image classification with respect to gender, skin tone, and age groups compared to specialised models trained on object-orientated datasets such as ImageNet. As uncurated data sources become increasingly multimodal, for example, the recent Radford et al. (2021) model operates on images and text, it is reasonable to assume other modalities will be introduced, such as uncurated video and audio data. Extending fair representations to exist in these cross-modal data domains will be an exciting challenge. Second, while the suggestion in Goyal et al. (2022) is that foundation models trained on sufficiently diverse data may produce fairer outcomes by default, pre-trained models are regularly fine-tuned for specific tasks. This can cause them to under-perform in the original task post fine tuning on out of distribution data (Kumar et al., 2022). Currently, it is unknown if the same holds true for foundation models and if so what the effect on the fairness-enhancing properties would be. It is likely that methods related to fine-tuning these large models to specific tasks while retaining, or improving, on fairness will become an active research area in the near future.

Throughout this thesis, there has been an assumption that the protected attribute labels are known for each sample. Work on Distributionally Robust Optimization aims to give good performance even under distribution shift. An extension of this framework, Group Distributionally Robust Optimization aims to improve the worst-case performance of a known population subgroup. Investigations have recently been carried out to loosen this restriction of requiring knowledge of subgroups based in part on clustering of samples to determine possible groups (Sohoni et al., 2020; Kehrenberg et al., 2022), or various instance reweighting procedures. These procedures can be either based on classification difficulty determined either by an adversarial model (Lahoti et al., 2020), by simply observing the loss after N-training steps (Liu et al., 2021), or using the distance from the average loss (Hashimoto et al., 2018) when training in a repeated setting. The challenge in all of these methods though is that they currently do not align to the widely-adopted fairness criteria discussed in section 2.5. However, given the practical challenges that may be present in obtaining protected attribute labels in some settings, I expect this avenue of research to continue.

Lastly, one of the main themes of this thesis has been to incorporate greater levels of transparency into methods that improve fairness. An underexplored area to promote this is the Bayesian framework, which allows prior expectations to be expressed and different types of uncertainty to be captured. The challenge is to align this framework with existing research avenues. Although there have been fairness-aware Bayesian approaches to model selection via hyper-parameter tuning (Perrone et al., 2021; Sim et al., 2021) and model evaluation (Ji et al., 2020), there is not an obvious approach to align this framework with fair representations. Pursuing this would be an exciting direction.

7.3 Conclusion

The main contribution of this work is providing methods to improve fairness of downstream systems by default, and doing so in a way that enables conversations with stakeholders about how the individual data samples have been altered during training, and importantly, when the model is deployed. The hope is that these conversations will lead to productive discussions about the assumptions and limitations of the system being emulated and how to codify procedures to monitor the impact a model may have. One such procedure to avoid accusation of Disparate Treatment (DT), may be to limit the feature space and design a model focusing on making predictions without needing sensitive attributes to be captured. Although this may help achieve short-term goals, none of the methods discussed in this thesis would be able to be accurately evaluated if that information were not captured in the first place. Furthermore, it would not be possible to investigate this topic if datasets containing this information were not openly available. It may seem counterintuitive, but to ensure the ethical machine learning goals of fairer, more interpretable, more transparent, and more accountable algorithms, we need protected attributes to be recorded. We may not necessarily use these features within our models, but if not captured and made available, we stand little chance of calming ethical concerns about the application of our field; particularly to areas with subjective outcomes.

A related problem is that one of the challenges in producing fairer models ML is the lack of diverse datasets that correspond to realistic tasks. Too often we have to use proxy outcomes and contrived problem-statements to emulate challenges being faced in real-world settings. This is often due to the practical and legal difficulties in sharing data, and also as a method to protect institutions and companies from harm. Clearly, there is no easy solution to this problem. My hope though is that more datasets relevant to real-world applications are made available. To tackle challenges faced today, the research community needs access to a greater number of diverse datasets that capture a wide range of biases.

When a model is optimised for accuracy based on the training data, it is not surprising that patterns within the training data are exploited. The problem with optimization procedures is exactly their success: they lead to outcomes that perform well compared to their objective. Other metrics that the objective was not designed to improve will likely be poorer. If there are additional constraints that need to be satisfied, these must be encoded into the objective. The problem is determining these additional constraints. The benefit of ML is not needing to define *how* a problem should be solved, but rather *what* problem requires a solution. Perhaps fairness then is the natural next step in ML — as well as defining what should be learned, it is important to define what should not.

It may be a reality that we will never be able to produce perfectly fair ML models, particularly during initial exploration of novel use-cases. However, this doesn't mean that fairness-enhancing methods should be abandoned, in fact, quite the opposite. Fairness interventions should be encouraged for all applications. Learning about fairness in more contexts leads to learning more about the capabilities and shortcomings of existing models, datasets, and problem statements resulting in new issues being raised. Thinking about these allows us to make improvements to all of the above in a cyclical fashion. Algorithmic Fairness then is a Sisyphean challenge to overcome, but the benefits that come with the smallest improvements are worthy of the effort.

BIBLIOGRAPHY

Adamski, Igor, Robert Adamski, Tomasz Grel, Adam Jędrych, Kamil Kaczmarek and Henryk Michalewski (2018). 'Distributed Deep Reinforcement Learning: Learn How to Play Atari Games in 21 minutes'. In: *High Performance Computing*. Ed. by Rio Yokota, Michèle Weiland, David Keyes and Carsten Trinitis. Cham: Springer International Publishing, pp. 370–388. ISBN: 978-3-319-92040-5.

Adel, Tameem, Zoubin Ghahramani and Adrian Weller (2018). 'Discovering Interpretable Representations for Both Deep Generative and Discriminative Models'. In: *International Conference on Machine Learning (ICML)*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Pmlr, pp. 50–59. URL: http://proceedings.mlr.press/v80/adel18a.html.

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford and Hanna M. Wallach (2018). 'A Reductions Approach to Fair Classification'. In: *International Conference on Machine Learning (ICML)*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Pmlr, pp. 60–69. URL: http://proceedings.mlr.press/v80/agarwal18a.html.

Agarwal, Alekh, Miroslav Dudik and Zhiwei Steven Wu (Sept. 2019). 'Fair Regression: Quantitative Definitions and Reduction-Based Algorithms'. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 120–129. URL: https://proceedings.mlr.press/v97/agarwal19d.html.

Agarwal, Sushant (2021). 'Trade-Offs between Fairness and Interpretability in Machine Learning'. In: *IJCAI 2021 Workshop on AI for Social Good*.

Albert, Edward Tristram (2019). 'AI in talent acquisition: a review of AI-applications used in recruitment and selection'. In: *Strategic HR Review*.

Ananny, Mike (2018). 'Seeing without knowing : Limitations of the transparency ideal and its application to algorithmic accountability'. In: DOI: 10.1177/1461444816676645.

Angwin, Julia, Jeff Larson, Lauren Kirchner and Surya Mattu (2016). *Machine Bias.* URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ardizzone, Lynton, Carsten Lüth, Jakob Kruse, Carsten Rother and Ullrich Köthe (2019). 'Guided Image Generation with Conditional Invertible Neural Networks'. In: arXiv: 1907.02392.

Arjovsky, Martin, Léon Bottou, Ishaan Gulrajani and David Lopez-Paz (2019). 'Invariant Risk Minimization'. In: arXiv: 1907.02893.

Artificial Intelligence, Select Committee on (2018). *AI in the UK: ready, willing and able?* URL: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm.

Awasthi, Pranjal, Matthäus Kleindessner and Jamie Morgenstern (Aug. 2020). 'Equalized odds postprocessing under imperfect group information'. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics.* Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 1770–1780. URL: https://proceedings.mlr.press/v108/awasthi20a.html.

Barocas, S and M Hardt (2017). *Fairness in Machine Learning.* URL: https://nips.cc/Conferences/2017/Schedule?showEvent=8734.

Barocas, S. and A. Selbst (2016). 'Big data's disparate impact'. In: *California Law Review* 104.3, pp. 671–732.

Barocas, Solon, Moritz Hardt and Arvind Narayanan (2019). *Fairness and Machine Learning.* http://www.fairmlbook.org.

BCS (2022). *BCS Code of Conduct.* URL: https://www.bcs.org/media/2211/bcs-code-of-conduct.pdf (visited on 17/08/2022).

Beaudry, Normand J. and Renato Renner (May 2012). 'An Intuitive Proof of the Data Processing Inequality'. In: *Quantum Info. Comput.* 12.5–6, pp. 432–441. ISSN: 1533-7146.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell (2021). 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922.

Bent, Jason R (2019). 'Is algorithmic affirmative action legal'. In: *The Georgetown Law Journal* 108, p. 803.

Bertrand, Marianne and Sendhil Mullainathan (2004). 'Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination'. In: *American economic review* 94.4, pp. 991–1013.

Beutel, Alex, Jilin Chen, Zhe Zhao and Ed H. Chi (2017). 'Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations'. In: arXiv: 1707.00075.

Beutel, Alex et al. (2019). 'Fairness in Recommendation Ranking through Pairwise Comparisons'. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, pp. 2212–2220. ISBN: 9781450362016. DOI: 10.1145/3292500.3330745. URL: https://doi.org/10.1145/3292500.3330745.

Black, Emily, Samuel Yeom and Matt Fredrikson (2020). 'FlipTest: Fairness Testing via Optimal Transport'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Fat* '20. Barcelona, Spain: Association for Computing Machinery, pp. 111–121. ISBN: 9781450369367. DOI: 10.1145/3351095.3372845.

Bommasani, Rishi et al. (2021). 'On the Opportunities and Risks of Foundation Models'. In: arXiv: 2108.07258.

Bower, Amanda, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargo and Suresh Venkatasubramanian (2017). 'Fair Pipelines'. In: FAT/ML '17. Halifax, NS, Canada. arXiv: 1707.00391.

Brown, Elizabeth and David I Perrett (1993). 'What gives a face its gender?' In: *Perception* 22.7, pp. 829–840.

Brown, Noam and Tuomas Sandholm (2018). 'Superhuman AI for heads-up no-limit poker: Libratus beats top professionals'. In: *Science* 359.6374, pp. 418–424. DOI: 10.1126/science.aao1733.

Bruce, Vicki, A Mike Burton, Elias Hanna, Pat Healey, Oli Mason, Anne Coombes, Rick Fright and Alf Linney (1993). 'Sex Discrimination: How Do We Tell the Difference between Male and Female Faces?' In: *Perception* 22.2. Pmid: 8474840, pp. 131–152. DOI: 10.1068/p220131. URL: https://doi.org/10.1068/p220131.

Butcher, Bradley, Vincent S. Huang, Christopher Robinson, Jeremy Reffin, Sema K. Sgaier, Grace Charles and Novi Quadrianto (2021). 'Causal Datasheet for Datasets: An Evaluation Guide for Real-World Data Analysis and Data Collection Design Using Bayesian Networks'. In: *Frontiers in Artificial Intelligence* 4. ISSN: 2624-8212. DOI: 10.3389/frai.2021.612551.

Calders, Toon and Sicco Verwer (2010). ‘Three naive Bayes approaches for discrimination-free classification’. In: *Data Mining and Knowledge Discovery* 21.2, pp. 277–292. ISSN: 1573-756x. DOI: 10.1007/s10618-010-0190-x.

Calmon, Flávio du Pin, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy and Kush R. Varshney (2017). ‘Optimized Pre-Processing for Discrimination Prevention’. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, pp. 3992–4001. URL: https://proceedings.neurips.cc/paper/2017/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html.

Castro da Silva, Bruno (2019). *UFRGS Entrance Exam and GPA Data*. Version V1. DOI: 10.7910/DVN/O35FW8.

Chiappa, Silvia (2019). ‘Path-Specific Counterfactual Fairness’. In: *AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 7801–7808. DOI: 10.1609/aaai.v33i01.33017801.

Choi, Yunjey, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim and Jaegul Choo (2018). ‘StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation’. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, pp. 8789–8797. DOI: 10.1109/cvpr.2018.00916.

Choi, Yunjey, Youngjung Uh, Jaejun Yoo and Jung-Woo Ha (2020). ‘StarGAN v2: Diverse Image Synthesis for Multiple Domains’. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Chouldechova, Alexandra (2017). ‘Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments’. In: *Big Data* 5.2. Pmid: 28632438, pp. 153–163. ISSN: 2167-6461. DOI: 10.1089/big.2016.0047.

Chuang, Ching-Yao and Youssef Mroueh (2021). ‘Fair Mixup: Fairness via Interpolation’. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=DNl5s5BXeBn.

Chzhen, Evgenii, Christophe Denis, Mohamed Hebiri, Luca Oneto and Massimiliano Pontil (2020). ‘Fair regression with Wasserstein barycenters’. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin. Vol. 33. Curran Associates, Inc., pp. 7321–7331. URL: https://proceedings.neurips.cc/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf.

Citron, Danielle Keats and Frank Pasquale (2014). 'The scored society: Due process for automated predictions'. In: *Wash. L. Rev.* 89, p. 1.

Clanuwat, Tarin, Mikel Bober-Irizar, Asanobu Kitamoto and Alex an Lamb (2018). 'Deep learning for classical japanese literature'. In: arXiv: 1812.01718.

Cohen, Gregory, Saeed Afshar, Jonathan Tapson and Andre Van Schaik (2017). 'EMNIST: Extending MNIST to handwritten letters'. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. Ieee, pp. 2921–2926.

Corbett-Davies, S and S Goel (2018). *Defining and Designing Fair Algorithms*. ICML 2018 Tutorial. URL: https://icml.cc/Conferences/2018/Schedule?showEvent=1862.

Creager, Elliot, Joern-Henrik Jacobsen and Richard Zemel (July 2021). 'Environment Inference for Invariant Learning'. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 2189–2200. URL: https://proceedings.mlr.press/v139/creager21a.html.

Creager, Elliot, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi and Richard S. Zemel (2019). 'Flexibly Fair Representation Learning by Disentanglement'. In: *International Conference on Machine Learning (ICML)*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Pmlr, pp. 1436–1445. URL: http://proceedings.mlr.press/v97/creager19a.html.

Cubuk, Ekin D., Barret Zoph, Dandelion Mane, Vijay Vasudevan and Quoc V. Le (June 2019). 'AutoAugment: Learning Augmentation Strategies From Data'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dedeo, Simon (2014). 'Wrong side of the tracks : Big Data and Protected Categories'. In: arXiv: 1412.4643.

Denton, Emily, Ben Hutchinson, Margaret Mitchell and Timnit Gebru (2019). 'Detecting Bias with Generative Counterfactual Face Attribute Augmentation'. In: *CoRR* abs/1906.06439.

Dheeru, Dua and Efi Karra Taniskidou (2017). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml.

Diakopoulos, Nicholas (2014). 'Algorithmic Accountability Reporting: On the Investigation of Black Boxes'. In: Columbia University. DOI: 10.7916/d8zk5tw2.

Dinh, Laurent, David Krueger and Yoshua Bengio (2014). 'NICE: Non-linear Independent Components Estimation'. In: *International Conference on Learning Representations (ICLR)*.

Donelan, The Rt Hon Michelle (Nov. 2021). *New levelling up plans to improve student outcomes.* URL: https://www.gov.uk/government/news/new-levelling-up-plans-to-improve-student-outcomes.

Doshi-Velez, Finale and Been Kim (Feb. 2017). 'Towards A Rigorous Science of Interpretable Machine Learning'. In: arXiv: 1702.08608.

Durugkar, Ishan P., Ian Gemp and Sridhar Mahadevan (2017). 'Generative Multi-Adversarial Networks'. In: *International Conference on Learning Representations (ICLR).* OpenReview.net. URL: https://openreview.net/forum?id=Byk-VI9eg.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel (2012). 'Fairness through Awareness'. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference.* Itcs '12. Cambridge, Massachusetts: Association for Computing Machinery, pp. 214–226. ISBN: 9781450311151. DOI: 10.1145/2090236.2090255.

Dwork, Cynthia and Christina Ilvento (2019). 'Fairness Under Composition'. In: *Innovations in Theoretical Computer Science (ITCS).*

Edwards, Harrison and Amos J. Storkey (2016). 'Censoring Representations with an Adversary'. In: *International Conference on Learning Representations (ICLR).* Ed. by Yoshua Bengio and Yann LeCun. arXiv: 1511.05897.

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger and Suresh Venkatasubramanian (2015). 'Certifying and Removing Disparate Impact'. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015.* Ed. by Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu and Graham Williams. Acm, pp. 259–268. DOI: 10.1145/2783258.2783311.

Feng, Rui, Yang Yang, Yuehan Lyu, Chenhao Tan and Yizhou an Sun (2019). 'Learning fair representations via an adversarial framework'. In: *ArXiv preprint* abs/1904.13341. URL: https://arxiv.org/abs/1904.13341.

Flores, Anthony W., Kristin Bechtel and Christopher T. Lowenkamp (2016). 'False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks'. In: *Federal Probation* 80, p. 38.

Foulds, James R., Rashidul Islam, Kamrun Naher Keya and Shimei Pan (2020). 'An Intersectional Definition of Fairness'. In: pp. 1918–1921. DOI: 10.1109/ICDE48307.2020.00203.

French, Geoff, Michal Mackiewicz and Mark Fisher (2018). 'Self-ensembling for visual domain adaptation'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rkpoTaxA-.

Friedler, Sorelle A., Carlos Scheidegger and Suresh Venkatasubramanian (2021). 'The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making'. In: *Commun. ACM* 64.4, pp. 136–143. ISSN: 0001-0782. DOI: 10.1145/3433949.

Friedler, Sorelle A., Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton and Derek Roth (2018). 'A comparative study of fairness-enhancing interventions in machine learning'. In: arXiv: 1802.04422.

Fu, S., H. He and Z. Hou (2014). 'Learning Race from Face: A Survey'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.12, pp. 2483–2509. DOI: 10.1109/tpami.2014.2321570.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand and Victor Lempitsky (2016). 'Domain-adversarial training of Neural Networks'. In: *Journal of Machine Learning Research* 17.1, pp. 2096–2030.

Gardner, Jacob R, Paul Upchurch, Matt J Kusner, Yixuan Li, Kilian Q Weinberger, Kavita Bala and John E Hopcroft (2016). 'Deep manifold traversal: Changing labels with convolutional features'. In: *European Conference on Computer Vision (ECCV)*.

Gatys, Leon, Alexander Ecker and Matthias Bethge (2016). 'A Neural Algorithm of Artistic Style'. In: vol. 16. DOI: 10.1167/16.12.326.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford (Nov. 2021). 'Datasheets for Datasets'. In: *Commun. ACM* 64.12, pp. 86–92. ISSN: 0001-0782. DOI: 10.1145/3458723.

Geifman, Yonatan and Ran El-Yaniv (2017). 'Selective Classification for Deep Neural Networks'. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel (2019). 'ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness'. In: *International Conference on Learning Representations (ICLR)*. OpenReview.net. URL: https://openreview.net/forum?id=Bygh9j09KX.

Girshick, Ross (Dec. 2015). 'Fast R-CNN'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

Global Future Council on Human Rights 2016-18 (2018). *How to prevent discriminatory outcomes in machine learning*. Tech. rep. World Economic Forum.

Goyal, Priya, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin and Piotr Bojanowski (2022). 'Vision Models Are More Robust And Fair When Pretrained On Uncurated Images Without Supervision'. In: arXiv: 2202.08360.

Gretton, Arthur, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf and Alexander J Smola (2007). 'A Kernel Approach to Comparing Distributions'. In: *Proceedings of the 22. AAAI Conference on Artificial Intelligence*. Max-Planck-Gesellschaft. Menlo Park, CA, USA: AAAI Press, pp. 1637–1641.

Gretton, Arthur, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf and Alexander Smola (2012). 'A Kernel Two-sample Test'. In: *Journal of Machine Learning Research (JMLR)* 13, pp. 723–773.

Gretton, Arthur, Olivier Bousquet, Alex Smola and Bernhard Schölkopf (2005). 'Measuring Statistical Dependence with Hilbert-Schmidt Norms'. In: *Algorithmic Learning Theory*. Ed. by Sanjay Jain, Hans Ulrich Simon and Etsuji Tomita. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 63–77. ISBN: 978-3-540-31696-1.

Grgic-Hlaca, Nina, Elissa M. Redmiles, Krishna P. Gummadi and Adrian Weller (2018). 'Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction'. In: *Proceedings of the 2018 World Wide Web Conference*. Ed. by Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas and Panagiotis G. Ipeirotis. Www '18. Lyon, France: International World Wide Web Conferences Steering Committee, pp. 903–912. ISBN: 9781450356398. DOI: 10.1145/3178876.3186138.

Hardt, Moritz, Eric Price and Nati Srebro (2016). 'Equality of Opportunity in Supervised Learning'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Harned, Zach and Hanna Wallach (2019). 'Stretching Human Laws to Apply to Machines: The Dangers of a "Colorblind" Computer'. In: *Florida State Univeristy Law Review* 47, p. 617.

Harrisburg University, Pennsylvania USA (May 2020). *Facial Recognition Software Predicts Criminality, Researchers Say*. URL: https://cacm.acm.org/careers/244713-facial-recognition-software-predicts-criminality-researchers-say/fulltext.

Hashimoto, Tatsunori, Megha Srivastava, Hongseok Namkoong and Percy Liang (Oct. 2018). 'Fairness Without Demographics in Repeated Loss Minimization'. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1929–1938. URL: https://proceedings.mlr.press/v80/hashimoto18a.html.

Hataya, Ryuichiro, Jan Zdenek, Kazuki Yoshizoe and Hideki Nakayama (Jan. 2022). 'Meta Approach to Data Augmentation Optimization'. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2574–2583.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun (2016). 'Deep Residual Learning for Image Recognition'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, pp. 770–778. DOI: 10.1109/cvpr.2016.90.

Heidari, Hoda, Michele Loi, Krishna P. Gummadi and Andreas Krause (2018). 'A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity'. In: arXiv: 1809.03400.

Hendrickx, Kilian, Lorenzo Perini, Dries Van der Plas, Wannes Meert and Jesse Davis (2021). 'Machine learning with a reject option: A survey'. In: *arXiv preprint arXiv:2107.11277*.

Higgins, Irina, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed and Alexander Lerchner (2017). 'beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework'. In: *International Conference on Learning Representations (ICLR)*. OpenReview.net. URL: https://openreview.net/forum?id=Sy2fzU9gl.

Hinnefeld, J. Henry, Peter Cooman, Nat Mammo and Rupert Deese (2018). 'Evaluating Fairness Metrics in the Presence of Dataset Bias'. In: arXiv: 1809.09245.

Hoffman, Judy, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros and Trevor Darrell (July 2018). 'CyCADA: Cycle-Consistent Adversarial Domain Adaptation'. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1989–1998. URL: https://proceedings.mlr.press/v80/hoffman18a.html.

Holland, Paul W. (1986). 'Statistics and Causal Inference'. In: *Journal of the American Statistical Association* 81.396, pp. 945–960. DOI: 10.1080/01621459.1986.10478354.

Holmstrom, Mark, Dylan Liu and Christopher Vo (2016). ‘Machine learning applied to weather forecasting’. In: *Meteorol. Appl*, pp. 1–5.

Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík and Hanna M. Wallach (2019). ‘Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?’ In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*. Ed. by Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox and Vassilis Kostakos. Acm, p. 600. DOI: 10.1145/3290605.3300830.

Hume, David (2000). *An enquiry concerning human understanding: A critical edition*. Vol. 3. Oxford University Press.

Hwang, Tim (2018). ‘Computational Power And The Social Impact Of Artificial Intelligence’. In: *Available at SSRN 3147971*, pp. 1–44.

Ioffe, Sergey and Christian Szegedy (2015). ‘Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift’. In: *International Conference on Machine Learning (ICML)*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 448–456. URL: http://proceedings.mlr.press/v37/ioffe15.html.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros (2017). ‘Image-to-Image Translation with Conditional Adversarial Networks’. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, pp. 5967–5976. DOI: 10.1109/cvpr.2017.632.

Jacobsen, Jörn-Henrik, Jens Behrmann, Richard S. Zemel and Matthias Bethge (2019). ‘Excessive Invariance Causes Adversarial Vulnerability’. In: *International Conference on Learning Representations (ICLR)*. OpenReview.net. URL: https://openreview.net/forum?id=BkfbpsAcF7.

Jacobsen, Jörn-Henrik, Arnold W. M. Smeulders and Edouard Oyallon (2018). ‘i-RevNet: Deep Invertible Networks’. In: *International Conference on Learning Representations (ICLR)*. OpenReview.net. URL: https://openreview.net/forum?id=HJsjkMb0Z.

Jaiswal, Ayush, Rex Yue Wu, Wael Abd-Almageed and Prem Natarajan (2018a). ‘Unsupervised Adversarial Invariance’. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi and Roman Garnett, pp. 5097–5107. URL: https://proceedings.neurips.cc/paper/2018/hash/03e7ef47cee6fa4ae7567394b99912b7-Abstract.html.

– (2018b). ‘Unsupervised Adversarial Invariance’. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/03e7ef47cee6fa4ae7567394b99912b7-Paper.pdf.

Ji, Disi, Padhraic Smyth and Mark Steyvers (2020). 'Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin. Vol. 33. Curran Associates, Inc., pp. 18600–18612. URL: https://proceedings.neurips.cc/paper/2020/file/d83de59e10227072a9c034ce10029c39-Paper.pdf.

Jiang, Heinrich and Ofir Nachum (2020). 'Identifying and Correcting Label Bias in Machine Learning'. In: *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. Pmlr, pp. 702–712. URL: http://proceedings.mlr.press/v108/jiang20a.html.

Johnson, Carrie (Jan. 2022). 'Flaws plague a tool meant to help low-risk federal prisoners win early release'. In: *NPR*. URL: https://www.npr.org/2022/01/26/1075509175/justice-department-algorithm-first-step-act?t=1644970124044.

Johnson, Justin, Alexandre Alahi and Li Fei-Fei (2016). 'Perceptual losses for real-time style transfer and super-resolution'. In: *European Conference on Computer Vision (ECCV)*.

Joshi, Shalmali an (2019). 'Towards Realistic Individual Recourse and Actionable Explanation'. In: arXiv: 1907.09615.

Jumper, John et al. (Aug. 2021). 'Highly accurate protein structure prediction with AlphaFold'. In: *Nature* 596.7873, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2.

Kallus, Nathan and Angela Zhou (2018). 'Residual Unfairness in Fair Machine Learning from Prejudiced Data'. In: *International Conference on Machine Learning (ICML)*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Pmlr, pp. 2444–2453. URL: http://proceedings.mlr.press/v80/kallus18a.html.

Kamiran, F. (2011). *Discrimination-aware Classification*. Tech. rep. Eindhoven: Technische Universiteit Eindhoven, p. 157. DOI: 10.6100/ir717576.

Kamiran, Faisal and Toon Calders (2009). 'Classifying without discriminating'. In: *2009 2nd International Conference on Computer, Control and Communication, IC4 2009*. ISBN: 9781424433148. DOI: 10.1109/ic4.2009.4909197.

– (2012). 'Data preprocessing techniques for classification without discrimination'. In: *Knowledge and Information Systems* 33.1, pp. 1–33. ISSN: 02191377. DOI: 10.1007/s10115-011-0463-8.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh and Jun Sakuma (2012). 'Fairness-aware classifier with prejudice remover regularizer'. In: *European conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 35–50.

Karimi, Amir-Hossein, Bernhard Schölkopf and Isabel Valera (2021). 'Algorithmic Recourse: From Counterfactual Explanations to Interventions'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 353–362. ISBN: 9781450383097. DOI: 10.1145/3442188.3445899.

Kearns, Michael J., Seth Neel, Aaron Roth and Zhiwei Steven Wu (2018). 'Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness'. In: *International Conference on Machine Learning (ICML)*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Pmlr, pp. 2569–2577. URL: http://proceedings.mlr.press/v80/kearns18a.html.

Kehrenberg, Thomas, Zexun Chen and Novi Quadrianto (2020a). 'Tuning Fairness by Balancing Target Labels'. In: *Frontiers in Artificial Intelligence* 3, p. 33. ISSN: 2624-8212. DOI: 10.3389/frai.2020.00033.

Kehrenberg, Thomas, Myles Bartlett, Viktoriia Sharmanska and Novi Quadrianto (2022). 'Addressing Missing Sources with Adversarial Support-Matching'. In: arXiv: 2203.13154.

Kehrenberg, Thomas, Myles Bartlett, Oliver Thomas and Novi Quadrianto (2020b). 'Null-Sampling for Interpretable and Fair Representations'. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox and Jan-Michael Frahm. Cham: Springer International Publishing, pp. 565–580. ISBN: 978-3-030-58574-7.

Kennefick, Ciara (2018). 'The Contribution of Contemporary Mathematics to Contractual Fairness in Equity, 1751–1867'. In: *The Journal of Legal History* 39.3, pp. 307–339. DOI: 10.1080/01440365.2018.1532657.

Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing and Bernhard Schölkopf (2017). 'Avoiding Discrimination through Causal Reasoning'. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf.

Kim, Byungju, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim and Junmo Kim (2019). 'Learning Not to Learn: Training Deep Neural Networks With Biased Data'. In: *IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR). Computer Vision Foundation / IEEE, pp. 9012–9020. DOI: 10.1109/cvpr.2019.00922.

Kim, Hyunjik and Andriy Mnih (2018). 'Disentangling by Factorising'. In: *International Conference on Machine Learning (ICML)*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Pmlr, pp. 2654–2663. URL: http://proceedings.mlr.press/v80/kim18b.html.

Kingma, Diederik P. and Jimmy Ba (2015). 'Adam: A Method for Stochastic Optimization'. In: *International Conference on Learning Representations (ICLR)*. Ed. by Yoshua Bengio and Yann LeCun. arXiv: 1412.6980.

Kingma, Diederik P. and Prafulla Dhariwal (2018). 'Glow: Generative Flow with Invertible 1x1 Convolutions'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi and Roman Garnett, pp. 10236–10245. URL: https://proceedings.neurips.cc/paper/2018/hash/d139db6a236200b21cc7f752979132d0-Abstract.html.

Kingma, Diederik P. and Max Welling (2014). 'Auto-Encoding Variational Bayes'. In: *International Conference on Learning Representations (ICLR)*. Ed. by Yoshua Bengio and Yann LeCun. arXiv: 1312.6114.

Klambauer, Günter, Thomas Unterthiner, Andreas Mayr and Sepp Hochreiter (2017). 'Self–Normalizing Neural Networks'. In: *Advances in Neural Information Processing Systems*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, pp. 971–980. URL: https://proceedings.neurips.cc/paper/2017/hash/5d44ee6f2c3f71b73125876103c8f6c4-Abstract.html.

Kleinberg, Jon, Sendhil Mullainathan and Manish Raghavan (2017). 'Inherent Trade-Offs in the Fair Determination of Risk Scores'. In: *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Ed. by Christos H. Papadimitriou. Vol. 67. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 43:1–43:23. ISBN: 978-3-95977-029-3. DOI: 10.4230/LIPIcs.ITCS.2017.43. URL: http://drops.dagstuhl.de/opus/volltexte/2017/8156.

Kumar, Ananya, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma and Percy Liang (2022). 'Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=UYneFzXSJWh.

Kusner, Matt J., Joshua R. Loftus, Chris Russell and Ricardo Silva (2017). 'Counterfactual Fairness'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, pp. 4066–4076. URL: https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Lahoti, Preethi, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang and Ed Chi (2020). 'Fairness without Demographics through Adversarially Reweighted Learning'. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin. Vol. 33. Curran Associates, Inc., pp. 728–740. URL: https://proceedings.neurips.cc/paper/2020/file/07fc15c9d169ee48573edd749d25945d-Paper.pdf.

Larson, Jeff, Surya Mattu, Lauren Kirchner and Julia Angwin (2016). 'How we analyzed the COMPAS recidivism algorithm'. In: *ProPublica (5 2016)* 9.1, pp. 3–3.

LeCun, Yann, Corinna Cortes and CJ Burges (2010). 'MNIST handwritten digit database'. In: *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2.

Levi, G. and T. Hassncer (2015). 'Age and gender classification using convolutional neural networks'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 34–42. DOI: 10.1109/cvprw.2015.7301352.

Li, Tian, Maziar Sanjabi, Ahmad Beirami and Virginia Smith (2020). 'Fair Resource Allocation in Federated Learning'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ByexElSYDr.

Liu, Evan Z, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang and Chelsea Finn (July 2021). 'Just Train Twice: Improving Group Robustness without Training Group Information'. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 6781–6792. URL: https://proceedings.mlr.press/v139/liu21f.html.

Liu, Liyuan, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao and Jiawei Han (2020). 'On the Variance of the Adaptive Learning Rate and Beyond'. In: *International Conference on Learning Representations (ICLR)*. OpenReview.net. URL: https://openreview.net/forum?id=rkgz2aEKDr.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz and Moritz Hardt (2019). 'Delayed Impact of Fair Machine Learning'. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by Sarit Kraus. ijcai.org, pp. 6196–6200. DOI: 10.24963/ijcai.2019/862.

Liu, Ziwei, Ping Luo, Xiaogang Wang and Xiaoou Tang (2015). 'Deep Learning Face Attributes in the Wild'. In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, pp. 3730–3738. DOI: 10.1109/iccv.2015.425.

Lohia, Pranay K., Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney and Ruchir Puri (2019). 'Bias Mitigation Post-processing for Individual and Group Fairness'. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2847–2851. DOI: 10.1109/ICASSP.2019.8682620.

Lopez-Paz, David, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf and Léon Bottou (2017). 'Discovering Causal Signals in Images'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, pp. 58–66. DOI: 10.1109/cvpr.2017.14.

Louizos, Christos, Kevin Swersky, Yujia Li, Max Welling and Richard S. Zemel (2016). 'The Variational Fair Autoencoder'. In: *International Conference on Learning Representations (ICLR)*. Ed. by Yoshua Bengio and Yann LeCun. arXiv: 1511.00830.

Madras, David, Elliot Creager, Toniann Pitassi and Richard Zemel (2019). 'Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data'. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Fat*'19. Atlanta, GA, USA: Association for Computing Machinery, pp. 349–358. ISBN: 9781450361255. DOI: 10.1145/3287560.3287564.

Madras, David, Elliot Creager, Toniann Pitassi and Richard S. Zemel (2018a). 'Learning Adversarially Fair and Transferable Representations'. In: *International Conference on Machine Learning (ICML)*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Pmlr, pp. 3381–3390. URL: http://proceedings.mlr.press/v80/madras18a.html.

Madras, David, Toniann Pitassi and Richard S. Zemel (2018b). 'Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi and Roman Garnett, pp. 6150–6160. URL: https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Mahendran, Aravindh and Andrea Vedaldi (2015). ‘Understanding deep image representations by inverting them’. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, pp. 5188–5196. DOI: 10.1109/cvpr.2015.7299155.

Maity, Subha, Debarghya Mukherjee, Mikhail Yurochkin and Yuekai Sun (Dec. 2021). ‘Does enforcing fairness mitigate biases caused by subpopulation shift?’ In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin. Vol. 35. Curran Associates, Inc., pp. 193–203. URL: https://proceedings.neurips.cc/paper/2021/hash/d800149d2f947ad4d64f34668f8b20f6-Abstract.htm.

Mary, Jérémie, Clément Calauzènes and Noureddine El Karoui (2019). ‘Fairness-Aware Learning for Continuous Attributes and Treatments’. In: *International Conference on Machine Learning (ICML)*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Pmlr, pp. 4382–4391. URL: http://proceedings.mlr.press/v97/mary19a.html.

McNamara, Daniel, Cheng Soon Ong and Robert C. Williamson (2019). ‘Costs and Benefits of Fair Representation Learning’. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’19. Honolulu, HI, USA: Association for Computing Machinery, pp. 263–270. ISBN: 9781450363242. DOI: 10.1145/3306618.3317964.

Merler, Michele, Nalini K. Ratha, Rogério Schmidt Feris and John R. Smith (2019). ‘Diversity in Faces’. In: arXiv: 1901.10436.

Miller, Tim (2019). ‘Explanation in artificial intelligence: Insights from the social sciences’. In: *Artificial Intelligence* 267, pp. 1–38. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2018.07.007.

Mirza, Mehdi and Simon Osindero (2014). ‘Conditional Generative Adversarial Nets’. In: arXiv: 1411.1784.

Molnar, Christoph (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. URL: https://christophm.github.io/interpretable-ml-book/.

Mooij, Joris M., Dominik Janzing, Jonas Peters and Bernhard Schölkopf (2009). ‘Regression by dependence minimization and its application to causal inference in additive noise models’. In: *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*. Ed. by Andrea Pohoreckyj Danyluk, Léon Bottou and Michael L. Littman. Vol. 382. ACM International Conference Proceeding Series. Acm, pp. 745–752. DOI: 10.1145/1553374.1553470.

Moro, Sérgio, Paulo Cortez and Paulo Rita (2014). 'A data-driven approach to predict the success of bank telemarketing'. In: *Decision Support Systems* 62, pp. 22–31. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2014.03.001.

Mozannar, Hussein and David A. Sontag (2020). 'Consistent Estimators for Learning to Defer to an Expert'. In: *International Conference on Machine Learning (ICML)*. Vol. 119. Proceedings of Machine Learning Research. Pmlr, pp. 7076–7087. URL: http://proceedings.mlr.press/v119/mozannar20b.html.

Mukherjee, Debarghya, Mikhail Yurochkin, Moulinath Banerjee and Yuekai Sun (July 2020). 'Two Simple Ways to Learn Individual Fairness Metrics from Data'. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 7097–7107. URL: https://proceedings.mlr.press/v119/mukherjee20a.html.

Munoz, Cecilia, Megan Smith and DJ Patil (2010). 'Big Data : A Report on Algorithmic Systems , Opportunity , and Civil Rights Big Data : A Report on Algorithmic Systems , Opportunity , and Civil Rights'. In: arXiv: 1011.1669.

Netzer, Yuval, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu and Andrew Y. Ng (2011). 'Reading Digits in Natural Images with Unsupervised Feature Learning'. In: *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*. URL: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

Oneto, L., Michele Donini, Massimiliano Pontil and Andreas Maurer (2020a). 'Learning Fair and Transferable Representations with Theoretical Guarantees'. In: *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 30–39.

Oneto, Luca, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer and Massimiliano Pontil (2020b). 'Exploiting MMD and Sinkhorn Divergences for Fair and Transferable Representation Learning'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin. Vol. 33. Curran Associates, Inc., pp. 15360–15370. URL: https://proceedings.neurips.cc/paper/2020/hash/af9c0e0c1dee63e5acad8b7ed1a5be96-Abstract.html.

Oord, Aäron van den, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals and Alex Graves (2016). 'Conditional Image Generation with PixelCNN Decoders'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon and Roman Garnett, pp. 4790–4798. URL: https://proceedings.neurips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html.

P, Deepak, Sanil V and Joemon M. Jose (June 2021). 'On Fairness and Interpretability'. In: arXiv: 2106.13271 [cs.CY].

Park, Song, Sanghyuk Chun, Junbum Cha, Bado Lee and Hyunjung Shim (2021). *Multiple Heads are Better than One: Few-shot Font Generation with Multiple Localized Experts.* arXiv: 2104.00887.

Pearl, Judea (2009). *Causality.* 2nd ed. Cambridge: Cambridge University Press. DOI: 10.1017/cbo9780511803161.

Pedregosa, F. et al. (2011). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pedreshi, Dino, Salvatore Ruggieri and Franco Turini (2008). 'Discrimination-aware data mining'. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08.* Ed. by Ying Li, Bing Liu and Sunita Sarawagi. Acm, pp. 560–568. ISBN: 9781605581934. DOI: 10.1145/1401890.1401959.

Perrone, Valerio, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi and Cédric Archambeau (2021). 'Fair Bayesian Optimization'. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* New York, NY, USA: Association for Computing Machinery, pp. 854–863. ISBN: 9781450384735. URL: https://doi.org/10.1145/3461702.3462629.

Pole, J.R. (1978). *The Pursuit of Equality in American History.* California Library Reprint Series. University of California Press. ISBN: 9780520032866. URL: https://books.google.co.uk/books?id=QZWQiMpgf84C.

Quadrianto, Novi and Viktoriia Sharmanska (2017). 'Recycling Privileged Learning and Distribution Matching for Fairness'. In: *Advances in Neural Information Processing Systems (NeurIPS).* Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, pp. 677–688. URL: https://proceedings.neurips.cc/paper/2017/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Quadrianto, Novi, Viktoriia Sharmanska and Oliver Thomas (2019). 'Discovering Fair Representations in the Data Domain'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Computer Vision Foundation / IEEE, pp. 8227–8236. DOI: 10.1109/cvpr.2019.00842.

Radford, Alec et al. (July 2021). 'Learning Transferable Visual Models From Natural Language Supervision'. In: *Proceedings of the 38th International Conference on Machine Learning.* Ed. by

Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

Rahimi, Ali and Benjamin Recht (2007). 'Random Features for Large-Scale Kernel Machines'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by John C. Platt, Daphne Koller, Yoram Singer and Sam T. Roweis. Curran Associates, Inc., pp. 1177–1184. URL: https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Rare Recruitment (2021). *Target Oxbridge*. Target Oxbridge. URL: https://targetoxbridge.co.uk/the%5C%5Fprogramme.html (visited on 01/08/2021).

Ravuri, Suman et al. (Sept. 2021). 'Skilful precipitation nowcasting using deep generative models of radar'. In: *Nature* 597.7878, pp. 672–677. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03854-z.

Redmond, Michael and Alok Baveja (2002). 'A data-driven software tool for enabling cooperative information sharing among police departments'. In: *European Journal of Operational Research* 141.3, pp. 660–678. URL: https://EconPapers.repec.org/RePEc:eee:ejores:v:141:y:2002:i:3:p:660-678.

Rezende, Danilo Jimenez and Shakir Mohamed (2015). 'Variational Inference with Normalizing Flows'. In: *International Conference on Machine Learning (ICML)*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1530–1538. URL: http://proceedings.mlr.press/v37/rezende15.html.

Ribeiro, Marco Túlio, Sameer Singh and Carlos Guestrin (2016). '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. Ed. by Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen and Rajeev Rastogi. Acm, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

Roh, Yuji, Kangwook Lee, Steven Euijong Whang and Changho Suh (2021a). 'FairBatch: Batch Selection for Model Fairness'. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=YNnpaAKeCfx.

Roh, Yuji, Kangwook Lee, Steven Whang and Changho Suh (2021b). 'Sample selection for fair and robust training'. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin. Vol. 34. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2021/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html.

Saito, Kuniaki, Kohei Watanabe, Yoshitaka Ushiku and Tatsuya Harada (June 2018). 'Maximum Classifier Discrepancy for Unsupervised Domain Adaptation'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saleiro, Pedro an (2018). 'Aequitas: A Bias and Fairness Audit Toolkit'. In: arXiv: 1811.05577.

Sattigeri, P., S. C. Hoffman, V. Chenthamarakshan and K. R. Varshney (2019). 'Fairness GAN: Generating datasets with fairness properties using a generative adversarial network'. In: *IBM Journal of Research and Development* 63.4/5, 3:1–3:9. DOI: 10.1147/JRD.2019.2945519.

Savani, Yash, Colin White and Naveen Sundar Govindarajulu (2020). 'Intra-Processing Methods for Debiasing Neural Networks'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin. Vol. 33. Curran Associates, Inc., pp. 2798–2810. URL: https://proceedings.neurips.cc/paper/2020/file/1d8d70dddf147d2d92a634817f01b239-Paper.pdf.

Shalit, Uri, Fredrik D. Johansson and David A. Sontag (2017). 'Estimating individual treatment effect: generalization bounds and algorithms'. In: *International Conference on Machine Learning (ICML)*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. Pmlr, pp. 3076–3085. URL: http://proceedings.mlr.press/v70/shalit17a.html.

Sharmanska, Viktoriia, Lisa Anne Hendricks, Trevor Darrell and Novi Quadrianto (2020). 'Contrastive Examples for Addressing the Tyranny of the Majority'. In: arXiv: 2004.06524.

Silver, David et al. (2017). 'Mastering the game of go without human knowledge'. In: *nature* 550.7676, pp. 354–359.

Sim, Rachael Hwee Ling, Yehong Zhang, Bryan Kian Hsiang Low and Patrick Jaillet (July 2021). 'Collaborative Bayesian Optimization with Fair Regret'. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 9691–9701. URL: https://proceedings.mlr.press/v139/sim21b.html.

Simons, Joshua, Sophia Adams Bhatti and Adrian Weller (2021). 'Machine Learning and the Meaning of Equal Treatment'. In: *Aies*.

Simonyan, Karen and Andrew Zisserman (2015). 'Very Deep Convolutional Networks for Large-Scale Image Recognition'. In: *International Conference on Learning Representations (ICLR)*. Ed. by Yoshua Bengio and Yann LeCun. arXiv: 1409.1556.

Singh, Arti, Baskar Ganapathysubramanian, Asheesh Kumar Singh and Soumik Sarkar (2016). 'Machine Learning for High-Throughput Stress Phenotyping in Plants'. In: *Trends in Plant Science* 21.2, pp. 110–124. ISSN: 1360-1385. DOI: https://doi.org/10.1016/j.tplants.2015.10.015.

Sohoni, Nimit, Jared Dunnmon, Geoffrey Angus, Albert Gu and Christopher Ré (2020). 'No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin. Vol. 33. Curran Associates, Inc., pp. 19339–19352. URL: https://proceedings.neurips.cc/paper/2020/file/e0688d13958a19e087e123148555e4b4-Paper.pdf.

Song, Jiaming, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao and Stefano Ermon (Apr. 2019). 'Learning Controllable Fair Representations'. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 2164–2173. URL: https://proceedings.mlr.press/v89/song19a.html.

Song, Le, Alex Smola, Arthur Gretton, Justin Bedo and Karsten Borgwardt (2012). 'Feature Selection via Dependence Maximization'. In: *Journal of Machine Learning Research (JMLR)* 13, pp. 1393–1434.

Sweeney, Latanya (May 2013). 'Discrimination in Online Ad Delivery'. In: *Commun. ACM* 56.5, pp. 44–54. ISSN: 0001-0782. DOI: 10.1145/2447976.2447990.

Thanh, Binh Luong, Salvatore Ruggieri and Franco Turini (2011). 'k-NN as an implementation of situation testing for discrimination discovery and prevention'. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*. Ed. by Chid Apté, Joydeep Ghosh and Padhraic Smyth. Acm, pp. 502–510. DOI: 10.1145/2020408.2020488.

Thomas, Oliver, Miri Zilka, Adrian Weller and Novi Quadrianto (2021). 'An Algorithmic Framework for Positive Action'. In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '21. –, NY, USA: Association for Computing Machinery. ISBN: 9781450385534. DOI: 10.1145/3465416.3483303.

Tolan, Songül (2019). 'Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges'. In: arXiv: 1901.04730.

Ulyanov, Dmitry, Vadim Lebedev, Andrea Vedaldi and Victor S. Lempitsky (2016). 'Texture Networks: Feed-forward Synthesis of Textures and Stylized Images'. In: *International Conference on Machine Learning (ICML)*. Ed. by Maria-Florina Balcan and Kilian Q. Weinberger. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1349–1357. URL: http://proceedings.mlr.press/v48/ulyanov16.html.

Ulyanov, Dmitry, Andrea Vedaldi and Victor S. Lempitsky (2017). 'Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, pp. 4105–4113. DOI: 10.1109/cvpr.2017.437.

Ustun, Berk, Alexander Spangher and Yang Liu (2019). 'Actionable Recourse in Linear Classification'. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Fat*'19. Atlanta, GA, USA: Association for Computing Machinery, pp. 10–19. ISBN: 9781450361255. DOI: 10.1145/3287560.3287566.

Vincent, James (2017). *DeepMind's AI became a superhuman chess player in a few hours, just for fun*. URL: https://www.theverge.com/2017/12/6/16741106/deepmind-ai-chess-alphazero-shogi-go.

Wachter, Sandra, Brent Mittelstadt and Luciano Floridi (2017). 'Transparent, explainable, and accountable AI for robotics'. In: *Science Robotics* 2.6, eaan6080. DOI: 10.1126/scirobotics.aan6080.

Wachter, Sandra, Brent Mittelstadt and Chris Russell (2018). 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR'. In: *Harvard Journal of Law & Technology* 31.2.

– (2020). 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI'. In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.3547922. URL: http://dx.doi.org/10.2139/ssrn.3547922.

Wadsworth, Christina, Francesca Vera and Chris Piech (2018). 'Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction'. In: arXiv: 1807.00199.

Wang, Jing, Jiahong Chen, Jianzhe Lin, Leonid Sigal and Clarence W. de Silva (2021). 'Discriminative feature alignment: Improving transferability of unsupervised domain adaptation by Gaussian-guided latent alignment'. In: *Pattern Recognition* 116, p. 107943. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2021.107943.

Wang, Yilun and Michal Kosinski (Feb. 2018). ‘Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.’ In: *Journal of Personality and Social Psychology* 114.2, pp. 246–257. DOI: 10.1037/pspa0000098.

Wang, Zeyu, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata and Olga Russakovsky (June 2020). ‘Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation’. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Wick, Michael L., Swetasudha Panda and Jean-Baptiste Tristan (2019). ‘Unlocking Fairness: a Trade-off Revisited’. In: *Advances in Neural Information Processing Systems (NeurIPS).* Ed. by Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox and Roman Garnett, pp. 8780–8789. URL: https://proceedings.neurips.cc/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html.

Winter, Eyal (2002). ‘The shapley value’. In: *Handbook of game theory with economic applications* 3, pp. 2025–2054.

Woodworth, Blake, Suriya Gunasekar, Mesrob I. Ohannessian and Nathan Srebro (July 2017). ‘Learning Non-Discriminatory Predictors’. In: *Proceedings of the 2017 Conference on Learning Theory.* Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, pp. 1920–1953. URL: https://proceedings.mlr.press/v65/woodworth17a.html.

Working Group (2017). *Machine learning: the power and promise of computers that learn by example.* Tech. rep. The Royal Society.

Wu, Xiaolin and Xi Zhang (2016). ‘Automated Inference on Criminality using Face Images’. In: *CoRR* abs/1611.04135. arXiv: 1611.04135. URL: http://arxiv.org/abs/1611.04135.

Xiang, Alice (2021). ‘Reconciling Legal and Technical Approaches to Algorithmic Bias’. In: *Tennessee Law Review* 88. URL: https://ssrn.com/abstract=3650635.

Xiao, Han, Kashif Rasul and Roland Vollgraf (2017). ‘Fashion-mnist: a novel image dataset for benchmarking machine learnin’. In: arXiv: 1708.07747.

Xiao, Taihong, Jiapeng Hong and Jinwen Ma (2018). ‘DNA-GAN: Learning disentangled representations from multi-attribute images’. In: *ICLR workshop.*

Xu, Depeng, Shuhan Yuan, Lu Zhang and Xintao Wu (2018). ‘FairGAN: Fairness-aware Generative Adversarial Networks’. In: *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018.* Ed. by Naoki Abe et al. Ieee, pp. 570–575. DOI: 10.1109/BigData.2018.8622525.

Yan, Bobby, Skyler Seto and Nicholas Apostoloff (2022). ‘FORML: Learning to Reweight Data for Fairness’. In: *CoRR* abs/2202.01719. arXiv: 2202.01719.

Yang, Jianfei, Han Zou, Yuxun Zhou, Zhaoyang Zeng and Lihua Xie (2020). ‘Mind the Discriminability: Asymmetric Adversarial Domain Adaptation’. In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox and Jan-Michael Frahm. Cham: Springer International Publishing, pp. 589–606. ISBN: 978-3-030-58586-0.

Yeom, Samuel and Michael Carl Tschantz (2018). ‘Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews’. In: arXiv: 1808.08619.

Yun, Sangdoo, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe and Youngjoon Yoo (Oct. 2019). ‘CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features’. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez-Rodriguez and Krishna P. Gummadi (2017a). ‘Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment’. In: *Proceedings of the 26th International Conference on World Wide Web*. Ed. by Rick Barrett, Rick Cummings, Eugene Agichtein and Evgeniy Gabrilovich. Www ’17. Perth, Australia: Acm, pp. 1171–1180. ISBN: 9781450349130. DOI: 10.1145/3038912.3052660.

– (2017b). ‘Fairness Constraints: Mechanisms for Fair Classification’. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*. Ed. by Aarti Singh and Xiaojin (Jerry) Zhu. Vol. 54. Proceedings of Machine Learning Research. Pmlr, pp. 962–970. URL: http://proceedings.mlr.press/v54/zafar17a.html.

Zaheer, Manzil, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov and Alexander J. Smola (2017). ‘Deep Sets’. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett, pp. 3391–3401. URL: https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Zemel, Richard S., Yu Wu, Kevin Swersky, Toniann Pitassi and Cynthia Dwork (2013). ‘Learning Fair Representations’. In: *International Conference on Machine Learning (ICML)*. Vol. 28. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 325–333. URL: http://proceedings.mlr.press/v28/zemel13.html.

Zhang, Brian Hu, Blake Lemoine and Margaret Mitchell (2018a). ‘Mitigating Unwanted Biases with Adversarial Learning’. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and*

Society. Aies ’18. New Orleans, LA, USA: Association for Computing Machinery, pp. 335–340. ISBN: 9781450360128. DOI: 10.1145/3278721.3278779.

Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin and David Lopez-Paz (2018b). ‘mixup: Beyond Empirical Risk Minimization’. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=r1Ddp1-Rb.

Zhang, Quanshi and Song-Chun Zhu (2018). ‘Visual interpretability for Deep Learning: a survey’. In: *Frontiers of Information Technology & Electronic Engineering* 19.1, pp. 27–39.

Zhu, Jun-Yan, Taesung Park, Phillip Isola and Alexei A. Efros (2017). ‘Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks’. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, pp. 2242–2251. DOI: 10.1109/iccv.2017.244.

Zliobaite, Indre (2015). ‘A survey on measuring indirect discrimination in machine learning’. In: arXiv: 1511.00148.